Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 18 (1994), 41-63.

A METHOD TO QUANTIFY THE ERROR DISTRIBUTION IN CONFUSION MATRICES

R.J.J.H. van Son

Abstract

In this paper a method is proposed to derive the effective number of different error classes from an identification experiment, combined with a method to determine the (non-) overlap in error classes between experiments. Both methods are based on standard information theory, reformulated to give more transparent results. Some examples from the literature are added at the end of the paper.

1 Introduction

Error-rates are generally used to summarize the results of identification and classification experiments with speech. Contrary to what its name suggests, the error-rate is not so much about errors but about correct responses. Although there is generally only a single correct response to a stimulus, there can be many *in*correct responses. The way the errors are distributed cannot be read from the error rate. As an example, I have included some hypothetical confusion matrices in this paper (Figures 1-4) which all share the same error rate, 33.3%, but have very different structures. For many purposes, it makes a lot of difference which of the potential errors are realized in an experiment, and how often.

In this paper I will propose a method to quantify the extent to which errors are "dispersed" or "clustered" in an experiment, or more precisely, in the confusion matrix that describes the results of the experiment. Furthermore, I will show that this method can also be used to quantify the *differences* between confusion matrices with respect to the error classes actually present in them. To derive measures of error dispersion and difference between confusion matrices, classical information theory is used to estimate the association between stimuli and responses. Most of the basic theory presented here can be found in any textbook on formal information theory (e.g., Khinchin, 1957; Sveshnikov, 1968; Press et al., 1988). However, the theory is partly reformulated to give it more relevance to speech research. Furthermore, some extensions are proposed that can help in interpreting the results of identification experiments.

Standard (textbook) information theory has been applied to confusion matrices before (a famous example is Miller and Nicely, 1955; but see also e.g., Blom, 1970). Applying standard information theory on identification experiments with speech, results in a quite opaque description of the outcome which can be difficult to interpret and communicate. Information theory was developed to solve problems of data transmission and storage. The principal aim in this field is that of economy of coding: maximizing correct throughput and storage while minimizing channel and storage capacity. Therefore, the theory is strong on points like optimal coding schemes and

	u	o	ɔ	ɑ	a	ε	e	ɪ	i	y	℘	œ	T
u	**44**	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	66
o	*2*	**44**	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	66
ɔ	*2*	*2*	**44**	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	66
ɑ	*2*	*2*	*2*	**44**	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	66
a	*2*	*2*	*2*	*2*	**44**	*2*	*2*	*2*	*2*	*2*	*2*	*2*	66
ε	*2*	*2*	*2*	*2*	*2*	**44**	*2*	*2*	*2*	*2*	*2*	*2*	66
e	*2*	*2*	*2*	*2*	*2*	*2*	**44**	*2*	*2*	*2*	*2*	*2*	66
ɪ	*2*	*2*	*2*	*2*	*2*	*2*	*2*	**44**	*2*	*2*	*2*	*2*	66
i	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	**44**	*2*	*2*	*2*	66
y	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	**44**	*2*	*2*	66
℘	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	**44**	*2*	66
œ	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	*2*	**44**	66
T	66	66	6	66	66	66	66	66	66	66	66	66	792

$\varepsilon = 1/3$

$L = 2.071$

$G = 2.071$

$d_r = 11$

$d_S = 11$

Figure 1: A hypothetical confusion matrix for Dutch monophthongs with errors distributed evenly over all possible response categories. Error-rate = 33.3% (i.e., $\varepsilon = 1/3$), incorrect responses are indicated in italics. To obtain a normalized confusion matrix, divide all entries by 792. *T* indicates the row and column totals, for other symbols, see text.

methods for data compression and error correction. These areas are mostly outside the scope of those involved in speech recognition research. In general, when working with speech, one has fixed coding schemes (e.g., a language) and channel capacities (e.g., maximal throughput). When performing identification experiments, one is interested in the distribution of errors and the differences between experimental conditions. On these points, the standard theory of information is weak.

Another point is the fact that both the digital "size" of a symbol sequence (e.g., sequences of stimuli or responses) and its information content in bits, scale logarithmically with the size of the symbol inventory that is used to code it. For example, a binary (i.e., *two*-value) representation of a random sequence of *16* vowels has to be only 4 (and not 8) times as long as the original in order to code the same sequence. The standard logarithmic measure of information makes it difficult to get a feeling of how the performance of subjects scales with the size of the experimental inventory. A formulation of information theory that is linear in the size of the symbol inventory used would often be more convenient. This is especially urgent when one wants to measure the spread of errors in terms of the number of categories used. Below, I will discuss this problem and propose a solution.

This paper will discuss identification experiments. In such experiments, a number of distinct stimuli are presented to subjects who are asked to label them as belonging to one of the categories of a closed set (e.g., one of the vowels). It is important to note that only the sequences of stimulus and response *labels* will be used. Additional information that can be extracted from the stimulus *sounds* is outside the scope of this paper, i.e., only the features that are expressed as labels or symbols in the experiment are used. Other features, like the sex, age, or identity of the speaker (in a vowel identification experiment), are ignored. Furthermore, the theory has two levels. On the level of *information content*, it ignores the way stimuli are mapped to responses, as long as the map is consistent, i.e., there is no notion of correct or incorrect labels. However, all basic measures of information are sensitive to the absolute value of the error rate. Therefore, at the next level, the information content is *normalized* for the error rate itself. This two-step process enables the separation of the analysis of classification *ambiguities* and the analysis of identification *errors*.

The relation between the response categories and the stimulus categories enables the experimentator to infer the relative importance of different stimulus conditions.

	u	o	ɔ	ɑ	a	ɛ	e	ɪ	i	y	ø	œ	T
u	**42**		*8*					*8*				*8*	66
o		**42**	*8*					*8*				*8*	66
ɔ			**50**					*8*				*8*	66
ɑ			*8*	**42**				*8*				*8*	66
a			*8*		**42**			*8*				*8*	66
ɛ			*8*			**42**		*8*				*8*	66
e			*8*				**42**	*8*				*8*	66
ɪ			*8*					**50**				*8*	66
i			*8*					*8*	**42**			*8*	66
y			*8*					*8*		**50**		*8*	66
ø			*8*					*8*			**42**	*8*	66
œ			*8*					*8*				**42**	66
T	42	42	138	42	42	42	42	138	42	42	42	138	792

$\varepsilon = 1/3$

$L = 1.647$

$G = 1.402$

$d_r = 4.550$

$d_s = 2.733$

Figure 2: A hypothetical confusion matrix for Dutch monophthongs with a strong response bias for the mid-open vowels (i.e., /ɔ ɪ œ/). Error-rate = 33.3% (i.e., $\varepsilon = 1/3$), incorrect responses are indicated in italics. To obtain a normalized confusion matrix, divide all entries by 792. *T* indicates the row and column totals, for other symbols, see text.

This relation between stimulus and response is completely determined by the confusion matrix which charts for each combination of stimulus (frequency = s_i) and response (frequency = r_j), the frequency of combined occurrence (p_{ij}). In the remainder of the paper the normalized frequency of occurrence will be used as an empirical definition of the probability of occurrence and both terms will be freely interchanged. Although a frequency of occurrence is not identical to a probability, treating the difference between them is outside the scope of this paper.

This paper will first give an overview of standard information theory (sections 2-3) followed by a simple example which will introduce the notion of error dispersion (section 4). Then follows a section on the quantification of the differences between sequences (section 5) and a more rigorous definition of the error-dispersion (section 6). Finally, the theory is tested on several examples from the literature (section 7). Those who are not interested in the mathematical foundations of the error-dispersion and error-differences can skip the sections 2-6 and go straight to the summary at the start of section 7.

2 Quantifying information: Entropy

The mean information per token of an uncorrelated symbol sequence is generally quantified as the *Shanon's entropy* of the sequence. The mean entropy (H), and therefore the mean information content, of a signal with *I* different symbols, each with a probability p_i of occurring, is

$$H = \sum_{i=1}^{I} -p_i \cdot {}^2\log(p_i) \qquad (1a)$$

The mean entropy of a sequence (equation 1a) reaches its maximum when the p_i have a uniform distribution (i.e., $H = {}^2\log(I)$). This entropy is also the average amount of information (in bits) necessary to distinguish the tokens in the sequence. It describes the minimum amount of space per token, necessary to store the sequence, or the minimum channel-width necessary to transmit the sequence error-free.

In speech research, these quantities have limited use. Equation 1a is formulated in a way that is quite opaque to its meaning for the average identification experiment. The entropy, H, is connected to the number of symbols that are required to code the sequence. In principle, every code that generates sequences with the same (or a higher) entropy can be used to describe a given sequence. The minimum number of symbols, averaged per token, needed to describe a given sequence is 2^H:

$$2^H = 2^{\sum_{i=1}^{I} -p_i \cdot {}^2\log(p_i)} = \frac{1}{\prod_{i=1}^{I}\left(p_i^{p_i}\right)} \tag{1b}$$

This quantity, 2^H, which is completely equivalent to the entropy, will be called the *effective number of symbols*, or the *effective number of categories*. It describes the number of symbols out of which the sequence is constructed, weighted by their importance for determining the information content or entropy of the sequence. The word *effective* can be justified by the fact that 2^H can be used to predict what fraction of the tokens in the sequence are included in the symbols or classes with the highest (or lowest) frequency of occurrence. Define W_α as the minimum number of classes, out of a total of I classes (i.e., symbols), that, combined, contain a fraction α of the total number of tokens ($0 \le \alpha \le 1$). Then for $\alpha \le 0.5$ and $I \le 34$, it generally holds that

$$W_\alpha \le \alpha \cdot 2^H \le \alpha \cdot I$$

(no counter-example was found, a fractional number of "classes" is interpreted as a fraction of the *frequency of occurrence* of the least frequent class). In other words, a number of $\alpha \cdot 2^H$ of the highest frequency symbols is responsible for at least a fraction of α of all observed tokens. For example, if I=12, H=2.5, and α=0.5, then the $0.5 \cdot 2^{2.5} = 2.82 < 3$ classes with the highest frequency of occurrence are responsible for more than half of all tokens in the sequence (actually, the sum of the two highest frequencies of occurrence and 0.82 times the next lower frequency, together, are larger than or equal to 0.5). If either $\alpha > 0.5$ or $I > 34$, there are some lopsided distributions for which the relation does not hold (these distributions have $p_1 \sim 0.25$ and $p_{i>1} \sim 0.75/(I-1)$.). Smooth distributions of the form $p_i \sim i^r$ and $p_i \sim q^{(i^r)}$, with r any number and q>0, all appear to follow the relation $W_\alpha \le \alpha \cdot 2^H$ if $\alpha \le 0.5$, irrespective of the size of the inventory I.

In general, the effective number of stimulus and response categories will be close to the actual number of categories present. The usefulness of equation 1b will appear when the relation between stimulus and response sequences are discussed.

In an identification experiment there is a correspondence between stimulus and response sequences, expressed by the confusion matrix. The confusion matrix contains all information present in the results of an identification experiment. The confusion matrix can be considered to list the frequencies of occurrence of stimulus-response *combinations* or *pairs*. When properly normalized (i.e., $\Sigma s_j = \Sigma r_j = \Sigma p_{ij} = 1$ over all i,j), the mean entropy of the individual stimulus-response pairs of the experiment can be calculated from the confusion matrix. The mean information content of the confusion matrix is, by definition, identical to the mean entropy of the sequence of stimulus-response pairs that resulted from the experiment. The entropy of the confusion matrix H_{CM} is expressed in equation 2.

$$H_{CM} = \sum_{i=1}^{I} \sum_{j=1}^{J} -p_{ij} \cdot {}^{2}\log\left(p_{ij}\right) \qquad (2)$$

With: I the number of stimulus classes, J the number of response categories, and p_{ij} the probability of response j to stimulus i.

The maximum value for the entropy of the confusion matrix (H_{cm}) can be calculated by assuming that stimulus and response are independent, i.e., $p_{ij} = s_i \cdot r_j$ (using equation 2):

$$\begin{aligned} H_{CM} &= \sum_{i=1}^{I} \sum_{j=1}^{J} -s_i \cdot r_j \cdot {}^{2}\log\left(s_i \cdot r_j\right) \\ &= \sum_{j=1}^{J} r_j \cdot \sum_{i=1}^{I} -s_i \cdot {}^{2}\log\left(s_i\right) + \sum_{i=1}^{I} s_i \cdot \sum_{j=1}^{J} -r_j \cdot {}^{2}\log\left(r_j\right) \\ &= \sum_{i=1}^{I} -s_i \cdot {}^{2}\log\left(s_i\right) + \sum_{j=1}^{J} -r_j \cdot {}^{2}\log\left(r_j\right) \\ &= H_{Stim} + H_{Resp} \end{aligned} \qquad (3)$$

With: I the number of stimulus categories, J the number of response categories, s_i the probability of stimulus i, r_j that of response j, and p_{ij} the probability of response j to stimulus i. H_{Stim} and H_{Resp} are the entropies of stimulus and responses.

When stimulus and response are independent, then the entropy of the confusion matrix is the exact sum of the entropy of the stimulus and the response sequences. When the distribution of both stimulus and response are uniform (maximum entropy), then the maximum entropy of the confusion matrix is ${}^{2}\log(I)+{}^{2}\log(J)$ (for I different stimulus and J different response categories).

The minimum entropy is obtained with perfect recognition, when $p_{i=j} = s_i$ and $p_{i \neq j} = 0$ (we assume that $0 \cdot {}^{2}\log(0)=0$). Then equation 2 reduces to equation 1 for the stimulus entropy. In the case of perfect recognition, both the entropy of the response sequence and the entropy of the confusion matrix are equal to the entropy of the stimulus sequence, i.e., $H_{CM} = H_{Stim} = H_{Resp}$.

The entropies are related because the stimulus and response sequences can both be obtained from the confusion matrix. It always holds that: $\max(H_{Stim}, H_{Resp}) \leq H_{CM} \leq H_{Stim} + H_{Resp}$.

3 Information transmitted to and lost from the responses

The mean entropy, technically the total amount of information present in stimulus and response sequences, is generally not the desired quantity. It contains information that is independent of the stimulus sequence, e.g. response biases and random "errors". What is important is the amount of information from the stimulus sequence used (or not) by the listeners, as can be extracted from the response sequence. The mean amount (or rate) of information transmitted from stimulus to response (T, in bits per token) can be stated as the difference between the maximum entropy possible and the actual entropy present in the confusion matrix, i.e., mean information transmitted is:

	u	o	ɔ	ɑ	a	ɛ	e	ɪ	i	y	ø	œ	T
u	**42**	*24*											66
o		**42**	*24*										66
ɔ			**42**	*24*									66
ɑ				**42**	*24*								66
a					**42**	*24*							66
ɛ						**42**	*24*						66
e							**42**	*24*					66
ɪ								**42**	*4*				66
i									**66**				66
y									*24*	**42**			66
ø							*24*				**42**		66
œ											*24*	**42**	66
T	42	66	66	66	66	66	90	66	114	42	66	42	792

$\varepsilon = 1/3$

$L = 0.928$

$G = 0.867$

$d_r = 1.019$

$d_s = 0.899$

Figure 3a: A hypothetical confusion matrix for Dutch monophthongs with a response shift towards vowels with a *higher* F_2 target. With respect to Figure 3b, $\delta_r = 0.859$ and $\delta_s = 1.014$. Error-rate = 33.3% (i.e., $\varepsilon = 1/3$), incorrect responses are indicated in italics. To obtain a normalized confusion matrix, divide all entries by 792. *T* indicates the row and column totals, for other symbols, see text.

$$T = H_{Stim} + H_{Resp} - H_{CM} \qquad \text{(T is the rate of transmission)} \qquad (4a)$$

This difference is the mean entropy of the response sequence that is already accounted for in entropy of the stimulus sequence.

The amount of information lost from the stimuli is the mean entropy of the confusion matrix minus the mean entropy of the response sequence, i.e.:

$$L = H_{CM} - H_{Resp} \qquad \text{(L is the rate of transmission \textit{loss})} \qquad (4b)$$

This is the mean entropy in the stimulus sequence, not accounted for in the response sequence, i.e., the entropy of the stimuli given the responses or H(stimulus|response). The actual meaning of L is much clearer when equation 4b is written in terms of the effective number of categories in the response sequence and confusion matrix (equation 1b):

$$2^L = \frac{2^{H_{CM}}}{2^{H_{Resp}}} = \frac{\text{Effective number of Stimulus} - \text{Response pairs}}{\text{Effective number of Response categories}} \qquad (4b')$$

Equation 4b' shows that the rate of information loss can be interpreted as the logarithm of the mean (effective) number of stimulus categories that can induce each response. It can also be called the logarithm of the *perplexity* of the stimuli with respect to the responses.

The mean, subject-related, "information" added to the responses (or *gained*) as a result of the errors and biases in the responses, is equal to the mean entropy of the confusion matrix minus the mean entropy of the stimulus sequence, i.e.:

$$G = H_{CM} - H_{Stim} \qquad \text{(G is the rate of transmission \textit{gain})} \qquad (4c)$$

This is the mean entropy in the response sequence, not accounted for by the stimulus sequence, i.e., the entropy of the responses given the stimuli or H(response|stimulus). Equivalent to the information loss, L (equation 4b'), the information gain, G, can be interpreted as the logarithm of the mean (effective) number of response categories for

	u	o	ɔ	ɑ	a	ɛ	e	ɪ	i	y	ø	œ	T
u	66												66
o	*24*	42											66
ɔ		*24*	42										66
ɑ			*24*	42									66
a				*24*	42								66
ɛ						42					*24*		66
e						*24*	42						66
ɪ								42				*24*	66
i								*24*	42				66
y	*24*									42			66
ø											42	*24*	66
œ			*24*									42	66
T	124	66	90	66	42	66	42	66	42	42	66	90	792

$\varepsilon = 1/3$

$L = 0.944$

$G = 0.867$

$d_r = 1.054$

$d_s = 0.899$

Figure 3b: A hypothetical confusion matrix with a response shift towards vowels with a *lower* F_2 target. With respect to Figure 3a, $\delta_r = 0.859$ and $\delta_s = 1.014$. Error-rate = 33.3% (i.e., $\varepsilon = 1/3$), incorrect responses are indicated in italics. To obtain a normalized confusion matrix, divide all entries by 792. *T* indicates the row and column totals, for other symbols, see text.

each stimulus or the logarithm of the *perplexity* of the responses with respect to the stimuli.

The effect the *distribution* of the errors has on the values of L and G are demonstrated in the (hypothetical) examples of Figures 1-4 which all share an error rate of 1/3. In general, the larger the number error categories in the confusion matrix is, the higher L and G will be. The more *uneven* the distribution of responses and stimuli, the higher L respectively G will be (compare L and G in Figures 2-3). However, L and G are also sensitive to the absolute error rates. In Tables 1 and 2 at the end of this paper, pairs of confusion matrices with different error rates and equivalent distributions of the errors do show quit large differences in the sizes of the transmission loss and gain.

Equations 4a-c can be combined to obtain:

$$H_{Stim} = T + L \qquad (4d)$$

$$H_{Resp} = T + G \qquad (4e)$$

$$L = G + H_{Stim} - H_{Resp} \qquad (4f)$$

In equation 4b, L measures the amount of stimulus-information that was lost from the responses, i.e., the degree to which the stimulus sequence can*not* be reconstructed from the response sequence. In equation 4c, G measures the amount of stimulus-independent entropy that was added to the responses by the subjects, i.e., the degree to which the response sequence can*not* be predicted from the stimuli. Note that T decreases and both L and G increase when the error rate increases (cf. Table 1 and 2 at the end of this paper).

Obviously, the rate of transmission, T, depends on the amount of information in the stimulus, i.e., the number of input classes. It is easy to see that, for a typical identification experiment:

$$H_{Stim} \sim H_{Resp} \sim H_{CM} \sim {}^{2}\log(\text{Number of Input Classes}).$$

This means that all entropies scale with the logarithm of the number of input classes. As a consequence (equation 4a) the transmission rate, T, also scales with the logarithm of this number. However, because the entropies of both the confusion matrix and the stimulus and response sequences scale with the same number (equation 4b), the rate of transmission *loss*, L, and *gain*, G, do not scale with the number of input classes. They only depend on the error-rate and the pattern of the errors. This is also obvious from the fact that L and G can be described in terms of the number of response categories per stimulus category or vice versa (equation 4b'). In general, any change in the stimulus entropy will be compensated by an equivalent change in the response entropy.

A small example will explain this principle. Consider a vowel identification experiment with 5 tokens, /y i I e ε/ and

$$H_{Stim} \approx H_{Resp} \approx {}^{2}\log(5).$$

In a typical experiment we will see $H_{CM} \geq {}^{2}\log(5)$. Now we extend this experiment and add the tokens /u o ɔ o a/ after which we see a comparable pattern of errors. We now expect that, in a first approximation, the new entropies will be,

$$H'_{Stim} \approx H'_{Resp} \approx {}^{2}\log(10) \text{ and } H'_{CM} \geq {}^{2}\log(10),$$

i.e.,

$$H' \approx H + {}^{2}\log(2)$$

if the new tokens "behave" like the old ones. Now the new rate of transmission is

$$\begin{aligned} T &= H'_{Stim} + H'_{Resp} - H'_{CM} \\ &\approx H_{Stim} + {}^{2}\log(2) + H_{Resp} + {}^{2}\log(2) - H_{CM} - {}^{2}\log(2) \end{aligned}$$

and therefore,

$$T' \approx T + {}^{2}\log(2).$$

That is, the rate of transmission scales as the entropies themselves. However, the loss of information is

$$L' = H'_{CM} - H'_{Resp} \approx H_{CM} + {}^{2}\log(2) - H_{Resp} - {}^{2}\log(2) = H_{CM} - H_{Res}$$

and, therefore, in a first approximation, we expect that $L' \approx L$, i.e., the loss of information does *not* scale as the entropies themselves. The same holds for the gain of information G.

4 An example

The theory described in the preceding sections will be illustrated with a simple, hypothetical example. In this example two assumptions will be made: First, the correct responses are evenly distributed over the stimuli, i.e., each stimulus has the same error-rate. Second, the errors are evenly distributed over a limited and fixed number of response labels (c.f., Figures 1, 3a-b), i.e., for each stimulus the errors are limited to some of the available response categories which all get the same number of responses.

The general error-rate will be indicated by ε, the rate of correct responses by (1-ε). The number of erroneous labels used for each stimulus, also called the dispersion of the errors, is d_s (e.g., if $d_s = 2$, an /ɔ/ type stimulus could be confused with /o o/ with equal probabilities). For example, $d_s = 11$ in Figure 1. The value of d_s would have been equal to 1 in Figures 3a and 3b if the /i/, respectively the /u/ stimuli would have been omitted.

Under the above assumptions, equation 2 reduces to (assume $p_{i=j} = (1\text{-}\varepsilon)\cdot s_i$; for each i: d_s responses have $p_{i\neq j} = \varepsilon/d_s \cdot s_i$; and the others have $p_{i\neq j} = 0$):

$$\begin{aligned}
H_{CM} &= \sum_{i=1}^{I} -d_s \cdot s_i \cdot \frac{\varepsilon}{d_s}\,{}^2\log\left(s_i \cdot \frac{\varepsilon}{d_s}\right) - s_i \cdot (1-\varepsilon)\,{}^2\log\left(s_i \cdot (1-\varepsilon)\right) \\
&= \sum_{i=1}^{I} -s_i \cdot \varepsilon \cdot \left({}^2\log(s_i) + {}^2\log\left(\frac{\varepsilon}{d_s}\right)\right) - s_i \cdot (1-\varepsilon) \cdot \left({}^2\log(s_i) + {}^2\log(1-\varepsilon)\right) \\
&= \sum_{i=1}^{I} -s_i \cdot (\varepsilon + (1-\varepsilon))\,{}^2\log(s_i) + s_i \cdot \varepsilon \cdot {}^2\log\left(\frac{\varepsilon}{d_s}\right) - s_i \cdot (1-\varepsilon)\,{}^2\log(1-\varepsilon) \\
&= \sum_{i=1}^{I} -s_i \cdot {}^2\log(s_i) - \left(\varepsilon \cdot {}^2\log\left(\frac{\varepsilon}{d_s}\right) + (1-\varepsilon)\,{}^2\log(1-\varepsilon)\right) \cdot \sum_{i=1}^{I} s_i \qquad \text{(5a)} \\
&= H_{Stim} - (1-\varepsilon)\,{}^2\log(1-\varepsilon) - \varepsilon \cdot {}^2\log\left(\frac{\varepsilon}{d_s}\right)
\end{aligned}$$

Thus:

$$\begin{aligned}
G &= H_{CM} - H_{Stim} \\
&= H_{Stim} - H_{Stim} - (1-\varepsilon)\,{}^2\log(1-\varepsilon) - \varepsilon \cdot {}^2\log\left(\frac{\varepsilon}{d_s}\right) \\
&= -(1-\varepsilon)\,{}^2\log(1-\varepsilon) - \varepsilon \cdot {}^2\log\left(\frac{\varepsilon}{d_s}\right) \qquad \text{(5b)}
\end{aligned}$$

$$L = H_{Stim} - H_{Resp} - (1-\varepsilon)\,{}^2\log(1-\varepsilon) - \varepsilon \cdot {}^2\log\left(\frac{\varepsilon}{d_s}\right)$$

With: I the number of stimuli, s_i the probability of stimulus i. ε is the error-rate and d_s the number of erroneous response categories per stimulus or the error-dispersion. Equation 4 was used to derive 5b.

The amount of information lost and gained due to the errors (ε) in the limit of $d_s = 1$ is (i.e., only a single error category for each stimulus):

$$L \approx G = -(1\text{-}\varepsilon){}^2\log(1\text{-}\varepsilon) - \varepsilon{}^2\log(\varepsilon) = H_\varepsilon$$

This quantity is called the "entropy of the error rate" (H_ε).

The approximation of the model used to derive equation 5a-b is only a crude one (below we will derive the relevant formulas for the general case). We will

	u	o:	ɔ	ɑ	a:	ɛ	e:	ɪ	i	y	ø:	œ	T
u	66												66
o:		33	*33*										66
ɔ		*33*	33										66
ɑ				33	*33*								66
a:				*33*	33								66
ɛ						66							66
e:							33	*33*					66
ɪ							*33*	33					66
i									66				66
y										66			66
ø:											33	*33*	66
œ											*33*	33	66
T	66	66	66	66	66	66	66	66	66	66	66	66	792

$\varepsilon = 1/3$

$L = 0.667$

$G = 0.667$

$d_{\varepsilon} = 0.593$

$d_S = 0.593$

Figure 4: A hypothetical confusion matrix with a complete loss of the *long-short* distinction for Dutch monophthongs. Error-rate = 33.3% (i.e., $\varepsilon = 1/3$). incorrect responses are indicated in italics. To obtain a normalized confusion matrix, divide all entries by 792. *T* indicates the row and column totals, for other symbols, see text.

demonstrate this "crudeness" by comparing an exact calculation with the approximation of equation 5a-b.

Assume that under certain conditions, long and short vowels cannot be distinguished (see figure 4). In Dutch there are 12 monophthongal vowels, 8 of which are part of long/short vowel pairs. In a balanced presentation with otherwise perfect identification the loss of the long/short distinction introduces theoretically 33.3% errors (4 out of 12 on average). According to equation 5b this amounts to a gain of information of G = 0.915 bits ($d_s = 1$).

We can calculate the exact gain of information from equation 2 (see figure 4). Both stimulus and response sequences have a uniform distribution and therefore, both the stimulus and response entropy are

$$H_{Stim}=H_{Resp}={}^2\log(12) \text{ bit.}$$

Four unpaired short vowels contribute a total of (66/792 = 1/12):

$$4\cdot(-1/12\cdot{}^2\log(1/12)) = 4/12\cdot{}^2\log(12)$$

from 4 cells in the confusion matrix. The eight confused vowels from long/short pairs contribute (33/792 = 1/24):

$$16\cdot(-1/24\cdot{}^2\log(1/2\cdot 1/12)) = 16/24\cdot({}^2\log(12)+{}^2\log(2))$$

from 16 cells in the confusion matrix (unbiased long/short confusion). This adds up to:

$$H_{cm} = {}^2\log(12)+2/3\cdot{}^2\log(2) = {}^2\log(12)+2/3 \text{ bit}$$

for the confusion matrix. Therefore, the information gained or lost due to the confusion is:

$$L = G = H_{cm} - H_{Stim} = 2/3 \approx 0.667 \text{ bit.}$$

A gain or loss of information of this size would correspond to a "random" error-rate of 18% if we use equation 5b (with d_s=1).

In the above examples with long/short confusions we see that the approximation of equation 5 over-estimates the loss of information when the errors are not distributed uniformly. Now equation 5b can be used to express the gain and loss of information, G and L, in terms of an effective number of error categories. In general, the results of an identification experiment will show a variable confusion between response categories. Neither the gain, G, nor the loss of information, L, is very transparent as to the pattern of errors and both depend on the size of the error rate. The number of response categories, "d_s" in equation 5, over which the errors are evenly distributed is a measure of the spread of the errors in the confusion matrix and is independent of the *size* of the error rate. This number, d_s, can be calculated from the error-rate, ε, and the gain (or loss) of information, G, as given in equation 5c:

$$d_s = \frac{\varepsilon}{1-\varepsilon}\cdot 2^{\frac{G+{}^2\log(1-\varepsilon)}{\varepsilon}} = \frac{\varepsilon}{1-\varepsilon}\cdot 2^{\frac{L+{}^2\log(1-\varepsilon)}{\varepsilon}} \cdot \frac{2^{\frac{H_{Resp}}{\varepsilon}}}{2^{\frac{H_{Stim}}{\varepsilon}}} \tag{5c}$$

$$d_s \approx \frac{\varepsilon}{1-\frac{\varepsilon}{2}}\cdot\frac{2^{\frac{G}{\varepsilon}}}{e} = \frac{\varepsilon}{1-\frac{\varepsilon}{2}}\cdot\frac{2^{\frac{L}{\varepsilon}}}{e}\cdot\frac{2^{\frac{H_{Resp}}{\varepsilon}}}{2^{\frac{H_{Stim}}{\varepsilon}}} \tag{5d}$$

In which d_s is the effective number of error categories, ε the error-rate, G the gain and L the loss of information and e the base of the natural logarithm (e ≈ 2.71828...). The approximation of 5d is quite accurate, to within 1% if ε≤0.4 and to within 10% if ε≤0.8.

Using 5c on the result of our example with a complete loss of the long short distinctions (figure 4), i.e., ε = 1/3 (33%) and

L = G = 2/3 (bits),

results in

$d_s = 0.593$.

The fact that d_s<1 illustrates the extreme clustering of the errors in the confusion matrix in this imaginary experiment. In general, d_s in equation 5 can be interpreted as the effective number of error classes per stimulus token.

5 The difference between confusion matrices

Generally, it will not be sufficient to compare the entropies between two experiments. It is often also important to know whether there is a real difference between the distributions of the responses. When listeners in two experiments use the same information, and respond in identical ways, then combining the responses from these two experiments should not change the entropy of the confusion matrix, at least after differences in stimulus and response sequences between both experiments are discounted.

It can be shown that the entropy of a superposition of N sequences, H_{pooled}, is always larger than or equal to the mean entropy of the sequences, $\overline{H}$. More specific, H_{pooled} equals:

$$H_{pooled} = \sum_{k=1}^{N} \alpha_k \cdot H_k + \eta \cdot \sum_{k=1}^{N} \alpha_k \cdot {}^2\log(\alpha_k) \tag{6a}$$

or

$$H_{pooled} = \overline{H} + \eta \cdot H_\alpha \tag{6b}$$

with

$$0 \leq \eta \leq 1,\ \sum_{k=1}^{N} \alpha_k = 1,\ \overline{H} = \sum_{k=1}^{N} \alpha_k \cdot H_k \text{, and } H_\alpha = \sum_{k=1}^{N} -\alpha_k \cdot {}^2\log \alpha_k$$

If $\alpha_k = 1/N$ then $H_\alpha = {}^2\log N$.

This means that the entropy of the combined sequences or confusion matrices equals the (weighted) mean of the individual entropies plus a fraction, $\eta \cdot H_\alpha$. Here η is the fraction of the tokens of the pooled sequence that can still be attributed (plausibly) to one of the original source sequences (e.g., confusions that are found in one experiment but not in the others).

If all sequences are realizations of the same "process", i.e., $p_{ki} \approx p_{pooled,i}$ then $\eta \approx 0$. When the tokens of all sequences exclude each other (i.e., $p_{ki} \cdot p_{li} = 0$ for *all* i and $k \neq l$) then $\eta = 1$. In all cases, η will be at least as large as the combined contributions of tokens that are unique to each experiment, i.e., $\eta \geq \sum p_{pooled,i}$ summed over all i for which $p_{ki} \neq 0$ for only a single k.

The parameters η_{Stim}, η_{Resp}, and η_{CM} describe the results from combining respectively, the stimulus sequences, response sequences, and confusion matrices of different "experiments". For these, the same relations hold as for the corresponding entropies:

$$\max(\eta_{Stim}, \eta_{Resp}) \leq \eta_{CM} \leq \eta_{Stim} + \eta_{Resp}.$$

For the confusion matrix, the parameter η_{CM} is related to the error-rates of the experiments. As the correct entries in a confusion matrix do not normally differ between experiments, η_{CM} measures the differences between the *errors* in the experiments. If the *correct* responses of the experiments overlap, then $\eta_{CM} \leq \varepsilon_{pooled}$, with the maximum reached when all errors in the individual experiments are different.

When the confusion matrix of the (virtual) "combined" experiment is calculated, the mean information transmitted, lost and gained from stimulus to response becomes:

$$\begin{aligned} T_{pooled} &= \overline{T} + \left(\eta_{Stim} + \eta_{Resp} - \eta_{CM}\right) \cdot H_\alpha \\ &= \overline{T} + \left(\eta_{Stim} - \lambda\right) \cdot H_\alpha \end{aligned} \tag{7a}$$

$$L_{pooled} = \overline{L} + \left(\eta_{CM} - \eta_{Resp}\right) \cdot H_{\alpha} \\ = \overline{L} + \lambda \cdot H_{\alpha} \qquad (7b)$$

$$G_{pooled} = \overline{G} + \left(\eta_{CM} - \eta_{Stim}\right) \cdot H_{\alpha} \\ = \overline{G} + \gamma \cdot H_{\alpha} \qquad (7b)$$

In which: $\lambda = \eta_{CM} - \eta_{Resp}$ and $\gamma = \eta_{CM} - \eta_{Stim}$, $0 \leq \lambda, \gamma \leq 1$, and $\overline{T}$, $\overline{L}$, and $\overline{G}$ are the weighted averages of the transmission rate, information loss, and gain, corresponding to $\overline{H}$ in equation 6.

This implies that the pooled information loss is the (weighted) mean information loss from the individual experiments plus a factor (λ) that includes differences in response (η_{Resp}) biases and differences in the way the subjects handled the stimuli in the two experiments (η_{CM}). The combined result of these fractions, λ, is related to the error-rate of one experiment with respect to the others, in a sense it is a weighted sum of the fraction of "errors" that are unique to each experiment.

For normal experiments, in which the stimulus sequences are generally matched when experiments are compared (i.e. $\eta_{Stimulus}=0$), λ can be understood as the information about the differences between confusion matrices of the experiments that cannot be deduced from the corresponding response sequences alone. For any set of experiments, if $\lambda<\eta_{Stim}$, the experiments are to a large extent incomparable. If $\lambda>\eta_{Stim}$, the experiments can be compared, but they are not identical and information is lost by pooling them (i.e., mixing all responses).

6 Error dispersion

The gain and loss of information (G and L) in an identification experiment can, in general, be attributed to the identification errors. The effect of errors on the size of the information gain and loss can be split into two "independent" contributions: The *size* of the error-rate itself, ε (e.g., Tables 1 and 2 below), and the *distribution* of the errors (e.g., Figures 1-4). All other things being equal, if the error-rate *increases* or the errors are distributed over *more* confusions, then the gain and loss of information will also *increase*. This dependence of G and L from the error-rate makes these quantities not well suited to describe the effects of the *dispersion* of the errors. The relation between G, L, and the error rate can be formalized by splitting the confusion matrix into a diagonal matrix with only the *correct* responses and an off-diagonal matrix with only the *in*correct responses. The corresponding entropies are called $H_{Correct}$ and H_{Error}. The entropy of the original confusion matrix H_{CM} can be expressed as a combination of the correct and incorrect responses and becomes (using equation 6, with $\eta = 1$, $\alpha_1 = 1-\varepsilon$, $\alpha_2 = \varepsilon$ and $H_{\alpha} = H_{\varepsilon}$):

$$H_{CM} = (1-\varepsilon) \cdot H_{Correct} + \varepsilon \cdot H_{Error} + H_{\varepsilon} \qquad 8a$$

In which: H_{CM} is the entropy of the confusion matrix, ε is the error rate, and $H_{Correct}$ and H_{Error} are the entropies of, respectively, the diagonal and off-diagonal (correct and incorrect responses) sub-matrices of the confusion matrix. The entropy of the error rate is $H_{\varepsilon} = -(1-\varepsilon)^2\log(1-\varepsilon) - \varepsilon^2\log(\varepsilon)$

In what follows we will concentrate on the information loss, L. Equations with the information gain, G, instead of the information loss can be derived using equation 4f.

The information loss, L, can be expressed as a combination of correct and incorrect responses using equation 4b:

$$\begin{aligned} L &= H_{CM} - H_{Resp} \\ &= (1-\varepsilon)\cdot\left(H_{Correct} - H_{Resp}\right) + \varepsilon\cdot\left(H_{Error} - H_{Resp}\right) + H_{\varepsilon} \end{aligned} \qquad 8b$$

As equation 8a, but with: L is the information loss and H_{Resp} is the entropy of the response sequence

In equation 8b we can see that the information loss L is split into two contributions, a contribution that depends only on the entropy of the error rate ($H_{\varepsilon} = -(1-\varepsilon)\cdot{}^{2}\log(1-\varepsilon) - \varepsilon\cdot{}^{2}\log\varepsilon$) and a contribution that also depends on the distribution of responses over the confusion matrix.

The "total" responses will be dominated by the correct responses. As a result H_{Resp} and $H_{Correct}$ will be almost equal and $(1-\varepsilon)\cdot(H_{Correct}-H_{Resp})\approx 0$. This means that the part of equation 8b that depends on the error distribution will be dominated by the factor $\varepsilon\cdot(H_{Error}-H_{Resp})$. The part between the brackets is independent of the error rate itself, it only depends on the distribution of the errors. It can loosely be interpreted as a "mean error entropy per response category". Note that the factor $(H_{Error}-H_{Resp})$ has the form of equation 4b, i.e. of the information loss L. However, it is *not* the information loss of the off-diagonal (error) matrix because it uses the response entropy of the complete *confusion matrix*, instead of the response entropy of the off-diagonal (error) matrix.

Define:

$$\begin{aligned} \ell_d &\equiv \frac{(1-\varepsilon)\cdot\left(H_{Correct} - H_{Resp}\right) + \varepsilon\cdot\left(H_{Error} - H_{Resp}\right)}{\varepsilon} \\ &\approx H_{Error} - H_{Resp} \end{aligned} \qquad 9a$$

In which: ε is the error rate, H_{Resp} is the entropy of the response sequence, and $H_{Correct}$ and H_{Error} are the entropies of, respectively, the diagonal and off-diagonal (correct and incorrect responses) sub-matrices of the confusion matrix. ℓ_d measures the effect of response errors on the loss of information. (the corresponding gain entity is called g_d)

The quantity ℓ_d is determined by the distribution of the incorrect responses. This is used to define the error-dispersion d_r (using equation 8b):

$$\begin{aligned} L &= \varepsilon\cdot\ell_d + H_{\varepsilon} \\ \ell_d &= \frac{L - H_{\varepsilon}}{\varepsilon} \\ d_r &= 2^{\ell_d} = \frac{2^{\left(\frac{L}{\varepsilon}\right)}}{2^{\left(\frac{H_{\varepsilon}}{\varepsilon}\right)}} = \frac{\varepsilon}{1-\varepsilon}\cdot 2^{\left(\frac{L+{}^{2}\log(1-\varepsilon)}{\varepsilon}\right)} \end{aligned} \qquad 9b$$

As equation 9a but with: L is the information loss rate, d_r is the error dispersion in effective error categories per response category, $H_{\varepsilon} = -\varepsilon\cdot{}^{2}\log(\varepsilon) - (1-\varepsilon)\cdot{}^{2}\log(1-\varepsilon)$ is the "entropy" of the error-rate

When equation 9b is compared to equation 5b (section 4), then it appears that ℓ_d is the excess information loss due to more than one error category per response (L-H_ε), averaged over the errors (i.e., $1/\varepsilon$). The last line of equation 9b is identical to equation 5c and can be approximated by equation 5d (using L instead of G).

$$d_r \approx \frac{\varepsilon}{1-\frac{\varepsilon}{2}} \cdot \frac{2^{\frac{L}{\varepsilon}}}{e} \qquad \text{(9c, see 5d)}$$

In which d_r isthe effective number of error categories per response category, ε the error-rate, L the loss of information and e the base of the natural logarithm (e = 2.71828...). The approximation of 5d is quite accurate, to within 1% if $\varepsilon \leq 0.4$ and to within 10% if $\varepsilon \leq 0.8$.

In equation 9b we see that the error-dispersion per response category, d_r, calculated from equation 8a is identical to the number of error categories per stimulus category, d_s, in the model calculations of equation 5, but now using L instead of G. This indicates that the error dispersion calculated according to equation 8 and 9 can indeed be interpreted as an effective number of error categories per response category (d_r, using L) or stimulus category (d_s, using G). This can be shown more explicitly by rewriting d_r using equation 9a and b, in:

$$\ell_d = \frac{(1-\varepsilon)\cdot\left(H_{Correct} - H_{Resp}\right) + \varepsilon\cdot\left(H_{Error} - H_{Resp}\right)}{\varepsilon}$$

$$d_r \equiv 2^{\ell_d} = \left(\frac{2^{H_{Correct}}}{2^{H_{Resp}}}\right)^{\frac{(1-\varepsilon)}{\varepsilon}} \cdot \frac{2^{H_{Error}}}{2^{H_{Resp}}} \qquad 10$$

Equation 10 implies that d_r equals the effective number of *correct* stimulus categories per response category times the number of *in*correct stimulus categories per response category, normalized for the error rate. Correspondingly, d_s is the (normalized) effective number of *correct* response categories per stimulus category times the number of *in*correct response categories per stimulus category. Both products measure the simultaneous dispersion of the errors along the stimulus and response "direction", and both are maximal if all stimuli and responses have the same error rate. Note that neither the stimuli, nor the responses are required to have one and only one correct category.

It should be emphasized that the model used to derive the equations 5a-c in section 4 is only an approximation for any real identification experiment. In contrast, equations 8-9 do *not* depend on approximations. The value of the error dispersion, d_r or d_s, is an exact measure of the contribution of the error distribution to the information loss or gain relative to the contribution of the errors *an sich*.

In Figures 1-4 and Tables 1 and 2, examples are given of d_r and d_s for some hypothetical and real data. In the Figures 1-4, there are 12 evenly distributed stimuli and, respectively, 132, 33, 11, and 8 error categories in the confusion matrix. The corresponding values of d_s would then be expected to be close to 11, 2.75, 0.917, and 0.667, respectively (i.e., number of confusions / number of stimuli). The actual values are 11, 2.73, 0.899, and 0.593. The increasing discrepancy between the expected and the actual values going from figures 1-4 is due to the increasingly uneven distribution of the *correct* responses. An uneven distribution of correct responses reduces the error dispersion. The effect of the distribution of the responses on the error dispersion can be seen when d_r is compared to d_s in Figure 2 and to a lesser extend in Figures 3a and

b. In these figures, the stimuli are distributed evenly, but the responses are not. In these examples this means that the entropy of the response sequences is less than that of the stimulus sequences. This results in markedly higher values for d_r (c.f., equation 10). If the entropy of the response sequences is higher than that of the stimulus sequences, then $d_r < d_s$ (e.g., there are more response possibilities than stimuli).

When confusion matrices are combined, i.e., experimental results are pooled, an expression for the new error-dispersion of the pooled results can be derived using equations 7b and 9b.

$$L_{pooled} = \sum_{k=1}^{N} (\alpha_k \cdot L_k) + \lambda \cdot H_\alpha \qquad \text{(11a, cf. 7b)}$$

$$\varepsilon_{pooled} \cdot \ell_{pooled} + H_{\varepsilon_{pooled}} = \sum_{k=1}^{N} \left(\alpha_k \cdot \varepsilon_k \cdot \ell_k + \alpha_k \cdot H_{\varepsilon_k}\right) + \lambda \cdot H_\alpha \qquad \text{(11b, cf. 9b)}$$

Define:

$$\delta_r = \frac{\lambda}{\varepsilon_{pooled}} - \frac{H_{\varepsilon_{pooled}} - \overline{H_\varepsilon}}{\varepsilon_{pooled} \cdot H_\alpha}$$

$$\delta_s = \frac{\gamma}{\varepsilon_{pooled}} - \frac{H_{\varepsilon_{pooled}} - \overline{H_\varepsilon}}{\varepsilon_{pooled} \cdot H_\alpha}$$

Then (substituting $\varphi_k = \alpha_k \cdot \varepsilon_k / \varepsilon_{pooled}$):

$$\ell_{pooled} = \sum_{k=1}^{N} (\varphi_k \cdot \ell_k) + \delta_r \cdot H_\alpha \qquad \text{(11c)}$$

$$d_{r_{pooled}} = 2^{\delta_r \cdot H_\alpha} \cdot \prod_{k=1}^{N} \left(d_{r_k}\right)^{\varphi_k} \propto N^{\delta_r} \cdot \left(\prod_{k=1}^{N} d_{r_k}\right)^{\frac{1}{N}} \qquad \text{(11d)}$$

Correspondingly:

$$d_{s_{pooled}} = 2^{\delta_s \cdot H_\alpha} \cdot \prod_{k=1}^{N} \left(d_{s_k}\right)^{\varphi_k} \propto N^{\delta_s} \cdot \left(\prod_{k=1}^{N} d_{s_k}\right)^{\frac{1}{N}} \qquad \text{(11e)}$$

with:

$$\varphi_k = \frac{\alpha_k \cdot \varepsilon_k}{\varepsilon_{pooled}}, \quad \overline{H_\varepsilon} = \sum_{k=1}^{N} \left(\alpha_k \cdot H_{\varepsilon_k}\right).$$

The proportionality of 11d is exact when $\alpha_k = 1/N$ and $\varepsilon_k = \varepsilon_{pooled}$, i.e., when all matrices are summed unweighted and all error-rates are equal.

From equation 11 it can be concluded that the error-dispersion of the combined confusion matrices (d_{pooled}) can be interpreted as the *geometrical* average of the individual error-dispersions (d_k), weighted by their contribution to the total error-rate (φ), times the number of matrices N (i.e., "experiments") to the power of

$$\delta_r = \left[\lambda - \left(H_{\varepsilon_{\text{pooled}}} - \overline{H}_{\varepsilon}\right) / H_{\alpha}\right] / \varepsilon_{\text{pooled}}$$

or

$$\delta_s = \left[\gamma - \left(H_{\varepsilon_{\text{pooled}}} - \overline{H}_{\varepsilon}\right) / H_{\alpha}\right] / \varepsilon_{\text{pooled}}.$$

The only factor in equation 11d that is independent of the error-rate is δ (which is "normalized" with respect to ε). This δ indicates the effective fraction of the pooled errors for which the error classes do not overlap between confusion matrices (cf. the interpretation of λ in section 5).

Figures 3a and 3b give an example of how the differences between confusion matrices influence the value of δ. The errors in both confusion matrices do not overlap, there is not a single shared confusion. As a result, $\delta_s \approx 1$, i.e., *all* errors are different. The value of δ_s is even somewhat larger than 1 due to a small difference between the *correct* responses in both confusion matrices. However, the value of $\delta_r < 1$ because the *responses* in both confusion matrices do not overlap completely (as do the stimuli). Any differences between the response sequences are always discounted from the value of δ_r. The same holds with respect to differences between stimulus sequences and the value of δ_s.

7 Examples: Error dispersion in vowel identification experiments*

The error rate does not describe the distribution of the errors in the confusion matrix. It is possible to give a quantitative description of the distribution of the errors using standard information theory. The information content, i.e., the entropy, of a confusion matrix (CM) can be described with three numbers, respectively, the entropies of the stimulus sequence (H_{Stim}), the response sequence (H_{Resp}), and the confusion matrix itself (H_{CM}):

$$H_{CM} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} -P_{ij} \cdot {}^{2}\log\left(P_{ij}\right)}{N} - {}^{2}\log(N) \qquad (12a)$$

$$H_{Stim} = \frac{\sum_{i=1}^{I} -S_i \cdot {}^{2}\log\left(S_i\right)}{N} - {}^{2}\log(N) \qquad (12b)$$

$$H_{Resp} = \frac{\sum_{j=1}^{J} -R_j \cdot {}^{2}\log\left(R_j\right)}{N} - {}^{2}\log(N) \qquad (12c)$$

*Readers can test their own examples on the World Wide Web URL: http://fonsg3.let.uva.nl:8001/Service/ErrorDispersion.html

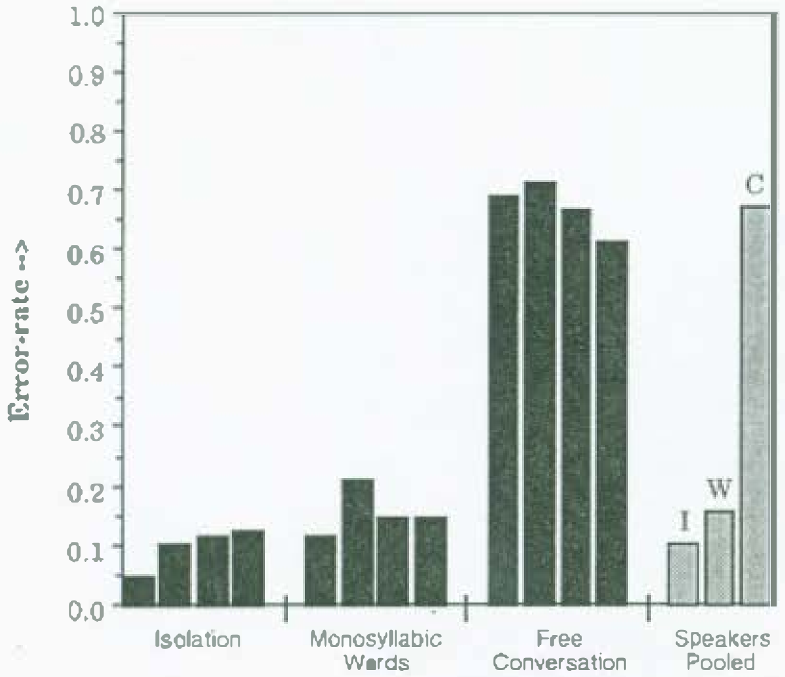

Figure 5: Error rate for vowels from four speakers from Koopmans-van Beinum (1980). Vowels were uttered in isolation (I), in monosyllabic, isolated words (W), and in unstressed syllables in free conversation (C). The filled bars indicate the results for the individual speakers, the grey bars for all speakers pooled. The error rate for vowels uttered in free conversation differs statistically significant from those uttered in isolation and monosyllabic words (Wilcoxon test, $p \leq 0.05$, two-tailed). The error rates of the latter two do not differ statistically significant from each other.

In equation 12a-e, S_i is the number of stimuli in class i, R_j the number of responses in class j, P_{ij} the number of stimuli in class i that elicited responses of class j, and N the total number of stimulus–response pairs.

Several measures using the information content of the confusion matrix and stimulus and response sequences are in use (c.f., Miller and Nicely, 1955), e.g.,

$T = H_{Stim} + H_{Resp} - H_{CM}$,

T/H_{Stim},

$L = H_{CM} - H_{Resp}$,

$G = H_{CM} - H_{Stim}$.

All these measures are sensitive to the absolute error rate, ε. However, it is possible to "normalize" the L and G measures in such a way that they become insensitive to the absolute error rate. When these normalized measures are transformed from the usual

Table 1: Confusion matrix and performance measures for identification of vowel realizations presented with and without their natural context. Taken from Kuwabara (1985)

#V# presentation ($\varepsilon = 20.9\%$)
$L = 0.825$, $d_r = 1.329$
$G = 0.814$, $d_c = 1.281$

VVV presentation ($\varepsilon = 3.8\%$)
$L = 0.271$, $d_r = 1.300$ $\delta_r = 0.068$
$G = 0.273$, $d_c = 1.311$ $\delta_c = 0.134$

	i	e	Λ	o	u	i	e	A	o	u	total
i	600	200				776	24				800
e	96	512	6	13	13	26	601		13		640
A		144	562	14				713		7	720
o			45	392	123	6		11	515	28	560
u				14	466				5	475	480
total	696	856	613	433	602	808	625	724	533	510	3200

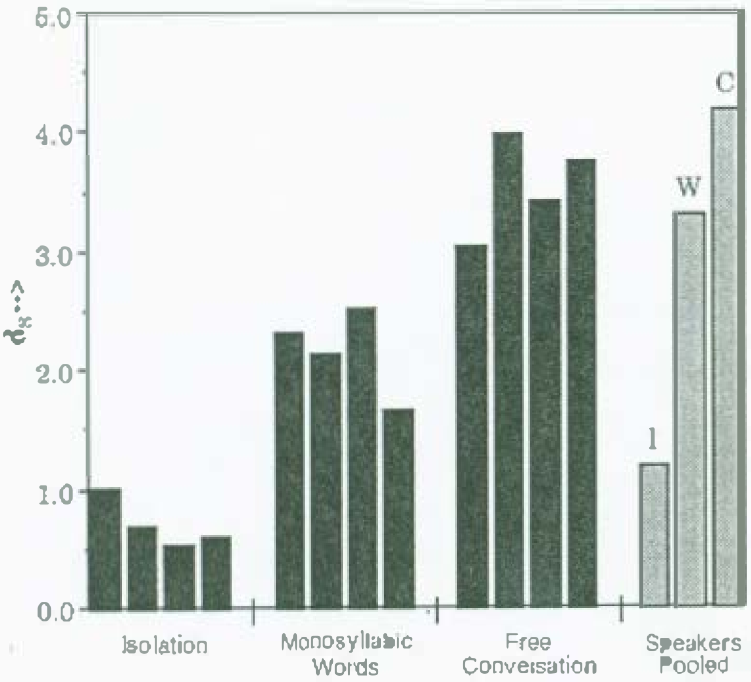

Figure 6: Error dispersion for vowels from four speakers from Koopmans-van Beinum (1980). Vowels were uttered in isolation (I), in monosyllabic, isolated words (W), and in unstressed syllables in free conversation (C). The filled bars indicate the results for the individual speakers, the grey bars for all speakers pooled. The differences between the error dispersions are statistically significant for all conditions (Wilcoxon test, p≤0.05, two-tailed).

logarithmic form to the linear form, the resulting measures represent the "effective number of error classes", or *error-dispersion*. In the previous sections, I proposed two measures of error dispersion, d_s and d_r. When ε is the error-rate, and $H_\varepsilon = -\varepsilon \cdot {}^2\log(\varepsilon) - (1-\varepsilon) \cdot {}^2\log(1-\varepsilon)$, then:

$$d_s = \frac{2^{\left(\frac{H_{[illegible]} H_{[illegible]}}{\varepsilon}\right)}}{2^{\left(\frac{H_\varepsilon}{\varepsilon}\right)}} \qquad (13a)$$

is the effective number of error categories per stimulus and

Table 2: Confusion matrix and performance measures for identification of vowel realizations presented with and without their natural context. Taken from Huang (1991)

	#V# presentation ($\varepsilon = 29.1\%$) $L = 1.223$, $G = 1.243$,	$d_r = 2.312$ $d_s = 2.426$				CVC presentation ($\varepsilon = 20.7\%$) $L = 0.965$, $G = 0990$,	$d_r = 2.167$ $d_s = 2.350$		$\delta_r = 0.019$ $\delta_s = 0.023$		
	i	ɪ	e	ɛ	ʌ	i	ɪ	e	ɛ	ʌ	total
i	706	100	33	3	18	764	65	23	1	7	860
ɪ	10	693	14	58	165	17	756	2	62	103	940
e	51	150	834	132	33	47	107	885	135	26	1200
ɛ	3	52	45	542	243	1	22	20	640	202	885
ʌ	2	33	4	175	446	4	24	0	71	561	660
total	772	1028	930	910	905	833	974	930	909	899	4545

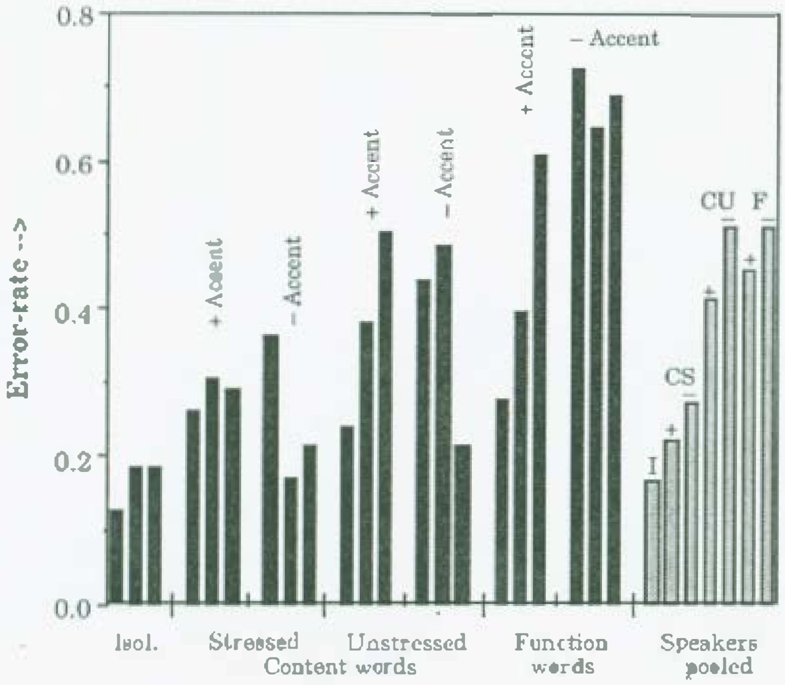

Figure 7: Error rates for vowels from three speakers from Van Bergem (1993). Vowel realizations were taken from monosyllabic words uttered in isolation (I), in stressed (CS) and unstressed (CU) syllables of content words, and in function words (F). Syllables were uttered with and without sentence accent (+/- Accent see text). The filled bars indicate the results for the individual speakers, the grey bars for all speakers pooled. The difference in error rates between vowels taken from stressed syllables in content words (CS) and function words (F) differ statistically significantly (ignoring +/–Accent, Wilcoxon test, p≤0.05, two-tailed). The other differences are not statistically significant. The presence or absence of sentence accent had no effect.

$$d_r = \frac{2^{\left(\frac{H_{CM}-H_{Resp}}{\varepsilon}\right)}}{2^{\left(\frac{H_{\varepsilon}}{\varepsilon}\right)}} \tag{13b}$$

is the effective number of error categories per response. In Figure 1-4, examples are presented of hypothetical confusion matrices. All confusion matrices have the *same* error rate of 1/3, but different *distributions* of the errors over the stimuli and responses. As a result of the differences in the distribution of the errors, the error dispersions are different. Whenever the distribution of *stimuli* and *responses* are different, the values of d_s and d_r are different for a single confusion matrix (c.f. Figures 2 and 3).

There also is a measure, δ, that describes differences in the error categories *between* confusion matrices. In the following the symbol $\overline{H}$ indicates the *mean* entropy of the N experiments. Furthermore, H' indicates the entropy and ε' the error-rate of the N *pooled* experiments (i.e., the combined confusion matrices). Furthermore, it is assumed that all confusion matrices are weighted equal (if not, use equation 11). The error-difference of the responses, δ_r, and stimuli, δ_s, are defined as:

$$\delta_r = \frac{\left(H'_{CM} - \overline{H}_{CM}\right) - \left(H'_{Resp} - \overline{H}_{Resp}\right)}{\varepsilon' \qquad H_\alpha} - \frac{\left(H'_{\varepsilon} - \overline{H}_{\varepsilon}\right)}{\varepsilon'^{2}\log(N)} \tag{14a}$$

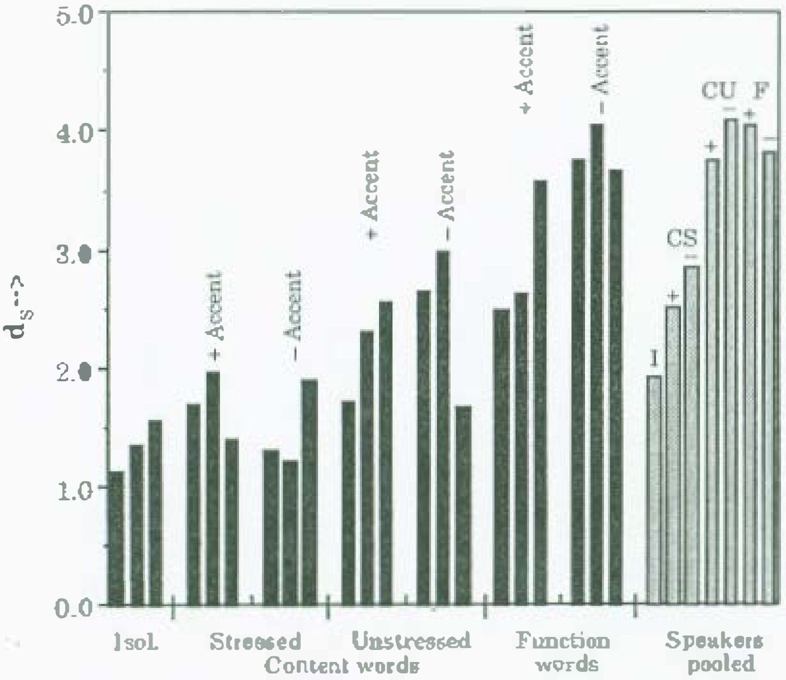

Figure 8: Error dispersions for vowels from three speakers from Van Bergem (1993). Vowel realizations were taken from monosyllabic words uttered in isolation (I), in stressed (CS) and unstressed (CU) syllables of content words, and in function words (F). Syllables were uttered with and without sentence accent (+/- Accent see text). The filled bars indicate the results for the individual speakers, the grey bars for all speakers pooled. Except for the results for vowels uttered in isolation (I), are all differences between syllable types are statistically significant (ignoring +/–Accent, Wilcoxon test, p≤0.05, two-tailed). The presence or absence of sentence accent had no effect.

$$\delta_S = \frac{\left(H'_{CM} - \overline{H}_{CM}\right) - \left(H'_{Stim} - \overline{H}_{Stim}\right)}{\varepsilon'} - \frac{\left(H'_{\varepsilon} - \overline{H}_{\varepsilon}\right)}{\varepsilon'^2 \log(N)} \qquad (14b)$$

H_α

The error-differences can be interpreted as the effective fraction of the errors that is outside the shared error categories, corrected for overall differences in the distribution of stimuli or responses. An example is given in Figures 3a and 3b. All confusions are different between these two confusion matrices, so $\delta_S \approx 1$ (note that the stimulus distributions are identical). Due to the differences between the distributions of the *responses*, which are discounted, the value of $\delta_T < 1$.

The measures, discussed above, have been calculated for data from papers concerning vowel identification. Papers concerning two questions in the field of vowel identification are discussed. The first question is about the effects of the presence of context on vowel identification. The second question is about the effects of stress and (informal) speaking style on vowel intelligibility.

Presenting vowel segments in their "natural" context enhances there intelligibility. Is this the result of a change in the *kind* of confusions that appear, i.e., the listeners "home in" on the correct vowel and there are less possible responses for each stimulus? Or is it the result of the listeners being better able to distinguish between ambiguous pairs, i.e., the confusions are the same, but the listeners can better select the correct response?

These questions were touched in the studies of Kuwabara (1985) and Huang (1991, 1992) which will be discussed here. They used vowel realizations excised from trisyllabic VVV sequences (Kuwabara, 1985) and CVC syllables (Huang, 1991, 1992) and presented them in isolation as well as in their original, syllabic, context.

Kuwabara (1985) used trisyllabic VVV sequences excised from short Japanese sequences. The vowels could be any one of the five Japanese vowels /i e ɑ o ɯ/ but the medial (target) vowel was always different from the two flanking vowel. Four subjects were asked to label the realizations (medial vowel only) as one of the five Japanese vowels. Each token was presented five times. Presented in isolation, the error-rate was 20%. Presented in context, the error-rate was only 4%. There were an equal number of stimulus/response pairs in both experiments.
From the confusion matrices in Table 1 it can be calculated that the error dispersions for both conditions are almost equal, $d \approx 1.3$, and the differences are small, $\delta \leq 0.13$. Thus, despite the enormous differences in the error *rate*, the *distribution* of the errors seems to be fairly similar, whether the medial vowels were presented with or without context. This means that the difference between the presentation of vowels with and without context is not so much the type of confusions, but the extent to which they are reported.

Huang (1991) presented consonant-vowel-consonant syllables to subjects as well as the excised vowels from these syllables (i.e., without the consonants). The results from four speakers were pooled in the confusion matrices presented in Table 2. Using the "raw" error-rates from the individual speakers it is possible to show that there is a statistically significant difference between the error-rates from presentation in context and presentation without context (Wilcoxon test, $p \leq 0.05$, two tailed). No statistical significant difference can be found for the error dispersions.

Again, from the confusion matrices it can be calculated that the error dispersions for vowels presented with and without context are almost equal, $d \approx 2.3$ and the differences small, $\delta \approx 0.02$. Therefore, it seems that the presence of context does help in recognition, but not by reducing the number of *different* confusions, but by reducing the number of times *each confusion* leads to an incorrect answer.

This inference, which seems obvious when inspecting the error-dispersion and difference, is difficult to justify by other means. A visual inspection of the confusion matrices could hint towards the role of context in these experiments, but this role would be hard to quantify.

With regard to the second question, about the influence of stress and speaking style on vowel intelligibility, two studies are discussed, one by Koopmans-van Beinum (1980) and one by Van Bergem (1993).

Koopmans-van Beinum (1980) presented vowel realizations from four speakers to listeners. She used vowels uttered in isolation, from monosyllabic words uttered in isolation and from unstressed syllables from free conversation. The resulting error rates and error dispersions are presented in figures 5 and 6.

Van Bergem (1993) presented vowels taken from identical syllables pronounced in isolation, as the stressed and unstressed syllables of content words and as a function word. All words, except those pronounced in isolation, where part of carrier sentences. The three speakers were unaware of which word was the target word. Sentences were structured to place a sentence accent on or next to the target syllable (unstressed syllables and function words cannot carry sentence accent) and, alternatively, the sentence accent was placed away from the target syllable. For each syllable there were 7 realizations for each of the three speakers. The resulting error rates and error dispersions are presented in figures 7 and 8.

For both experiments, the error *dispersion* can separate the conditions better than the error *rate*. Using the error dispersion, there are statistically significant differences between all conditions (Wilcoxon test, $p \leq 0.05$, two-tailed). The only exception are the Isolated vowels of Van Bergem (1993), mostly because there were only three values available. Using only the error rate, it is not possible to decide, based on *these* data, that there is a difference in the identification of vowels caused by speaking style or word stress/syllable type. But by using the error dispersion, it *can* be inferred that changes in stress and speaking style change the *number* of ambiguities as well as the overall intelligibility.

8 Conclusions

From the examples given it can be seen that the error dispersion is an independent measure of the spreading of the errors over individual stimulus and response categories. It is possible to use the error dispersion to distinguish response patterns in conditions where the error rate does not distinguish them, just as it is possible to spot equivalencies when the error rate indicates only distinctions.

Acknowledgements

I would like to thank Dick van Bergem and Florien Koopmans-van Beinum for making available the data of their identification experiments. I would also like to thank Louis ten Bosch and Louis Pols for their helpful comments.

References

Blom, J.G. (1970). 'The application of information theory to vowel recognition experiments', *Proceedings of the Sixth International Congress of Phonetic Sciences*, edited by B. Mála. M. Romportl. P. Janora, Prague 1967, 189-196.

Huang, C.B. (1991): 'An acoustic and perceptual study of vowel formant trajectories in American English', Ph.D. Thesis, Massachusetts Institute of Technology, USA (Research Laboratories of Electronics, Technical report no. 563, Cambridge, MA).

Huang, C.B. (1992): 'Modelling human vowel identification using aspects of formant trajectory and context' in *Speech perception, production and linguistic structure*, edited by Y. Tohkura. E. Vatikiotis-Bateson & Y. Sagisaka (Ohmsha, Tokyo; IOS Press, Amsterdam), 43-61.

Khinchin, A.I. (1957). *Mathematical foundation of information theory*, translated by R.A. Silverman and M.D. Friedman, Dover Publications Inc., New York NY, pp120.

Koopmans-van Beinum, F.J. (1980): 'Vowel contrast reduction, an acoustic and perceptual study of Dutch vowels in various speech conditions', Ph.D. Thesis, University of Amsterdam, The Netherlands, pp163.

Kuwabara, H. (1985): 'An approach to normalization of coarticulation effects for vowels in connected speech', *Journal of the Acoustical Society of America* 77, 686-694.

Miller, G.A., Nicely, P.E. (1955). 'An analysis of perceptual confusions among some English consonants', *Journal of the Acoustical Society of America* 27 (2), 338-352.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P. (1988). *Numerical recipes in C*, Cambridge University Press, Cambridge MA, second edition 1992, 632-635.

Sveshnikov, A.A. ed. (1968). *Problems in probability theory, mathematical statistics and theory of random functions*, W.B. Saunders Company Philadelphia, 157-170.

Van Bergem, D.R. (1993). 'Acoustic vowel reduction as a function of sentence accent, word stress, and word class', *Speech Communication* 12, 1-23.