Integration Efficiency for Speech Understanding Within and Across Sensory Modalities

Ken W. Grant[1] and Steven Greenberg[2]
Walter Reed Army Medical Center, Army Audiology and Speech Center, Washington, DC 20307-5001, [grant@tidalwave.net]
International Computer Science Institute[2] 1947 Center Street, Berkeley, CA, 94704, [steveng@icsi.berkeley.edu]

BACKGROUND AND SUMMARY

The ability to understand speech relies in part on our capacity to integrate spectro-temporal information from different frequency regions of the speech spectrum. This is especially true for multichannel hearing aids and cochlear implants where speech information is divided into separate spectral bands and subjected to different types and degrees of signal processing. The time frame over which this integration occurs may reflect different levels of processing: one which operates over relatively short time spans (ca. 50 ms) and is involved in detailed phonetic analysis of the signal and another operating at a more abstract level of syllable length units where the time span is about 250 ms. Because spoken conversation normally involves face-to-face interaction resulting in both auditory and visual information being used to decode the speech signal, it is important to determine whether the time frames for spectro-temporal integration and the efficiency at which integration proceeds is the same for auditory-only speech presentations and for auditory-visual speech presentations, and whether integration efficiency is compromised by hearing impairment. In this paper, we discuss the relative efficiency of these processes by comparing the ability of normal-hearing and hearing-impaired subjects to integrate narrow bands of speech when presented under auditory-only and auditory-visual conditions. Specifically, nonsense syllables (/a/-consonant-/a/) spoken by a female talker were filtered into two or four 1/3-octave wide bands of speech using an FIR filter whose slope exceeded 100 dB/octave. The four-band auditory condition consisted of filter passbands of 298-375 Hz, 750-945 Hz, 1890-2381 Hz, and 4762-6000 Hz presented concurrently. Two additional auditory conditions were made by combining either bands 1 and 4 (the two fringe bands) or bands 2 and 3 (the two middle bands). For auditory-visual conditions, subjects viewed a video image of the talker presented synchronously with either the two fringe bands or the two middle bands. A sixth condition consisting of visual-only speech recognition was also tested. Integration efficiency was determined by using Braida's Prelabeling Model of Integration (Braida, 1991) to predict subject responses in the four-band auditory condition as well as for the two auditory-visual conditions. Comparisons of within-modality (auditory only) and across-modality (auditory-visual) integration efficiency showed that both normal-hearing and hearing-impaired subjects had little trouble integrating auditory and visual information but that hearing-impaired listeners demonstrated problems integrating spectral information across widely separated auditory frequency channels.

DESIGN

- STIMULI:

  Consonant recognition (vCv) using spectrally sparse acoustic stimuli consisting of 2 or 4 narrow spectral slits (1/3-octave) separated by at least one octave. The consonant set included /b,p,g,k,d,t,m,n,f,v,θ,ð,s,z,ʃ,ʒ,tʃ,dʒ/ surrounded by the vowel /a/.

  - `IOOI = Band 1 (298-375 Hz) + Band 4 (4762-6000 Hz)`
  - `OIIO = Band 2 (750-945 Hz) + Band 3 (1890-2381 Hz)`
  - `IIII = Band 1 + Band 2 + Band 3 + Band 4`

- THREE PRESENTATION MODES:
  - **Auditory (IOOI, OIIO, IIII)**
  - **Visual**
  - **Auditory-Visual (IOOI, OIIO)**

- SUBJECTS:
  - **4 Normal Hearing, 4 Hearing Impaired**

	RIGHT EAR									LEFT EAR								
	250	500	1000	1500	2000	3000	4000	6000	8000	250	500	1000	1500	2000	3000	4000	6000	8000
JJL	5	10	5	15	15	15	0	5	25	5	10	0	10	10	15	5	5	5
JEN	15	15	10	10	15	20	15	15	20	10	5	5	10	10	15	10	20	15
JOC	5	0	5	5	0	5	10	5	20	10	5	5	5	0	0	10	10	10
MTC	5	5	5	5	15	15	5	10	10	5	0	0	5	20	10	5	0	15
JES	20	20	20	15	35	55	65	70	70	20	15	15	35	40	95	95	100	90
DJF	20	25	40	45	55	70	80	80	80	20	20	25	40	35	60	65	80	80
DGW	10	25	50	50	45	50	60	105	120	20	15	45	45	40	55	60	75	75
ECC	40	40	45	70	70	75	75	80	75	30	35	45	65	70	70	70	75	70

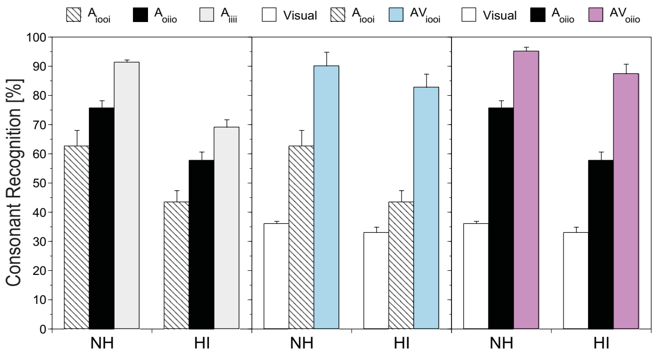

FIGURE 1. Average consonant recognition scores for the 6 presentation conditions. Data are for 4 NH and 4 HI subjects. Note that the 4-band audio scores for HI subjects (A_{iiii}) are significantly lower than AV scores (Av_{iooi} or Av_{oiio}).

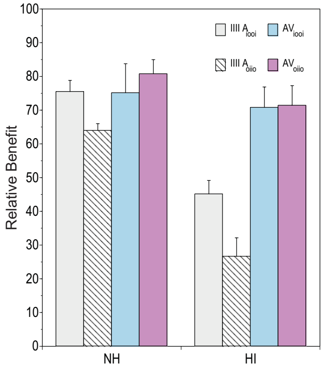

FIGURE 2. Auditory and auditory-visual relative benefit (A-RB, AV-RB) scores obtained by combining information across auditory and visual channels. Relative benefit is defined as the percent correct improvement relative to the total possible improvement given the recognition performance obtained in some reference condition (after Sumby and Pollack, 1954). For both Auditory and Auditory-Visual conditions, two reference conditions were considered (i.e., A_{iooi} and A_{oiio}). Thus,

$$
\begin{aligned}
A\text{-}RB_{iooi} &= (A_{iiii}-A_{iooi})/(100-A_{iooi})\\
A\text{-}RB_{oiio} &= (A_{iiii}-A_{oiio})/(100-A_{oiio})\\
AV\text{-}RB_{iooi} &= (AV_{iooi}-A_{iooi})/(100-A_{iooi})\\
AV\text{-}RB_{oiio} &= (AV_{oiio}-A_{oiio})/(100-A_{oiio})
\end{aligned}
$$

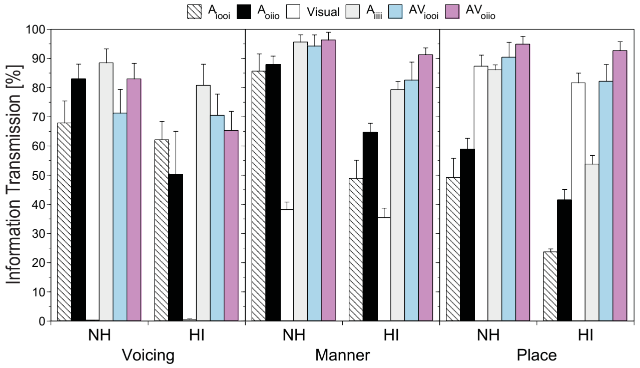

FIGURE 3. Percent information transmitted in the different auditory and auditory-visual test conditions for the speech features voicing, manner-of-articulation, and place-of-articulation. Hearing-impaired subjects showed substantially lower than normal place and manner scores for all auditory conditions. More similar performance across the two groups of subjects were seen in the auditory-visual conditions, presumably because transmission of place cues were essentially recovered through the integration of hearing and speechreading.

The poorer performance in relative benefit and feature transmission for HI subjects, most notably in the three auditory test conditions can be due to poor hearing (i.e., a lack of information), poor processing of auditory information (i.e., a problem integrating all available information), or some combination of these two. To separate the effects of degraded or absent auditory "*information*" from poor "*information processing*" we analyzed each subject's data using an optimum processor model of integration (Braida, 1991). The model predicts both overall accuracy and error patterns in the combined test conditions (i.e., A_{iiii}, AV_{iooi}, and AV_{oiio}) from the constituent confusion matrices (i..e., A_{iooi}, A_{oiio}, and Visual).

MODEL FITS

- Optimum-procesor models (Braida, 1991) describes consonant reception in terms of a multidimensional extension of the theory of signal detection.
- Confusion matrices are subjected to Multidimensional scaling, resulting in estimates of stimulus centers and response centers.
- In general, the distance between two stimulus centers, d'(i,j), determines the observer's ability to distinguish between the two consonant, S_i and S_j.
- For multichannel presentations (e.g., A_{IIII}, AV_{IOOI}, and AV_{OIIO}, integration assumes that the cue densities in the multichannel condition are the "Cartesian products" of the densities corresponding to the separate test conditions (e.g., A_{IOOI}, A_{OIIO}, and Visual). Thus, cues are combined optimally and there is no interference (e.g., masking or distraction) across conditions.

$$
\begin{aligned}
d'_{A_{iiii}} &= \sqrt{d'_{A_{iooi}}(i,j)^2 + d'_{A_{oiio}}(i,j)^2}\\
d'_{AV_{iooi}} &= \sqrt{d'_{A_{iooi}}(i,j)^2 + d'_{V}(i,j)^2}\\
d'_{AV_{oiio}} &= \sqrt{d'_{A_{oiio}}(i,j)^2 + d'_{V}(i,j)^2}
\end{aligned}
$$

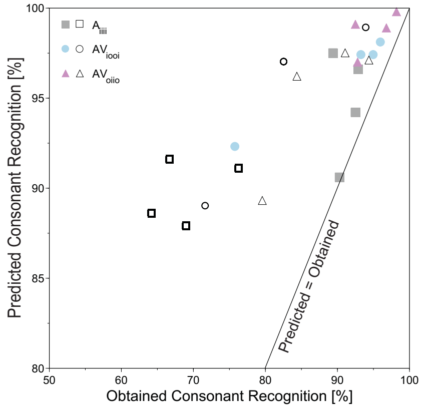

FIGURE 4. Obtained and predicted consonant recognition scores for A_{iiii}, AV_{iooi}, and AV_{oiio} test conditions. Scores for HI subjects are shown by open symbols. The line indicates the region where obtained and predicted scores are equal. All scores to the left of this line indicate where subjects fell short of optimal performance. Note that the largest discrepancies between predicted and obtained scores are for HI subjects in the auditory listening conditions (open square symbols). Thus, for these subjects, the mutual information extracted from the A_{iooi} and A_{oiio} conditions was sufficient to predict a much higher A_{iiii} score than obtained (by approximately 20 percentage points). This suggests that for these HI subjects, information across spectral channels was not integrated optimally.

To estimate the integration efficiency for NH and HI subjects for the different multichannel presentation conditions (i.e., A_{iiii}, Av_{iooi}, and Av_{oiio}), we calculated the ratio (in percent) between the obtained and predicted consonant recognition scores.

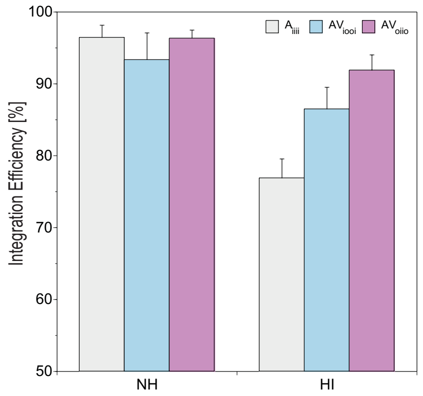

FIGURE 5. Integration efficiency (IE) for NH and HI subjects. For NH subjects, integration of speech information across auditory and visual modalities and across auditory spectral channels is equally robust. For HI subjects however, the integration efficiency for combining spectral information from widely spaced frequency bands is greatly impaired. The difference in IE measures between NH and HI subjects was significant only for the A_{iiii} condition.

DISCUSSION

The results from modeling fits using optimum processor models of integration show that HI subjects have difficulty combining information across widely spaced frequency channels independent of their ability to extract information from the separate 2-band conditions. One possible explanation for this result is offered below. Because HI subjects had extensive high frequency hearing loss, we might assume that the audibility of band 4 was severely compromised. When band 4 was presented in combination with band 1, there was no signal energy in frequency channels just below 4700 Hz. It is possible that information contained in band 4 could have been partly obtained by off-frequency listening (using less impaired auditory channels between 2-3 kHz). However, when all four bands were presented, these same off-frequency channels were now pre-occupied with their own input signals (bands 2 and 3) and might have been less able to extract the same degree of relevant information from band 4. Thus, the A_{iooi} condition may provide useful high-frequency information only when there is no signal energy present in the adjacent lower-frequency regions. This hypothetical scenario is not too unlike what might be inferred from the studies by Hogan and Turner (1998) and Doherty and Turner (1996) showing that HI subjects have difficulty extracting and integrating high-frequency speech information when presented as part of a broadband speech signal. This idea could be tested by comparing A_{iiii} and A_{iiio} conditions (where we predict that the A_{iiio} would be as good as the A_{iiii}) as well as by comparing A_{iooi} and A_{iooo} conditions (where we predict that the A_{iooi} condition would be better than the A_{iooo} condition).

REFERENCES

Braida, L.D. (1991). "Crossmodal integration in the identification of consonant segments," Quarterly J. Exp. Psych. **43**, 647-677.

Doherty, K.A., and Turner, C.W. (1996). "Use of a correlational method to estimate a listener's weighting function for speech," J. Acoust. Soc. Am. **100**, 3769-3773.

Hogan, C.A., and Turner, C.W. (1998). "High-frequency audibility: Benefits for hearing-impaired listeners," J. Acoust. Soc. Am. **104**, 432-441.

Sumby, W. H., and Pollack, I. (1954). "Visual contribution to speech intelligibility in noise," J. Acoust. Soc. Am. **26**, 212-215.