THE RELATIONSHIP BETWEEN PERCEPTUAL AND ACOUSTIC ANALYSIS OF SPEECH SOUNDS.

Kerrie Lee[1] and Phillip Dermody[2]

[1]School of Communication Disorders
University of Sydney

[2]Speech Communication Research
National Acoustic Laboratories

ABSTRACT - The study investigates the relationship between the perceptual identification of CV syllables and acoustic analyses of these sounds. The perceptual data for the speech sounds are used to show the relative robustness of the different speech syllables. Acoustic analyses including traditional measures of overall energy, duration and position of spectral peaks as well as measures based on spectral slope and ratio measures between spectral peaks were performed on the same syllables. The results suggest that acoustic characteristics such as energy and duration are not simply related to perceptual robustness of speech sounds and that perceptual robustness is related to relational measures between spectral peaks in the signal.

INTRODUCTION

Research data on the acoustic correlates of perceptual performance for speech sounds have been largely concerned with the description of a very limited set of stimuli in each study and there are very few investigations that have attempted to provide general acoustic features that might be used as correlates of perceptual results. For instance, one approach has been to use multidimensional scaling analysis of perceptual data to define the acoustic dimensions which group speech sounds (Wish and Carroll, 1974; Soli, Arabie and Carroll, 1986). In these studies acoustic measures of speech sounds were compared with published perceptual data of Miller and Nicely (1955).

An alternative approach is to determine what acoustic factors might be responsible for the relative robustness of some sounds compared to others in the same acoustic conditions. Dubno and Levitt (1979) used this approach to provide data about the acoustic correlates of perceptual performance. In their study the percent correct performance for spoken syllables was used as the dependent variable in a linear regression analysis for predicting which acoustic features play a role in syllable identification. An advantage of the Dubno and Levitt study is that the actual speech sounds used in the assessment of perceptual performance were also used for the acoustic analysis.

Dubno and Levitt (1981) investigated 15 variables derived from their acoustic analysis of their speech sounds. The variables related to spectral and intensity dimensions of the speech signal. Overall, their results suggest that the acoustic variables chosen for analysis did not correlate very highly with perceptual performance. The results for the unvoiced stops and fricatives in both quiet and noise conditions indicate some relationship but the effect seems limited to those classes of sound in an /i/ vowel context. For these stimuli the acoustic variables that were successful in predicting the perceptual results included: the consonant energy, the consonant duration and bandwidth and the 1st and 2nd spectral peaks of the consonant in quiet and the highest consonant spectral peak, the consonant-noise ratio, consonant energy and duration, and the vowel peak frequency and overall energy.

Dynamic variables derived from the acoustic analysis which focused on the second formant transition were not as successful as the static measures of energy and duration for perceptual performance prediction and it seems from this result that the acoustic information about the consonant could be more important for correlation with perceptual performance than information about the transition into the vowel.

On the basis of Dubno and Levitt's results and because of the difficulties of adequately defining and measuring dynamic changes in the speech signal only static measures of acoustic structure are considered for use with the perceptual data reported here. This also means that the present data can be compared with previous studies.

The second problem of how many variables to include for analysis is also problematic. For the present study, traditional measures including speech sound energy and duration were investigated, as well as some new ways to characterise some aspects of the relativity of speech sound variables. Stevens (1987) has discussed the need to see acoustic analysis in relational terms rather than in absolute values of measures. An acoustic analysis of robustness needs to include measures of intensity and the spectral characteristics of the speech signal. Intensity can be considered not only in terms of an overall measure of amplitude of the signal, but also in terms of intensity of different frequency bands. For example, Tibbitts (1989) showed that the slope of the spectra in four bands: 0-2000Hz, 2000-4000Hz, 4000-6000Hz and 6000-8000Hz correlated with perceptual performance based on speakers rated as having high or low intelligibility.

Spectral characteristics of the signal can be described by an overall measure such as the frequency of the highest energy peaks in the signal (e.g., Dubno and Levitt, 1981). In addition, relational measures based on the overall measure may help to characterise the signal. Miller (1987) has suggested that the relationship between energy peaks in terms of the distance in frequency might identify the characteristics of the auditory-perceptual space of listeners identifying speech sounds. Duration was included as it is an obvious acoustic parameter on which consonants show large differences.

The final problem in looking for acoustic correlates of perceptual performance is what parts of the speech signal of consonant vowel pairs should be measured for the analysis. There are two broad aspects of the signal that can be labelled consonant and vowel as well as an intermediate portion that is the transition between these regions. In the present analysis, guided by the results of Dubno and Levitt (1981) who found that the most significant correlations between the consonant information and perceptual performance were static acoustic properties of the consonant segment, the choice was made to include measurements of the portion that included only the consonant segment.

PERCEPTUAL RESULTS

The stimuli used for perceptual testing were 108 CV syllables spoken by an adult male. The CVs were comprised of 18 consonants /p t k b d g s f θ z v ð l r m n h w/ in 6 vowel contexts (I, i, a, o, U, u/). The sounds were edited and amplitude equalised using a dB RMS measure. The sounds were presented to 12 listeners in a quiet room meeting audiometric standards. The stimuli were presented at a number of presentation levels using a randomised presentation of stimulus items. Each listener heard 8 repetitions of each stimulus at each presentation level. The data was analysed and the conditions that resulted in quartile performance scores were used for comparison with the acoustic analysis.

ACOUSTIC ANALYSIS MEASURES

Static Measures

The stimuli were edited into two segments labelled consonant and vowel. The consonant portion was defined as the part of the speech signal which ended at the point at which the speech signal showed a shift in amplitude and an onset of regularity or periodicity.

Several static properties of the acoustic structure of the consonant portion (CP) were used. These included the intensity of the CP, measures using the root mean square (RMS) amplitude, as well as duration in msec (called MSEC). In addition, the CP was analysed using a standard FFT. For the FFT analysis, the consonant portion from onset to the end of the segment was down-sampled to 18169 samples per second (i.e., half the original sampling rate) which could be easily implemented by removing every second data point. Prior to down-sampling the speech signals were filtered using a digital filter with zero phase shift. The resulting speech files, which were of varying lengths, depending on the length of the consonant, were then analysed using an FFT (with hamming window) over the entire length of the file. This analysis was displayed (0-8.5KHz) and the three highest peaks were measured by the experimenter using a cursor and the centre frequencies recorded. These three values were then ordered from lowest to highest in the frequency range (called spectral peak SP 1, SP2 and SP3).

Relational Measures

The slopes of the spectrum of each CP were derived from the FFT described above. The 8kHz frequency range was divided into four sections (0-2Khz; 2-4kHz; 4-6KHz; and 6-8KHZ). The 2KHz slope, or average variation in dB over a 2KHz region, was identified as a trend in the base of the spectrum over that range. This estimate of the slope followed the base of the spectrum or overall spectral shape within that range. The amplitude of the frequency bands (in dB RMS) was also included as a measure of the relative contributions of different frequency regions to the speech sound identification. The slope values are referred to as SLP 1-4 and the amplitude of the frequency bands as RMS 1-4.

A second measure was based on the ratio of the frequencies between the three highest spectral peaks. That is, the difference in Hz between each of the peaks was expressed as a ratio (i.e., SP1/SP2 is called SP1-2; SP1/SP3 is called SP1-3; SP2/SP3 is called SP2-3).

DATA ANALYSIS

The 16 acoustic measures were included in a multiple regression analysis using the percent correct score for each consonant averaged over the 3 test presentation levels which resulted in an average performance correct for the all consonants of 25, 50 and 75% correct for each vowel context as the dependent variable.

RESULTS

The results for the analysis for all the speech sounds collapsed over intensity levels are presented in Table 1. These results indicate that about 40% of the variance in the identification scores can be explained by the set of acoustic variables used in the present study.

An analysis of the correspondence of acoustic variables across all the signal presentation levels indicates very similar associations at each level with the energy measures, peak measures and slope measures, all contributing differentially to the performance at each level.

Variable	Multiple R	Multiple R^2
RMS	0.342	0.117
SP1-3	0.577	0.333
SP3	0.598	0.358
SLP1	0.614	0.377
SP2-3	0.635	0.404
RMS3	0.643	0.413
SLP2	0.649	0.422
RMS1	0.664	0.440

Table 1. Results for the regression analysis between the PI function perceptual data and acoustic analysis measures, showing the acoustic variables which contributed to the regression.

The results from the present analysis suggest that it is not possible to explain the individual performance of different speech sounds at different levels by a simple addition of static acoustic factors. However, because the results of Dubno and Levitt (1981) also indicate that vowel context may play an important role in the regression analysis (in their case the stops and fricatives paired with /i/ showed good correspondence with acoustic variables while other consonants and different vowel contexts did not). A regression analysis was therefore carried out which looked for vowel effects across all presentation levels combined. The results of these analyses are presented in Table 2.

Vowel	Multiple R	Multiple R^2
/ae/	0.97	0.94
/e/	0.95	0.90
/^/	0.92	0.85
/u/	0.73	0.53
/I/	0.68	0.46
/o/	0.64	0.42

Table 2: Results of regression analyses showing the value of the Multiple R and Multiple R^2 when the perceptual results for each vowel context are run separately against the acoustic variables.

The results in Table 2 suggest that the acoustic variables have considerable power in explaining the variance of the individual speech sound identifications in the /ae/,/e/ and /^/ vowel contexts but only moderate prediction of the identification performance from acoustic variables in the /u/, /I/ and /o/ vowel environments.

In the /ae/ vowel context the variables that are included in the regression equation are i) overall RMS amplitude of the consonant segment (RMS) ii) the ratio of spectral peaks 2 and 3; iii) spectral peak SP1 and SP3; iv) the slope of the frequency bands 1 and 4; iv) RMS energy in the frequency bands 1 through 4.

The significant variables for the /e/ vowel are: i) RMS ii) ratio of spectral peaks 1 to 3; ratio SP2 to 3 and ratio SP1 to 2; iii) the slope of the frequency bands 1 and 2; iv) the frequency of spectral peak 3 and 2;v) the RMS energy in frequency band 3.

The variables predicting identification of consonants with the /^/ vowel include: i) the ratio of the spectral peaks 1 to 3 and ratio of SP1 to SP2; ii) RMS iii) the slope of the frequency bands 3 and 4; iv) the duration in milliseconds of the consonant segment; v) the frequency of the spectral peaks 2 and 1; vi) RMS energy in frequency band 3.

The results for the /u/ vowel include only four variables which are: i) the ratio of spectral peaks 1 to 2; ii) the overall RMS energy iii) the RMS energy in frequency band 2; iv) the frequency of spectral band 2.

The consonant identification scores in the /I/ context are predicted only by i) RMS energy in the consonant segment ii) the ratio of spectral peaks 1 to 2.

DISCUSSION

For the vowel contexts in which the acoustic variables are reasonably successful in predicting the variance of the identification scores the energy of the consonant portion, the spectral peaks, their ratios and the slope of the spectrum represent acoustic variables that listeners might use to perceive speech sounds presented at low intensity levels.

The fact that acoustic variables derived from the consonant segment of the syllable are successful in some prediction of identification scores in some sounds but not in others suggests that the cues for some sounds may be found in the transition rather than the consonant portion. That is, different portions of the syllables might be used for identification at low presentation levels. Phoneme identification in some vowel contexts (in this case notably for the /ae,e,^/ vowel contexts) is based largely on the consonant segment while in other vowel contexts (for instance in /o/ in the present results) other segments of the signal also need to be used for accurate identification.

The present results suggest that the commonsense view that speech sound identification is dependent on the audibility of the phone segment is supported in the present study with the overall RMS energy of the consonant segment playing a significant role in the prediction of the individual speech sounds. However, it is also clear that other factors related to the relationship of the slope of different frequency bands as well as the relative position of energy peaks in the consonant portion also play an important role in differentiating consonants from each other as well as the same consonant in different vowel contexts.

REFERENCES

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

Miller, G., and Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. "Journal of the Acoustical Society of America", 27, 338-352.

Miller, J.D. (1987) Auditory perceptual processing of speech sounds. In "Auditory Processing of

Complex Sounds". Edited by W. Yost and C. Watson. Hillsdale,New Jersey:Lawrence Erlbaum, Assoc.

Soli, S., Arabie, P. and Carroll, J. (1986) Discrete representation of perceptual structure underlying consonant confusions. "Journal of the Acoustical Society of America", 79, 826-837.

Stevens, K. (1987) Relational properties as perceptual correlates of phonetic features. In "Proceedings of the XIth International Congress of Phonetic Sciences". Volume 4. Tallinn. pp. 352-355.

Tibbitts, J. (1989) "A Digital Signal Processing Technique to Improve the Intelligibility of Speech for the Hearing Impaired in Quiet". PhD dissertation, School of Electrical Engineering, University of Sydney.

Wish, M. and Carroll, D. (1974) Applications of individual differences scaling to studies of human perception and judgement. In E. Carterette and M. Friedman (Eds) "Handbook of Perception". Volume II. New York: Academic Press. pp. 449-491.