The effect of spectral resolution on speech perception in a multi-talker babble background

Bom Jun Kwon and Christopher W. Turner

Department of Speech Pathology and Audiology, University of Iowa, Iowa City, IA 52242

Abstract: Does finer spectral resolution always provide an advantage for speech recognition? In this experiment, speech was divided into separate bands within which spectral information was removed by replacing the fine structure with noise; the temporal envelope cues of each band remained intact. Normal-hearing subjects identified consonants in /C/a for speech and babble processed with 1, 2, 4, and 8 bands, in addition to the unprocessed speech and babble, for a range of S/N ratios. For high S/N ratios our results agree with earlier results, in that performance improved with increases in the number of bands, and higher transmission of information was observed for voicing and manner than place of articulation. However, as S/N decreased, the advantage of better spectral resolution was reduced. This suggests that for speech in a competing babble background, increased spectral resolution not only increases the saliency of speech, but at the same time makes background babble a more effective distracter. Thus, the benefit provided by fine spectral resolution depends upon the task and S/N ratio.

INTRODUCTION

It has been demonstrated that speech information with only gross spectral resolution is sufficient in understanding speech since cochlear implantation with relatively few channels is quite successful. Van Tasell et al(1) and Turner et al(2) showed that even when spectral information was not available, the identification of consonants was above chance by taking advantage of only temporal information. Shannon et al(3) showed that with speech processed through only 4 bands normal listeners could identify phonemes quite well. This study deals with speech in a babble noise background; is 4 bands of resolution enough, and how does performance deteriorate with increasing of the level of noise babble? Similar questions were addressed by Fu(4) who used white noise as a masker, which will briefly be compared to the present results.

STIMULI AND SIGNAL PROCESSING

64 consonants(4 speakers, 16 for each) in /C/a context were used as stimuli. The partitioning scheme of each band was nearly logarithmic. Boundaries of bands were 150, 260, 450, 830, 1500, 2600, 4500, 8300, 15000 Hz, for 8-band processing, odd-numbered boundaries were used for 4-band, and 1500 Hz was the boundary for 2-band. The background noise was a babble recorded by 12 talkers. Stimuli were bandpass-filtered, multiplied by the speech-correlated noise(5) in order to make the spectrum in each band flat while the temporal envelope remained intact, then the bands were added together. 5 normal hearing subjects participated, and performance was tested across a range of S/N ratios for each band condition.

DATA AND RESULTS

Fig 1 shows the mean performances as a function of S/N ratio for each band condition. Fig 2 is plotted as a function of number of bands for each S/N ratio. Not surprisingly, as S/N ratio goes down, the score decreases in each band condition. Generally the more processing bands, the higher the score, although performances of 4- and 2-band do not show large differences. In contrast to the notion that 4-band processing provides enough speech information in a quiet background, it is shown that more bands are needed as S/N ratio decreases. However, it is also shown in these plots that the relationship between the number of bands and S/N ratio is not necessarily monotonic. In very low S/N ratio conditions, performance for the unprocessed speech was poorer than that at 8-band, which is against the general rule that better performance is consistently obtained with more bands.

WHICH CONSONANTS ARE DISTINGUISHABLE IN WHAT CONDITIONS?

Distinctive feature analysis showed that the voicing and manner information were transmitted relatively well even in conditions of low S/N ratios and small numbers of bands, compared to the place information, which is consistent with previous results(2-3) for speech in quiet. Once listeners identified the distinctive features (except for the place features), it was analyzed how well they could make correct identification of place information(Table 1). While the identification of stop consonants is generally easier than that of (a)fricatives, the minimum number of bands required to make the correct answer differs case by case. For example, listeners needed more than 4 bands in order to be able
to distinguish /m/ from /n/, while they needed at least 8 bands for /k/ vs. /l/. Listeners failed to distinguish between /f/ and /6/, the most confusable pair in the experiment, even at the 8-band condition.

![Graph showing performance as a function of S/N ratio](image1)

**FIGURE 1.** Mean performance as a function of S/N ratio with parameter of number of bands (iso-band curves).

![Graph showing performance as a function of number of bands](image2)

**FIGURE 2.** Mean performance as a function of number of bands for each S/N ratio (iso-S/N-ratio curves).

<table>
<thead>
<tr>
<th>group</th>
<th>p-t-k</th>
<th>b-d-g</th>
<th>m-n</th>
<th>f-θ-s</th>
<th>v-θ-z</th>
<th>k-t</th>
<th>z-ζ</th>
<th>f-θ</th>
</tr>
</thead>
<tbody>
<tr>
<td>unprocessed</td>
<td>98</td>
<td>97</td>
<td>99</td>
<td>86</td>
<td>94</td>
<td>98</td>
<td>100</td>
<td>86</td>
</tr>
<tr>
<td>8 bands</td>
<td>93</td>
<td>89</td>
<td>100</td>
<td>61</td>
<td>71</td>
<td>93</td>
<td>99</td>
<td>48*</td>
</tr>
<tr>
<td>4 bands</td>
<td>71</td>
<td>54</td>
<td>81</td>
<td>50</td>
<td>49*</td>
<td>63*</td>
<td>73*</td>
<td>57*</td>
</tr>
<tr>
<td>2 bands</td>
<td>69</td>
<td>64</td>
<td>69*</td>
<td>42*</td>
<td>56</td>
<td>66*</td>
<td>60*</td>
<td>49*</td>
</tr>
<tr>
<td>1 bands</td>
<td>42*</td>
<td>44*</td>
<td>51*</td>
<td>42*</td>
<td>36*</td>
<td>62*</td>
<td>51*</td>
<td>52*</td>
</tr>
</tbody>
</table>

**DISCUSSION**

The counterintuitive phenomenon of better performance at 8-band than the unprocessed speech at poor S/N ratios was not observed in Fu's study (4), which implies that a competing speech from multi-talker source (so-called babble noise), as a masker, works differently than the white noise does. In other words, unprocessed competing background speech may distract the listener's attention more than 8-band. In Fu's experiment where white noise was used as a masker, the number of processing bands of noise was not a critical factor, producing essentially parallel performance curves for each condition of S/N ratio, which is not the case for the babble noise as seen in fig 2.

In conclusion, it is generally true that the more processing bands, and the higher the S/N ratio, the better performance listeners can expect, with some exceptions. The question of how many bands we need in understanding speech can be specifically answered only on the basis of contexts and S/N ratios.

**ACKNOWLEDGMENT**

This study was supported in part by NIDCD.

**REFERENCES**