OBJECTIVE PREDICATES OF WORD INTELLIGIBILITY

Chris J. James
Hearing Health Care Research Unit
Department of Communicative Disorders
Elborn College, University of Western Ontario
London, Ontario, Canada, N6G 1H1.

ABSTRACT

There have been many attempts in the past to systematically attribute the subjective properties of speech to objective and physical characteristics. The aim of this study was to look in more detail at how objective properties interact to contribute to the shape of intelligibility versus presentation level functions for individual words. The interaction of such features as speech level, familiarity and spectral consistency were found to be complex and enlightening. Spectral consistency was measured using the variance of spectral flatness σ and the length of words. These two measures were a novel addition to the standard measures. Although the results shown by no means explain all the variation of speech intelligibility they do provide some insight into the play of factors for the fairly simple case of mono-syllabic words.

SOMMAIRE

Jusqu'à ce jour, de nombreux efforts ont été déployés afin d'attribuer de façon systématique les propriétés subjectives de la parole à des caractéristiques objectives et physiques. Le but de cette étude était de scruter plus en détails comment les propriétés objectives contribuent à la forme d'intélligibilité versus les fonctions du niveau de présentation pour les mots isolés. L'interaction de ces caractéristiques telles que le niveau de la parole, la familiarité et l'uniformité spectrale s'est avérée complexe et révélatrice. L'uniformité spectrale a été mesurée en utilisant la variance de l'égalité spectrale σ et la longueur des mots. Ces deux mesures représentent un ajout innovateur aux mesures classiques. Même si les résultats présentés ne permettent pas d'expliquer toutes les variations de l'intélligibilité de la parole, ils fournissent quelques indices sur le rôle des facteurs pour le cas relativement simple des mots mono-syllabiques.

1. INTRODUCTION

In this paper we seek to assess the relative importance of objective speech measurements in predicting subjective word thresholds and word intelligibility. We use the twelve Boothroyd (1968a and 1968b) lists of words specified by Markides (1978) as our material. Some of the measurements described in this paper were obtained in other studies, the specific sources are described in the text. Here we seek to see how the difficulty of perception of an individual word versus level of presentation is related to properties such as familiarity, energy (speech level) and frequency characteristics.

The analysis in this paper takes the form of correlation of subjective or behavioral data with objective properties. The subjective data was collected in a previous study aimed to identify a subset of words from the Boothroyd lists to be used in an adaptive test (James, 1992b). Some details of the nature of these data are presented herein. The choice and treatment of the objective measurements is based on previous reports in the literature and some further intuitive reasoning by the author. In this case the objective measurements are treated as indicators of the amount of information given to the listener for correct identification of the target word. There are two basic aims of this work: The first is to assess the relative importance of physical

measurements and linguistic measurement and how best to treat these properties. The second is to find good objective predictors of the intelligibility characteristics of new speech material.

2. SPEECH MEASUREMENTS

Before embarking on any exploratory statistical analysis, it is useful to discuss the relative importance of the various objective factors in determining the subjective results. We can divide the objective measurements described below into three distinct *types*: those which describe level, spectral characteristics and linguistic usage. We might expect any measure of a certain *type* to be interrelated with other measures of the same *type* and independent of measures of a different *type*. However, it may be deduced that some data of apparently different *types* may be interrelated, for example speech level and mean spectral flatness (see below) are likely to be related measures.

In this study we are relating thresholds of stimulus words (the level at which a word is on average 50% intelligible) and the intelligibility of words at given presentation levels. The relative importance of objective *types* in predicting these subjective characteristics may be inferred by considering the effect of presentation level: It is fairly obvious that small differences in speech signal level will be of little importance at supra-threshold levels. Alternatively, small differences such as these may increase or decrease the level of thresholds, perhaps by the actual difference in speech level. (It is noted here that the manner in which the recording levels were set on the source tape will contribute to measured levels of the words, however these levels were preserved in the subjective study and in the measurements presented here.)

From previous studies (for example, James et al., 1992a), we know that certain Boothroyd words are more robust to degradation than others. This would seem to indicate that certain frequency characteristics are more robust than others. For example, fricative consonants of low intensity will be less audible compared to say the liquids that occur in diphthongs (such as "veil" and "fail."). In these words there are characteristic shifts in the vowel which forms the part of the speech spectrum with the highest energy. There might also be other clusters/structures which prove more robust, or alternatively more fragile to changes in presentation level. We can also state that the degree of consistency of any particular feature will lend to its perception. Here we might expect some interaction of steady state spectral characteristics with their duration and in a limited sense with the word's total duration. It is also true that shorter words will, in general, be of lower total energy.

Finally, we speculate about the importance of familiarity in the perception of speech. This has a history of study, for example Broadbent (1967) and Morton (1969), showing that more frequently used words are of higher intelligibility. Wayland et al. (1989) in addition show the effect of limiting the amount of information available to subjects by gating the test words. We may conclude a similar result for speech level: The further above threshold we get the more acoustic information is available. At higher levels, the effects of familiarity are reduced because the cohort of confusable words is reduced.

2.1. Speech Material

The speech material used in this study was the selection of twelve Boothroyd (1968a and 1968b) lists of ten mono-syllabic words, recordings made by ISVR, Southampton. For the purposes of analysis, the words were digitally transcribed and stored in a 12-bit digital format on a Masscomp 5450 computer. The speech data was in a form which enabled use of Audlab signal analysis software and enabled speech to be replayed via a reconstruction filter (Kemo vbf/22) into any external measurement device such as a measuring amplifier or Sound Level Meter (SLM). In this study the relative signal levels to the calibration tone on the original test tape are preserved both for subjective presentation level calibration and for speech level measurements..

2.2. Speech Level

Speech level measurements were obtained in two ways: By calculation from the digitised word samples stored on the Masscomp computer, and by replaying the words via a reconstruction filter set at 8000 Hz cutoff frequency into a Brüel and Kjær SLM (2204), via an in line pre-amplifier (Brüel and Kjær UA0196). Using calculation and an SLM various measures of speech level for a given sample were obtained:

The root mean square power (by calculation) P_{rms}.

$$P_{rms} = \sqrt{\left(\frac{\sum_{n} P_n^2}{n} \right)}$$

Where P_n are calculated from the square of the voltage of the speech signal.

The mean square power level (by calculation) L_{Prms}.

$$L_{Prms} = \log(P_{rms})$$

The total energy of the speech sample (by calculation) P_{tot}.

$$P_{tot} = \sum_n P_n$$

The total energy level (by calculation) L_{Ptot}.

$$L_{Ptot} = \log(P_{tot})$$

The "A" weighted "Impulse" level (SLM) L_{Aimp}.

The "A" weighted "Fast" level (SLM) L_{Afast}.

2.3.   Spectral Analysis

Spectral analysis was in the form of the calculation of normalised Spectral Flatness from the digitised speech samples. More detailed information on the characteristics of particular sections of speech may be obtained using spectral analysis, particularly obtained via the FFT. The spectrum of speech obtained from fourier transform gives us the power-frequency distribution of the signal averaged over a frame of time. Some work has been done on the relation of spectral shape to perception previously by Dubno and Levitt (1981) and more recently by Lee and Dermody (1992) for segments of speech. However these have studied the properties of the discriminability of speech sounds, and the results are not easily translated into the complete word context. Thus here we aim to measure how the signal changes along its duration (see Figure 1). Much of the acoustic information in speech originates from changes in the spectrum and thus it is an important thing to measure. The time-amplitude variation of the various frequency components in speech signals is invariably complicated. We could, perhaps, go into great detail on these variations, however there will be great differences between the frequency characteristics of different articulations of the same word on a microscopic scale, both between speakers and by the same speaker. It is therefore useful to look at the general shape and variation of speech spectrum.

One measure which lends itself to calculation is called Spectral Flatness σ (Jayant and Noll, 1984), this gives us a single measure of the spectral shape of the signal. σ is obtained from a single spectrum and in this case is normalised, a value of 1 would indicate a flat or white noise spectrum and a value of 0 a completely random spectrum.

Spectral Flatness, σ, is defined as the uniformity of the frequency distribution of the signal thus:

$$\sigma_{t-t_0} = \frac{\exp\left(\sum_n [\ln(P_n)]/n\right)}{\sum_n (P_n)/n}$$

Where P_n is the mean power over the frequency interval n, in the duration $t-t_0$.

With this measure we could use the whole duration of the speech sample or use spectral flatness as indicator of change of spectrum by looking at the flatness between successive frames. The latter is more useful if we wish to look at the variation or consistency of the spectrum of the signal.

An example of spectral flatness analysis is given in Figure 1. The top graph shows the time-amplitude variation for the word "cheese", the lower graph shows the corresponding spectral flatness plotted against time frame. We can see the three distinct regions representing "ch", "ee" and "zz", the first consonant, vowel and final consonant. The most dominant feature is the low spectral flatness over the duration of the vowel, this is due to the strong harmonic structure of the vowel sound. The regions representing the consonants produce much smaller dips in the flatness curve, these are due to the shaped noise characteristics of "ch" and "zz". It is interesting to note that vowel sounds are of low spectral flatness and high energy.

The normalised spectral flatness was calculated for frames of 12.8 ms at 3.2 ms intervals using a version of the Audlab "fft" program (sfm) modified by the author. Thus, between about 160 and 200 data points of spectral flatness were obtained for each word. The mean and variance of these data were then calculated to give the parameters σ_{mean} and σ_{var} for each word. With this approach it was possible to get a measure of both the average shape, and consistency of the shape of the spectrum of individual words. No transformation was applied to σ in the calculations since σ_{mean} and σ_{var} were largely independent (Howell, 1982).

2.4.   Duration

The duration t_f for each word was obtained in terms of the number of frames used for each sample in the Spectral Flatness analysis (above).
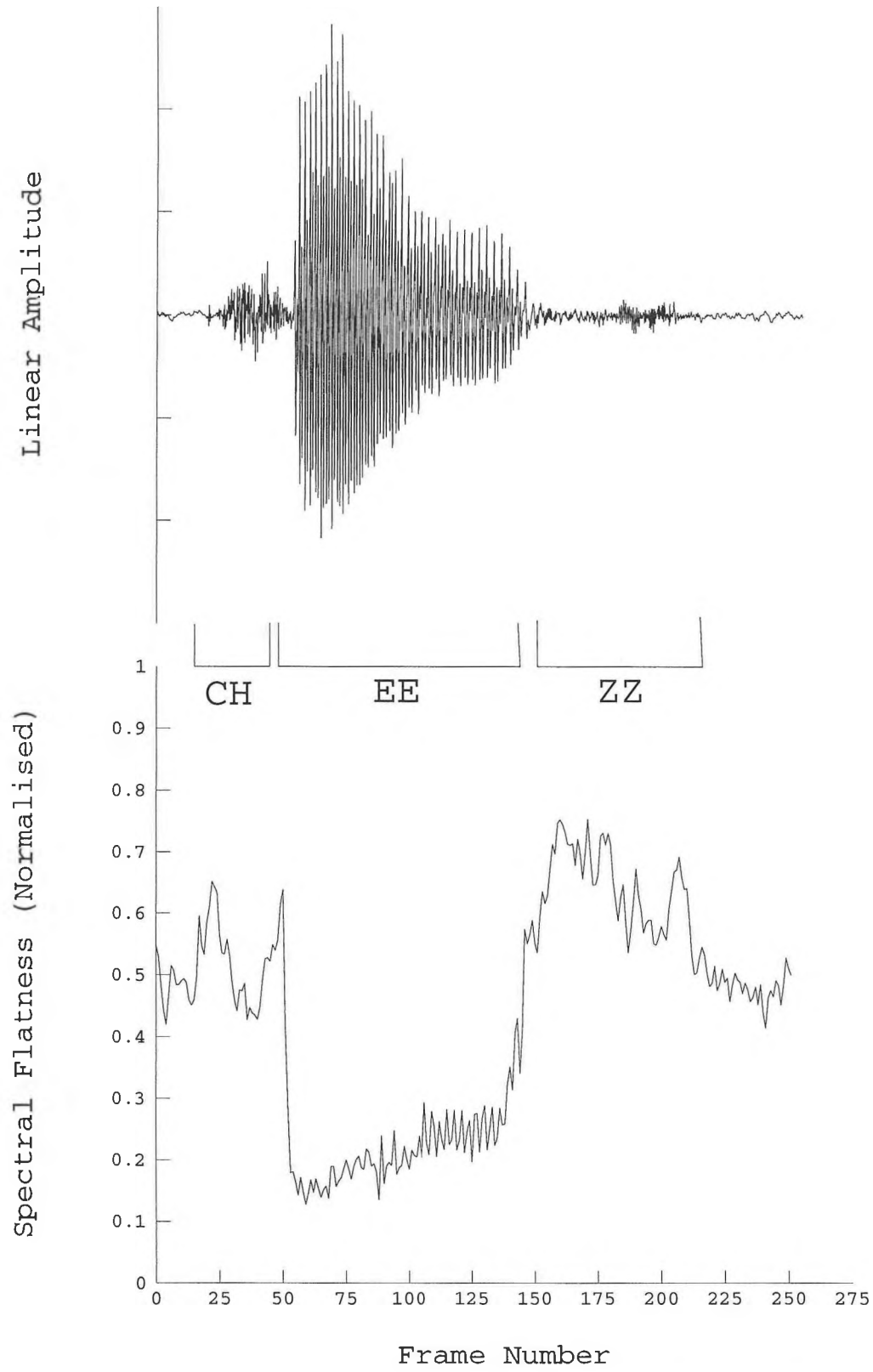
FIG. 1. Spectral flatness σ versus frame for the word "cheese" (lower plot). The upper plot is the corresponding time-domain waveform. Frame length 12.8 ms, frame shift 3.2 ms.

2.5.   Familiarity/Frequency of Usage

In most studies of the relationship of word frequency to intelligibility, word counts have been collected from written sources (for example, Rosenzweig and Postman. 1957, Broadbent, 1967, Lyregaard, 1976, and Hood and Pool, 1980) such as those produced by Thorndike and Lorge (1944) and more recently by Francis and Kučera (1982). In this study frequency of usage data was obtained from teletext subtitles (James, 1991), this was data was used because it represents spoken English. Data were collected on the usage of not only the target words, words in the Boothroyd lists, but also for words which were confusable with the target words. Morton's (1969) logogen model supports the idea of "competing" responses, and so some indication of relative familiarity of the target word to other candidates would be useful. The word usage data used here were treated in several ways:

Absolute count of word usage U.

The logarithm of absolute count of word usage $LogU$, as applied by Broadbent, and Lyregaard.

The ratio V_{c1+c2}, where:

$$V_{c1+c2} = \frac{U}{\sum_{c1} U + \sum_{c2} U}$$

i.e. the ratio of the usage of the target word to the sum of usage of the words in the cohort. In this case the response set was defined by all those words which either front rhyme ($c2$ confusion, as in cat and cap) or end rhyme ($c1$ confusion, as in they and lay).

The logarithm of the ratio V_{c1+c2}, $LogV_{c1+c2}$.

The ratio V_{c1c2}, where:

$$V_{c1c2} = \frac{U}{\sum_{c1c2} U}$$

here the response set $c1c2$ is defined by any words which have a common vowel to the target word (Taken from the dictionary as specified in James 1991).

The logarithm of V_{c1c2}, $LogV_{c1c2}$.

2.6.   Word Intelligibility

Seventeen subjects with no history of hearing impairment and between the ages of twenty one and twenty six, were chosen from the local University population. A 6 dB decrement/increment two-by-two paradigm was used to adaptively control the presentation level of 120 Boothroyd words (James, 1992b): The presentation level for the next two words is based on the results from the previous two such that two errors produce a step up in presentation level, one error no change, and no errors a step down. The order of presentation of words was randomised for each subject. The presentation rate was unpaced, a new word only being initiated after the subject's response to the previous one had been recorded. The presentation of each word was cued by a tone followed by a short gap. The subjects were asked to repeat each word as they heard it and told that the words would vary in loudness. Even if they heard only part of a word, or a word that did not make sense, or even a single sound, they were asked to re-iterate it. The response u for each word was recorded as correct (1) or incorrect (-1) by the experimenter. Thus the presentation levels of words are scattered around some overall threshold for each subject.

The data from this subjective experiment are expressed here in terms of Robustness Indices for each word, for several reasons:

> "Intelligibility" versus presentation level functions are to be generated for individual words as opposed to some fractional scoring of a list of items. One can only score a response as correct or as an error for a given item.

> The above being the case it seems appropriate to include some measure of the uncertainty of the result (variance) in the expression of the intelligibility function (see below definition of R_L).

> Data points u obtained for a word at a given level L are used in obtaining Robustness at other levels. This deviates from the conventional way in which "intelligibility" is obtained. However, this makes more efficient use of the available raw data provided that conditions for the results u are met as specified below.

Robustness Indices R_L were calculated for each word as below for a range of presentation levels L:

$$R_L = \frac{\sum_{k=1}^{n} u_L(k)}{\sum_{k=1}^{n} (u_L(k) - NM_L)}$$

The denominator in is rounded up to 0.01 for all values less than 0.01 for the purposes of computation. The "normalised mean" NM_L is defined:

$$NM_L = \frac{\sum_{k=1}^{n} u_L(k)}{n}$$

and u_L are defined for subject-word results u:

$$\textit{If } u(k) = -1 \quad \textit{and} \quad L \leq L(k) \quad \Rightarrow u_L(k) = -1$$
$$\textit{else if } u(k) = 1 \quad \textit{and} \quad L \geq L(k) \quad \Rightarrow u_L(k) = 1.$$

Note: If neither of the conditions in above is satisfied then the data point is not used (and is not included in n).

Thus

$$-\infty < R_L < +\infty$$

For example, in future trials

$$R_L > 0 \Rightarrow p(u_L^+) = 1,$$
$$R_L < 0 \Rightarrow p(u_L^+) = 0,$$

and

$$R_L \equiv 0 \Rightarrow p(u_L^+) = \tfrac{1}{2}.$$

Where $p(u_L^+)$ is the probability of a positive outcome for a subsequent trial with presentation level L. We can plot the "Robustness" R_L versus presentation level L function for each word.

Thus, the more positive the value of R_L the more intelligible the word at level L. A zero value of R_L indicates 50% intelligibility (i.e. either outcome, right or wrong, is indeterminate), and a large and negative R_L low intelligibility.

The Robustness Index R versus presentation level L function for each word was described in a variety of ways. A threshold is obtained at the level at which the Robustness Index is zero, that is where future outcomes have zero predictability. Due to the nature of the functions, graph modelling is applied to the data to obtain measures such as threshold and slope of the graph. Two methods were employed: A straight line fit with intercept and gradient and a Fermi distribution fit with two parameters, midpoint and width. The former and latter correspond in each case to threshold and rate of change of Robustness with level. Two parameters which similarly correspond are *average* and *difference*, these are discussed in more detail in James (1992b), but are also defined below. In addition we look at the Robustness Index for a set of fixed L:

Intercept of linear regression for Robustness versus presentation Level, R_{int}.

Slope of linear regression for Robustness versus presentation Level R_{grad}.

The mid point of the Robustness curve versus level as calculated from a Fermi distribution fit, R_{mid}.

The width of the Robustness curve versus level as calculated from Fermi fit, R_{width}.

The *average* value L_{av}, where:

$$L_{av} = \frac{L_{max}^- + L_{min}^+}{2}$$

where L_{max}^- is the maximum level at which the word was perceived incorrectly across all subjective trials, and L_{min}^+ the minimum level at which the word was perceived correctly across all trials.

The *difference* L_{diff}:

$$L_{diff} = |L_{max}^- - L_{min}^+|$$

The Robustness R of the word at a particular level L, R_L.

3. DATA ANALYSIS

The data were organised into files containing lists of values for each measured parameter. These were ordered according to the original order of the lists and words. This allowed use of various analysis programs to ascertain the degree of interrelation between the different measures. Comparisons, within and between *types*, between measures were made using Linear Regression (Hays, 1963) implemented in the "Unixstat" suite of programs running on a Masscomp 5450 minicomputer. These calculations were also confirmed using the "Minitab" functions "regress" (least squares) and "rregress" (ranked regression) running on a Hewlett Packard 9000s/800 computer. The final regression analysis for prediction of subjective parameters was also performed using the "Unixstat" programs. All these programs provided correlation matrices for any combination of the parameters.

4. RESULTS

4.1. Objective measurements

Summary statistics for a selection of ten objective measures were calculated, these are divided into four *types*; level, spectrum, duration and word usage. There is little to compare about the level and spectrum parameters except to say that the average impulse level L_{Aimp} (indicating peaks) is higher than that obtained with the SLM set on "Fast" (L_{Afast}). The variance in the total power L_{Ptot} is proportionally less than that for L_{Prms} . The longest word "wide" is over twice as long as the shortest word "jot". The highest energy word was "goes" using L_{Prms}, L_{Ptot} and L_{Aimp}, and "dodge" using L_{Afast}. "hutch" was the lowest energy word using both calculated measures and "cheek", "keys" and "shoot" were of lowest level using both SLM settings.

The word with the highest average spectral flatness σ_{mean} was "fish", and with the lowest "veil". Words with highest and lowest variability σ_{var} in spectral shape were "vice" and "hip".

The lowest word count $LogU$ of -0.693 corresponds to a value of $U = 0.5$, that is a zero count word (to the precision of counting). This rounding was also used in calculating the other usage parameters such as $LogV_{c1+c2}$ and $LogV_{c1c2}$. This rounding will introduce errors into these values which may have effects in the regressions. (Also of note is that the minimum for $LogV_{c1c2}$ is -11.51, unfortunately the lower limit on the precision of the ratio calculation (Minimum V_{c1c2} < 0.00001). The zero count words were "haze", "hoof", "thatch", "hutch", "thieve" and "rove", the highest count was for "have" .

Within the *types* level and word usage there was a large degree of correlation (r > 0.5). This is expected with the level parameters, which have only subtle differences in the treatments of the measurements. The smallest correlations were between L_{Atot} and L_{Aimp} within the level *type* (r = 0.545). It is of interest that the mean spectral flatness σ_{mean} exhibited a degree of correlation to the level parameters (r > 0.4, except for L_{Aimp}). σ_{var} and t_{frame} appeared, to a large degree, to be unrelated to any other *type*, and σ_{mean} and σ_{var} were independent (r < 0.185).

We had hoped by use of the word usage measures $LogU$, $LogV_{c1+c2}$ and $LogV_{c1c2}$, to isolate the effects of overall word count from those due to cohorts of the target words. However for these cases $LogU$, $LogV_{c1+c2}$ and $LogV_{c1c2}$ are highly correlated (r > .75), thus it is hard to justify the inclusion of V_{c1+c2} and V_{c1c2}. above absolute word usage $U.$.

4.2. Subjective Parameters

R_{int}, R_{mid} and L_{av} are thresholds related in terms of experimental presentation level, and R_{width} and L_{diff}, express the widths or slopes of the word intelligibility curves. The value R_{grad} may roughly be equated to the width of the intelligibility function in decibels (the limen) if multiplied by 2000 (Robustness range -1000 to 1000).

The word "laze" was measured as having the highest threshold using R_{mid} (57.1 dB) and L_{av} (98.0 dB), and "thighs" the highest using R_{int} (58.6 dB). "Fog" was shown to have the lowest threshold by R_{int}, 16.1 dB, and second lowest by L_{av}. In this case the word with the lowest was "man", 18.0 dB. Using R_{mid}, "bone" had the lowest threshold at 29.0 dB. The words with the flattest intelligibility curves was "poach" by R_{grad} and R_{width}, and "laze" (an old favourite) by L_{diff}. The words with most rapidly increasing intelligibility were "thieve" by R_{grad}, and "will" by R_{width}. Numerous words had L_{diff} equal to zero.

Within the threshold predicting *type* there was high correlation (r > 0.7), R_{grad} and R_{width} were well correlated (r > 0.5). This was not the case within the slope predicting *type*. This is to be expected as the straight line approximates the Fermi fit in this situation. There were various relations between the two *types* and of note is the degree of relatedness between R_{width} and R_{int} (r = 0.513). This may be due to some artifact in the two methods of describing the shape of the intelligibility curves.

Table 1. Summary statistics for the subjective data expressed in terms of Robustness Index R_L at various presentation levels L.

Variable	R_{30}	R_{35}	R_{40}	R_{45}	R_{50}	R_{55}	R_{60}
Minimum	-800	-500	-500	-200	-100	-50	0
Maximum	50	350	750	750	850	850	850
Mean	-344	-93	30	182	351	401	420
Std. Dev.	230	197	237	242	249	231	227

Table 2. Correlation of a selection of subjective parameters versus objective predictors, and total regression taking all selected predictors into account. The stars (*) indicate the degree of significance; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For the individual objective predictors, a significant result implies that: taking into account all the other predictors, the predictor in question significantly contributes to the overall regression (Hays, 1963).

Variable	Total	L_{Afast}	σ_{mean}	σ_{var}	$LogU$	t_{frames}
R_{int}	***0.518	-0.212	0.295	*0.219	***0.352	0.160
R_{mid}	***0.517	*-0.252	0.271	*0.200	***0.353	0.148
L_{av}	**0.417	-0.213	0.222	0.113	**-0.314	0.094
R_{grad}	0.256	-0.135	0.109	0.077	-0.119	-0.034
R_{width}	0.207	-0.028	0.179	0.112	-0.164	0.051
L_{diff}	0.220	0.218	-0.115	0.120	0.055	0.097

Table 3. Correlation of R_{mid} versus various combinations of a selection of objective predictors, and the total regression taking all selected predictors into account. The stars (*) indicate the degree of significance; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (see text). Note that the removal of certain predictors from the regression not only affects the total regression, but also the importance (significance) of other predictors.

Total	L_{Aimp}	L_{Afast}	σ_{mean}	σ_{var}	$LogU$	t_{frames}
***0.517	-------	*-0.252	0.271	*0.200	***-0.353	0.148
***0.511	-------	***-0.252	-------	**0.200	***-0.353	0.148
***0.494	-------	*-0.252	0.271	**0.200	***-0.353	-------
***0.504	*-0.199	-------	*0.271	*0.200	***-0.353	0.148
***0.472	**-0.199	-------	-------	**0.200	***-0.353	0.148
***0.483	*-0.199	-------	0.271	*0.200	***-0.353	-------
***0.477	-------	-------	**0.271	0.200	***-0.353	0.148

4.3. Summary statistics of R versus L data

Table 1 gives summary statistics for Robustness Index R_L (as specified above) over a range of presentation levels L. The greater the value of R_L the more intelligible the word at level L, where a zero value of R_L indicates 50% intelligibility (i.e. either outcome, right or wrong, is indeterminate), and a large and negative R_L low intelligibility. From the means of the data we can deduce that the mean threshold across all words is in the region of 35 to 40 dB. Of note is the comparatively large standard deviation of R_Ls to their ranges and means.

4.4. Regression of subjective versus objective

In this section we list the end results of this study, that is the prediction of subjective parameters by objective measurements. Table 2 shows the correlation between a selection of subjective parameters and a selection of objective predictors. The contributions of objective predictors are combined to give a total prediction for each subjective parameter, the correlation between the combined regression against the subjective parameter is then obtained. The significance of the contribution from each objective predictor is also shown, the hypothesis is: The predictor does not contribute to the overall regression taking into account all other predictors. The stars indicate the degree of significance of the correlations. Thus a significance level of $p < 0.05$ implies that there is only a 5% probability that the predictor in question does not contribute to the overall regression taking all other predictors into account. We can see that $LogU$ has the greatest and a unique contribution to the prediction of the subjective parameters indicating threshold, followed by L_{Afast} and σ_{var}.

4.5. Contributions and combinations of objective predictors

Table 3 allows comparison between regressions of R_{mid} using seven different combinations of objective predictors. We can see the overlaps in contributions made to the total regression by different predictors, i.e. L_{Afast} and σ_{mean} do not uniquely contribute to the overall regression when used together. The correlation coefficients are repeated for each combination so one can see how the individual contributions from predictors make up the total regression.

4.6. R_Ls versus objective predictors

The correlation of R_Ls calculated from subjective data versus a selection of predictors is shown in Table 4. These comparisons indicate the relative importance of predictors at different levels. The stars (*) denote significant contributions from individual predictors and the final regression. We can see that in all cases the predictor t_{frames} never makes significant contributions in the presence of the other predictors. L_{Afast}, σ_{var}, and $LogU$ are significant predictors at all levels except $L = 45$ dB, most noticeably σ_{var} is not significant (and indicates very little predictive capability) at $L = 35$ and 40 dB. The least significant ($0.01 > p > 0.001$) total regression was obtained with $L = 45$ dB, where neither L_{Afast} or σ_{var} are significant predictors (p's > 0.05). Stepping only 5 dB up from this level gave the highest regression coefficient with the three predictors, L_{Afast}, σ_{var}, and $LogU$, all making very highly significant contributions to the multiple regression (p < 0.001).

5. DISCUSSION

The object of this study has been to investigate the relative importance or significance of particular objective measures in the prediction of subjective thresholds, and word intelligibility characteristics. Many of the objective measures described in earlier sections of this paper were not used in the final selection used in the results presented above. The criterion for selection as a useful predictor was that the parameter must be either uncorrelated with other parameters, or that a group of correlated parameters produced the most significant prediction. Certain parameters were merely mathematically "treated" versions of others (i.e. L_{Prms}, and P_{rms} etc.) and these were compared within *types* with parameters with similar units (i.e. L_{Prms} with L_{Afast}).

The following discussion is divided into sections by parameter *type*, with a penultimate section bringing all these together. In the final section we look at the multiple regression model obtained with a "good" set of predictors and discuss the properties of particular words which stray from this model.

5.1. Speech level and σ_{mean}

Much attention in the past few decades has been given to the effective measurement of speech levels, and the equalisation of speech levels between words and between sentences used in speech audiometry. Most of the conclusions of these kinds of studies (Fuller and Whittle, 1982, Steeneken and Houtgast, 1979, and Tschopp, 1991a and 1991b) have been that subjective measures are not simply related to speech levels, but that reasonable accuracy may be obtained using measures such as L_{Afast}.

Here we will consider the relative merits of L_{Afast} and L_{Aimp} in the prediction of speech thresholds. We can also bring into the discussion the parameter σ_{mean} which was well

correlated with measures of speech level: σ_{mean} was well correlated with L_{Afast} but not so much with L_{Aimp}. We can therefore conclude that when σ_{mean} and L_{Afast} are both included in the multiple regression versus R_{mid} (a subjective threshold) that the significance of their individual contributions will be reduced. In fact the apparent overlapping of contributions from L_{Afast} and σ_{mean} is such that the inclusion of σ_{mean} has very little effect on the total regression (Table 3), this is true to a lesser degree when using L_{Aimp} without σ_{mean}. Removing both direct measures of level still leaves a good total regression. L_{Afast} gives the most significant prediction of threshold measures (Tables 2 and 3) in all situations.

In the treatment of Robustness R at level L, we have used only L_{Afast} as a predictor of the level *type*. The mean threshold of the words used here were judged to be in the region of 40 dB, with a standard deviation of order 5 dB. Table 4 shows the significance of L_{Afast} in predicting mean performance (i.e. R_L) in the range 30 to 60 dB.

Below 40 dB L_{Afast} is not significantly useful in the prediction of performance, for example R_{30} (the value of R at $L = 30$ dB), this is expected since the standard deviation of L_{Afast} is only 2.6 dB and the range approximately ± 6 dB. The information imparted to the listening individual at this level is going to be severely affected in all cases. Moving up in level, to R_{35} and R_{40} speech level becomes highly significant. This we should expect in the critical region around threshold, since the effects of any factor which marginally affects the amount of information made available will be amplified by the action of lexical contexts (James et al., 1992a).

At $L = 45$ dB, the speech level again becomes unimportant as all the differences in information giving due to speech level are largely equalised at this listening level. At 50 dB and above the speech level is again significant. This might be due to the perception of those words with little contextual information (perhaps we could term them *fragile*) as being more wholly dependent on the energy of speech features and having low intrinsic context. This idea can be confirmed if the words with high familiarity are separated from those with low familiarity. We will treat this as a probabilistic division and use the simple word count of data as the criterion of division of our sample. That is we will divide the sample into two sets, those with the highest counts and those with the lowest. The words were divided into two sets, the first (a) words with $LogU > 3.478$, and the second (b) $LogU \leq 3.478$, where the mean of $LogU$ for the whole sample was 3.478. We now calculate the regression of L_{Afast}, σ_{var}, $LogV_{c1+c2}$ and t_{frames} for the two sets (a) and (b) against R_{40}, R_{45} and R_{50}, the results are shown in Tables 5.

The results for regression against R_{40} for both sets, show that for the high familiarity set (a), $LogV_{c1+c2}$ is the more significant predictor, whereas for the low familiarity set (b), L_{Afast} is more useful. At 45 dB, the correlation L_{Afast} versus R is very low for set (a) but significant for (b). For R_{50}, the importance of speech level is greater for the low context set (b) than for the high context set (a). This confirms the hypothesis that speech energy is much more important for the recognition of low predictability items (b) than for high predictability items (a).

Table 4. Correlation of subjective R_L's versus objective predictors, and total regression taking all selected predictors into account. The stars (*) indicate the degree of significance; * p < 0.05, ** p < 0.01, *** p < 0.001 of individual predictors and for the total regression.

Variable	Total	L_{Afast}	σ_{var}	$LogU$	t_{frames}
R_{30}	***0.477	**0.227	**-0.203	***0.325	-0.134
R_{35}	***0.417	**0.243	-0.065	**0.294	-0.152
R_{40}	***0.394	***0.271	-0.063	**0.253	-0.101
R_{45}	**0.340	0.135	-0.123	**0.272	-0.084
R_{50}	***0.521	***0.230	***-0.263	***0.330	-0.155
R_{55}	***0.494	**0.187	***-0.248	***0.338	-0.156
R_{60}	***0.497	**0.218	**-0.213	***0.351	-0.171

Tables 5. Correlation of subjective R_Ls around threshold versus objective predictors, and total regression taking all selected predictors into account, for (a) words with $LogU > 3.478$, 64 words (top), and (b) words with $LogU \leq 3.478$, 56 words (bottom). The stars (*) indicate the degree of significance; * p < 0.05, ** p < 0.01, *** p < 0.001 of individual predictors and for the total regression.

(a) High usage ($LogU > 3.478$)

Variable	Total	L_{Afast}	σ_{var}	$LogV_{c1+c2}$	t_{frames}
R_{40}	*0.443	*0.175	-0.093	**0.286	-0.151
R_{45}	0.353	-0.016	*-0.250	0.200	-0.081
R_{50}	**0.492	*0.090	***-0.372	0.088	*-0.203

(b) Low usage ($LogU \leq 3.478$)

Variable	Total	L_{Afast}	σ_{var}	$LogV_{c1+c2}$	t_{frames}
R_{40}	*0.421	**0.393	-0.011	0.175	-0.052
R_{45}	0.361	*0.350	-0.062	0.141	-0.105
R_{50}	**0.502	**0.406	*-0.131	0.242	-0.127

5.2. The significance of σ_{var}

The variance of spectral flatness σ_{var} has been described above as a measure of the dynamics of the spectrum, and for this sample is uncorrelated with the mean spectral flatness σ_{mean}. As a predictor of threshold, σ_{var} is a significant predictor of subjective threshold (Tables 2 and 3), in that the lower the variation of spectral shape the lower the threshold. We also surmise that the transmission of finely detailed structure, indicated by high σ_{var}, would more greatly be affected by reduced listening levels, and the correct perception of the word would require the reception of this fine detail.

The correlation of σ_{var} versus R_Ls, indicates that σ_{var} generally works above threshold ($L = 50$ dB). However, at low levels (i.e. $L = 30$ dB), the variance of spectrum may provide necessary clues for the perception of high predictability words. In fact for set (a), σ_{var} is a highly significant (correlation coefficient = -0.327 ,p < 0.001) predictor of R_{30} this is not true for set (b) (correlation coefficient = -0.055, p = 0.337). Unlike speech level, σ_{var} is a significant predictor of performance R just above threshold ($L = 45$ dB) for high frequency words, becoming very significant at $L = 50$ dB.

5.3. Word familiarity and word intelligibility

It is clear from Tables 2 to 5 that word usage, represented here by $LogU$ is a very significant predictor of word intelligibility. It is obviously unrelated to any of the other objective parameter *types* since it is in essence not a "physical" parameter of speech. It is interesting to note the differences in word intelligibility characteristics between the sets of words (a) and (b) (as defined in Section 5.1) are due to the effect of probabilistic biases which we might predict from the word counts. In Section 5.1 we used $LogU$ as the criterion measure because it simply indicates a probability of response where information from auditory cues is severely limited. It was seen that for low usage words at low levels ($L = 40$ dB), speech energy (L_{Afast}) had the greatest effect on identification. However, for high usage words the cohort size measured by $LogV_{c1+c2}$ was significant and speech level only to a lesser degree.

5.4. Word duration t_{frames}

Throughout the regression analysis t_{frames} has been incorporated because we found this parameter largely uncorrelated with every other (except for L_{Ptot}, which we did not use in any of the regressions above). In most examples the duration seems to be a useful predictor but fails to be actually significant in the presence of other "good" predictors in all but one example (Table 5a).

The trend of intelligibility R_L at level L tends to go: Longer word, then lower intelligibility. Since the trend predicted by t_{frame} is relatively weak, it is difficult to justify any particular reasons for the trend. We can take a similar approach to that in Section 5.1, by separating the sample data into two groups on the basis of one parameter. We can now define two new groups; (c) where $t_{frames} > 272$, and (d) $t_{frames} \leq 272$, where the mean of t_{frames} for the whole sample is 272.

Some summary statistics for the two groups of words (c) and (d) are given in Tables 6.

Tables 6 show us that (c) and (d) have similar distributions of characteristics, thus when comparing the two groups further we can rule out effects other than due to the dividing parameter, duration t_{frames}.

In Tables 7 we show the results of linear regressions of R_{mid} versus a set of "good" parameters as previously described. The most striking difference between the two regressions is the relative importance of σ_{var} and t_{frames}. For the longer words (c), t_{frames} gives a significant trend whereas σ_{var} is not such a good predictor. The reverse is true for the short words (d). Though weak, this result suggests that shorter words require greater uniformity in their frequency characteristics (i.e. smaller σ_{var}) than longer words to achieve lower thresholds. It is therefore important to maintain a reasonable information rate whatever the total duration for low thresholds or good intelligibility, or "squashing too much information in too short a time gives higher thresholds".

Tables 6. Summary statistics for some parameters of (c) words with $t_{frames} > 272$, 62 words, and (d) words with $t_{frames} \leq 272$, 58 words. Note that there is little difference between the two sets (c) and (d) in terms of the distributions of R_{mid}, L_{Afast}, σ_{var} and $LogU$.

(c) $t_{frames} > 272$

Variable	R_{mid}	L_{Afast}	σ_{var}	$LogU$	t_{frames}
Mean	42.14	121.6	0.026	3.23	296
Std. Dev.	8.01	2.8	0.008	2.18	17.31

(d) $t_{frames} \leq 272$

Variable	R_{mid}	L_{Afast}	σ_{var}	$LogU$	t_{frames}
Mean	40.26	121.3	0.024	3.74	242
Std. Dev.	5	2.5	0.009	2.17	19.58

Table 7. Correlation of R_{mid} versus objective predictors, and total regression taking all selected predictors into account, for (c) words with $t_{frames} > 272$, 62 words, and (d) words with $t_{frames} \leq 272$, 58 words . The stars (*) indicate the degree of significance; $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ of individual predictors and for the total regression.

Variable	Total	L_{Afast}	σ_{var}	$LogU$	t_{frames}
R_{mid} (c)	***0.536	*-0.271	0.122	**-0.404	*0.173
R_{mid} (d)	**0.500	*-0.246	**0.276	*-0.266	0.033

5.5. The combination of predictors

In the discussion Sections 5.1 to 5.4 we examined the relationship between objective and subjective parameters. Using a "good" combination of predictors, L_{Afast}, σ_{var}, $LogU$ and t_{frames} for example, we can calculate a linear regression equation for R_{mid}, where $\hat{R}_{mid}$ is the estimated value of R_{mid}. From this equation we can see that a change in speech level

$$\hat{R}_{mid} = -0.841\,L_{Afast} + 229\,\sigma_{var} - 1.09\,LogU + 0.033\,t_{frames} + 132$$

L_{Afast} gives an almost one for one change in word threshold R_{mid}. Given that the range of L_{Afast} was 13 dB, we obtain a predicted range of 11 dB in the subjective threshold R_{mid}. Similarly we can approximate ranges of 9 dB due to σ_{var}, 11 dB due to $LogU$ and 7 dB due to t_{frames} in R_{mid}. This gives us an indication of the magnitude of effects due to contributions from individual predictors.

Figure 2 shows a scatter plot of $\hat{R}_{mid}$ versus R_{mid}. The line $\hat{R}_{mid} = R_{mid}$ is marked giving the line of "perfect" correlation. Points corresponding to particular words are also indicated. Of great interest are those words which deviate most from the line of perfect regression, such as "thumb" and "ways". These are words which conform least to the objective prediction of intelligibility given here. The word "ways" has a much greater subjective threshold R_{mid} than estimated threshold $\hat{R}_{mid}$. It is hard to explain this phenomenon in terms of the quantities we have measured, instead we have to look more closely at the structure of the word. In this case the author was able to look back at the responses recorded in the study described in James (1992a) which looked at word list intelligibility under certain conditions. It was possible to consider the response set for the three plurals included in the sample set of "Boothroyd" words; keys, thighs and ways. Subjects tended to lose the perception of the "zz" in "keys", making "key" and there were also two non-word responses of "keeve". The word "thighs" was commonly confused with "five" or just "thigh". Looking at confusions for "ways", the author found many occurrences of the word "wave", in fact 19 out of a total number of 36 presentations! It is likely under the conditions imposed in that study (and at low presentation levels) that sound of the plural "zz" was lost or masked leaving the subject with an elongated version of "way" (try saying "way" and then "ways", and compare the length!), as this speech sound is too long to "fit" into "way", then it is only natural to attempt an alternative, in this case "wave" seems a reasonable (and popular) alternative.
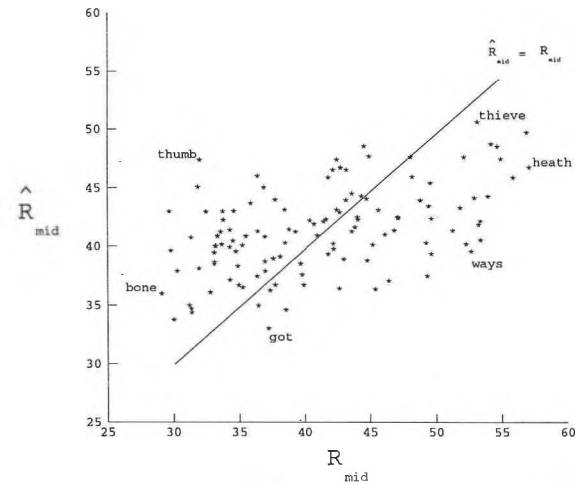

FIG. 2. Scatter plot of $\hat{R}_{mid}$ versus R_{mid}, where $\hat{R}_{mid}$ is the estimation of R_{mid} from regression analysis.

At the other extreme lies the word "thumb" which has a much lower subjective threshold than predicted. To some extent the word "thumb" could be likened to a diphthong (for example "veil"); it has two voiced parts with different tones, the vowel "uh" and the nasal "mm", these features may make it robust. However, if the "um" combination is a "robust" feature, why should "thumb" not still be confused with words within its response cohort (e.g $c1+c2$), such as rhyming words "sum" and "dumb"?

For the two exceptions above we have looked in more detail at the perception of speech features. In the first case we were able to call on experience and use previous results to provide some clues for the deviation of subjective thresholds from those predicted. It must be remembered that from the selection of physical measures we have chosen in this study, there are not likely to be a "best" set of predictors which can explain all the variances of subjective data in all cases. At this point we can apply previously obtained results to the behaviour of the entire test set: we can simply calculate the correlation between the number of incorrect responses for each word from the previous study (James et al., 1992a) and the subjective threshold R_{mid}. Not surprisingly this gives us a very high correlation (coefficient = 0.808), much higher than that obtained from the best multiple regressions listed above. This result gives us insight into the consistency of results between experiments and reliability of the word recognition task. Perhaps the "membership" of response cohorts will be similar in all but a few cases, and those few cases could be assessed by further experiment. Explaining the properties of words which fall furthest from the predictions is not the least difficulty in this kind of study. One could explain, perhaps, the behaviour of all words compared to the "ideal" ones (those on the line of perfect

regression), implying that the presence of any particular word used in audiometric material should be justified; i.e. not too unfamiliar, not too familiar, not a plural, not a verb etc. Under these restrictions the compilation of test material such as used in phonemically balanced designs becomes more difficult. At this point we can look back at work done by Hood and Poole (1980) which gives us useful information about the consistency of word intelligibility between speakers and in relation to familiarity as indicated by frequency of occurrence. They state that the effect of the speaker outweighs consideration of familiarity, this is something which we were unable to account for in this study. But they do offer us an olive branch in that measurements with one speaker were generally fairly consistent for many of the words. Inconsistent words are attributed to intra-subject and perhaps subject-speaker-word interaction. Perhaps the time and frequency analysis used here would encompass these causes of variance. Hood and Poole also stated that they and others (e.g. Tobias, 1964) regard such measures as phonetic balance to be unnecessary in the design of speech perception tests, this is now largely accepted.

Thus an alternative to theoretically structured designs then is to use the implicit properties of individual words: if the response set for a given word under given conditions is well defined, then we can define word intelligibility in terms of the size and properties of this response set. Thus we can define speech hearing impairment in terms of "mistakes". This might have some use in the selection and tuning of hearing aids. According to the kind of mistake we can predict the "conditions" imposed by impairment by comparison with response sets from experiments utilising a range of enforced conditions. This is an opened ended approach, unlike forced-choice methods, and by its nature it takes into account all effects of word usage, word confusion, speech spectrum and level.

6. CONCLUSIONS

The trends indicated by the analysis used in this study should apply to mono-syllables outside the sample used here. It is a useful result that word usage $LogU$ plays an important part in the prediction of thresholds (i.e. R_{mid}), and therefore it may be used as a guide to selecting test sets in the future, before making recordings of material. Other measurements, for example L_{Afast}, σ_{var} and t_{frames}, may only be made on specific material. It would be interesting to study the general trends for these quantities for a number of different speakers and examples from the same speakers (assuming that they are all trying to make recordings to audiological standards). For example, is the property "speech level" characteristic for a given word or does it only measure level for a given example (as would be expected). Similar arguments could be applied to the other measures used here: One could predict that the speech spectrum and duration would be more characteristic for a given word.

The results presented here pose questions about the most rational method of selecting material for the measurement of speech hearing. Certainly the use of known response sets in the analysis of response errors would be useful. There is also scope for the use of smaller sets of test material with particular properties, as obtained by experiment, aimed at making specific measurements of hearing acuity.

This study has shown that the intelligibility characteristics of individual words can be predicted to a fair degree by objective measurements. Some interaction is apparent between the sources of information, namely physical and linguistic (or contextual). The measure σ_{var} seems to be a useful complement to the other parameters and indicates the importance of information rate or feature rate in speech perception.

7. ACKNOWLEDGEMENTS

8. REFERENCES

Boothroyd, A. (1968), Statistical Theory of the Speech Discrimination Score, *Journal of the Acoustical Society of America*, **43**(2), 363-367.

Boothroyd, A. (1968), Developments in Speech Audiometry, *Sound*, **2**, 3-10.

Broadbent, D.E. (1967), Word-Frequency Effect and Response Bias., *Psychological Review*, **74**(1), 1-15.

Dubno, J. and Levitt, H. (1981) Predicting consonant confusions from acoustic analysis. *Journal of the Acoustical Society of America*, **69**, 249-261.

Francis, W. and Kučera, H. (1982), *Frequency analysis of English usage: Lexeican and grammar.* Pub. Houghton Mifflin, Boston.

Fuller, H.C and Whittle, L.S. (1982), The Measurement of speech levels for Audiometry. *Proceedings of the Institute of Acoustic, U.K.*, Spring 1982, Paper **D.3.5**.

Hays, W.L. (1963), *Statistics*. Chapter 14. Holt, Reinhart and Winston, UK.

Hood, J.D. and Poole, J.P. (1980), Influence of Speaker and Other Factors Affecting Speech Intelligibility. *Audiology* **19**, 434-455.

Howell, D.C. (1982), In *Statistical Methods for Psychologists*, Chapter 11, Duxbury Press.

James, C.J. (1991), The Frequency of Usage of 120 Boothroyd Words. *British Society of Audiology Newsletter*, May 1991, 13-14.

James, C.J., Bowsher, J.M. and Simpson, P.J. (1992a), Digitization effects and the non-equivalence of isophonemic word lists: A cautionary tale. *Acustica* **75**(4), 45-57.

James, C.J. (1992b), Word Intelligibility Experiments. *In: The Application of Computers to Speech Audiometry. Ph.D Thesis, University of Surrey, UK*. Chpt. 7.

Jayant, N.S. and Noll, P. (1984), *Digital coding of waveforms*, Prentice-Hall, New Jersey.

Lee, K and Dermody, P. (1992), The relationship between perceptual and acoustic analysis of speech sounds. *Proceedings of the Australian Speech Science and Technology Association*, December 1992, 14-19.

Lyregaard, P.E. (1976), *On the relation between recognition and familiarity of words*. National Physical Laboratory Report Ac 78.

Markides, A. (1978), Speech discrimination functions for normal hearing subjects with AB isophonemic word lists, *Scandinavian Audiology* **7**, 239-245.

Morton, J. (1969), Interaction of information in word recognition, *Psychological Review* **76**/2, 165-178.

Rosenzweig, M.R and Postman, L. (1957), Intelligibility as a function of usage, *Journal of Experimental Psychology* **54**(6), 412.

Steeneken, H.J.M. and Houtgast, T. (1979), *Comparison of some methods for measuring speech levels*. report no. IZF 1978-22, Institute for Perception TNO, Soesterburg, Netherlands.

Thorndike, E.L. and Lorge, I., (1944), *The Teacher's Word Book of 30,000 Words*. Bureau of Publications, Teach. Coll., Columbia University, New York.

Tobias, J.V. (1964), On phonemic analysis of speech discrimination tests. *Journal of Speech and Hearing Research*. **7**, 98-100.

Tschopp, K., Beckenbauer, T. and Harris, F.P. (1991a), Objective Measures of Sentence Level with respect to Loudness, *Audiology* **30**, 113-122.

Tschopp, K., and Fastl, H. (1991b), On the Loudness of German Speech Material Used in Audiology, *Acustica* **73**, 33-34.

Wayland, S.C et al. (1989), Recognition of isolated words: The dynamics of cohort reduction, *Applied Psycholinguistics* **10**(4), 475-487.