INTRODUCTION

Severely hearing impaired or deaf individuals are dependent on their lipreading (speech reading) skills to receive speech. However, within the English language there are many syllables and phonemes that look similar to a lipreader. For example, the syllables /pɑ/, /bɑ/, and /mɑ/ are extremely difficult for persons using lipreading to distinguish because the initial phonemes appear similar.

Our approach for providing automated speech cues to aid lipreading involves processing the speech signal to produce continuous parameters that can be presented visually to a user.

The challenge is to develop a visual display that presents speech parameters to a lipreader in a useful way. **Our study was aimed at preliminary development and testing of a new continuous**

parameter display. The new display was based on a higher order polygon representation of speech spectra and sound level.

I. SPECTRAL REPRESENTATION SPEECH

Sound Database

The primary database used to test this new display consisted of several pairs of speech samples --vowels and consonant-vowel sequences-- that are ambiguous to a lipreader. An attempt was made to include pairs with voicing, nasality, and other distinctions as categorized by Fisher and Logemann (1971). The speech sample pairs used in our study are listed in **Table 1**.

Database Analysis

A spectrum was obtained for a windowed and padded 30-ms segment of a speech sample to provide spectra at about 16-Hz intervals. This process was advanced through the speech sample at 5-ms intervals to provide spectra at a rate of 200 per second.

Visual Spectral Representation

The spectrum was divided into ten bands of equal intelligibility (based on work by Beranek, 1947). Spectral energy in these bands were weighted so to have equal spectral energy in the long term (based on data collected by Byrne et al., 1994). Each spectral band and corresponding weighting factor are shown in **Table 2**.

The spectral energy of a 30-ms speech

segment was divided by the ten bands to form ten band energies. These band energies were weighted then normalized to unity to provide a weighted band energy (*WBE*) having a value of less than one. These weight band energies formed the basis for the spectral polygon display.

Spectral Polygon (Decagon) Creation

The polygon spectral display was designed around a regular decagon as shown in **Fig. 1**. Axes are drawn from the decagon's center to the decagon's vertices as shown in the figure. The *WBE* for each of the ten spectral bands is plotted along its respective axis (energy increases from zero at the center of the decagon toward a value of one at each vertex). An irregular decagon is created

when the ten spectral band energies, each marked on its corresponding axis, are connected with lines, as seen in the figure. The “spectral decagon” or “deca-spectrum” changes shape at 5-ms intervals as new spectral information is obtained. Overall sound level is represented by line shade and line width, which are consistent throughout a single decagon.

II. TESTING THE SPECTRAL DECAGON DISPLAY

The spectral display consisted of sequences of time-evolving spectral decagons shown in real time as obtained at 5-ms intervals from each speech sample (**Table 1**). The **fifty subjects** used in the testing of this display were presented with pairs of spectral decagon sequences obtained from the ambiguous speech pairs listed in Table 1. **Each pair was presented in one of the four possible combinations** AA, BB, AB,

BA, where A and B each represent one of the samples in a pair. In the discrimination task, **subjects were asked to record a same/different decision by pressing a specified computer key after viewing a pair of spectral decagon sequences**. Before beginning the discrimination task, subjects were presented several spectral pairs with immediate feedback so they could become familiar with the task. After this,

subjects completed three similar discrimination tests.

Test by Test Description

In Test 1, subjects were presented with 72 display pairs (18 pairs in each of the four possible combinations) for discrimination. The order of presentation of pairs was randomized individually for each subject. After a few minutes of rest, a subject was given Test 2, which was similar to Test 1 but with a different randomization of the pair presentation order (inter-test correlation less than 0.1.). After a further short rest, a subject was given Test 3, again with a different randomization order. Subjects typically took 30-40 minutes to complete the entire process.

III. RESULTS

The three test results for each of the fifty subjects were analyzed from two different viewpoints. **First**, the overall results for each of the three tests were examined. **Second**, the results were examined from the perspective of discrimination of individual speech sample pairs. This led to an average of all subject responses for each of the 18 pairs and was expressed as a percent correct response (**See:** Fig.s 2, 3, and 4).

Discussion

A. Test to Test Results

The overall percent correct responses for all subjects on each of the three tests are shown in **Fig. 2.** This figure shows a trend of improvement in subjects' ability to discriminate the spectral display pairs from test to test. As subjects progressed from test to test, they apparently learned to discriminate on additional display characteristics.

B. Pair Test Results and Discussion

Deca-spectral pairs were presented in two ways: either the spectra in the pair were the same or they were different. **Fig. 3** shows cumulative responses of all subjects to "same" pair presentations in percent correct. The lowest percentage correct at 79% occurred for /shɑ/jhɑ/, pair 17, while /bɑ/mɑ/, pair 9, had the highest at 97%. Thus, when the spectral patterns are the same they are consistently judged to be the "same".

The "different" pair discrimination results of the eighteen pairs are shown in **Fig. 4**. Strengths and weaknesses of the display are more apparent in the "different" than in the "same" discrimination task. That eleven pairs were clearly discriminable by the subjects is an apparent strength of the display.

The three vowel comparisons, /I/i/, /ɑ/æ/, and /u/ʊ/ (pairs 1, 2, and 3), were

discriminable by subjects. This might be expected for the vowel samples where the deca-spectra shown in **Fig.5** were mostly constant and the display was of relatively long duration.

The consonants /k/ and /g/ were combined with the three vowels /ɑ/, /i/ and /u/ for the contrasts, /kɑ/gɑ/, /ki/gi/ and /ku/gu/ (pairs 4, 5 and 6). Subjects discriminated /kɑ/gɑ/ with 90% accuracy. Discrimination was reduced to 45% for /ku/gu/ and failed for /ki/gi/ (Negative values result from the correction for guessing.). A plausible explanation for these different results can be found in the stylized spectrograms of Liberman, DeLattre and Cooper (1957). There are large spectral transitions of sufficient duration in both /kɑ/ or /gɑ/ (low, back vowel) that permit subjects to discriminate the voiced/unvoiced contrast. The spectral transitions in /ki/ and /gi/ (high, front vowel) are much more subtle and so probably fail to permit discrimination. In fact, subjects judged them the same more often than not because they didn't notice the slight transition shown by the spectra as they went from consonant to vowel. The /ku/gu/ (high, back vowel) transitions fall between these two cases.

Now consider the discrimination results of the three samples /pɑ/, /bɑ/ and /mɑ/. The poor discrimination (25% correct) of /pɑ/bɑ/ (pair 7) illustrates a weakness of the display. Some insight can be gained by comparing their spectrograms and deca-spectra in **Figures 6 and 7**. The early appearing (0 to 75-ms) deca-spectra, though different in the two figures, occur only briefly and are

weak in intensity. On the other hand, the discrimination (better than 90% correct) of /pɑ/mɑ/ and /bɑ/mɑ/ (pairs 8 and 9) is very good (see Fig. 4). The longer duration (0 to 150-ms) deca-spectra shown in Fig. 8 for /mɑ/ apparently provided a discriminable contrast with those of /pɑ/ and /bɑ/. The nasal contrast in pairs 8 and 9 came through clearly on the display while the unvoiced/voiced contrast in pair 7 did not.

Similar results are found for the pairs /tɑ/dɑ/, /tɑ/nɑ/, and /dɑ/nɑ/. The poor discrimination (35% correct) of /tɑ/dɑ/ (pair 10) illustrates a weakness of the display similar to that of /pɑ/bɑ/. The discrimination (better than 90% correct) of pairs /tɑ/nɑ/ and /dɑ/nɑ/ (pairs 11 and 12) is very good (see **Fig. 4**). The longer duration deca-spectra for /nɑ/ apparently provided a discriminable contrast with those of /tɑ/ and /dɑ/. The nasal contrast in pairs 11 and 12 came through clearly on the display while the unvoiced/voiced contrast of pair 10 did not.

When /sɑ/ was paired with /shɑ/, /chɑ/ and /jhɑ/ (pairs 13, 14, and 15, respectively) discrimination of better than 90% correct resulted. The spectral decagons of /sɑ/, shown in **Fig. 9**, are quite different from those of /shɑ/, /chɑ/ and /jhɑ/ which probably form the basis for the good discrimination. However, when /shɑ/, /chɑ/ and /jhɑ/ are paired the discrimination is much poorer because of their similar spectra. The pair /shɑ/jhɑ/ is not discriminated which is probably due to lack of duration and spectral information for making the voiced/unvoiced distinction, a result similar to /pɑ/bɑ/ and /tɑ/dɑ/ results seen above. The better

discrimination (better than 60% correct) for the /shɑ/chɑ/ and /chɑ/jhɑ/ is likely due to differences in duration that overcome the similar spectra.

IV. CONCLUSION

The spectral decagon shows considerable promise as an aid to lipreading in that 11 of the 18 pairs were discriminated with better than 90% correct responses.

Two of the remaining pairs were discriminated correctly more than 65% of the time. Of the five remaining pairs (/ki/gi/, /ku/gu/, /pɑ/bɑ/, /tɑ/dɑ/ and /shɑ/jhɑ/), all involved a voiced/unvoiced distinction to be made

and all resulted in poor discrimination. In all of these cases the consonants appear in initial position and stop gap information is not present. In may be that better discrimination would result in vowel-consonant-vowel context such as /ɑbɑ/ɑpɑ/.

An obvious modification to the display would eliminate representing sound level via effective display intensity (line width and intensity) through the use

of constant line width and intensity. This may serve to make spectral decagon sequences of some consonants more perceptible and ultimately more discriminable and is a topic currently being examined

A final note concerns the lack of extended exposure to the display. Further training might include pointing out subtleties in the display to be observed, as well as additional exposure

to the display. A significant amount of training might improve a subject's ability to discriminate the various spectral patterns. This is suggested in the improvement in results from Test 1 (82% correct) to Test 3 (85% correct).

REFERENCES

Beranek, Leo L. (1947). "Design of speech systems," Proc. of the I. R. E. Sept. 1947: 365-380.

Byrne et al. (1994). "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108-2120.

Fisher, H. and J. Logemann (1971). "The Fisher-Logemann Test of Articulation Competence," The Riverside Publishing Company.

Liberman, A. M., P. Delattre, and F. S. Cooper (1957). "Some results of research on speech perception," J. Acoust. Soc. Am. **29**, 117-123.