Acta Oto-Laryngologica
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713690940

The number of spectral channels required for speech recognition depends on the difficulty of the listening situation

Robert V. Shannon; Qian-Jie Fu; John Galvin Iii

Online Publication Date: 01 April 2004
To cite this Article: Shannon, Robert V., Fu, Qian-Jie and Galvin Iii, John (2004) 'The number of spectral channels required for speech recognition depends on the difficulty of the listening situation', Acta Oto-Laryngologica, 124:3, 50 - 54
To link to this article: DOI: 10.1080/03655230410017562
URL: http://dx.doi.org/10.1080/03655230410017562

Acta Otolaryngol 2004; Suppl 552: 50–54

Taylor & Francis healthsciences

The Number of Spectral Channels Required for Speech Recognition Depends on the Difficulty of the Listening Situation

ROBERT V. SHANNON, QIAN-JIE FU and JOHN GALVIN III

From the House Ear Institute, Los Angeles, CA, USA

Shannon RV, Fu Q-J, Galvin J III. *The number of spectral channels required for speech recognition depends on the difficulty of the listening situation*. Acta Otolaryngol 2004 Suppl 552: 50–54.

Cochlear implants provide a limited number of electrodes, each of which represents a channel of spectral information. Studies have shown that implant recipients are not receiving all of the information from the channels presented to their implant. The present paper provides a quantitative framework for evaluating how many spectral channels of information are necessary for speech recognition. Speech and melody recognition data from previous studies with cochlear implant simulations are compared as a function of the number of spectral channels of information. A quantitative model is applied to the results. Speech recognition performance increases as the number of spectral channels increases. A sigmoid function best describes this increase when plotted as a function of the log number of channels. As speech materials become more difficult, the function shifts to the right, indicating that more spectral channels of information are required. A model proposed by Plomp provides a single index to relate the difficulty of the task to the number of spectral channels needed for moderate recognition performance. In conclusion, simple sentence recognition in quiet can be achieved with only 3–4 channels of spectral information, while more complex materials can require 30 or more channels for an equivalent level of performance. The proposed model provides a single index that not only quantifies the number of functional channels in a cochlear implant, but also predicts the level of performance for different listening tasks.

INTRODUCTION

Cochlear implants (CIs) reduce the spectral information in speech to streams of modulated electrical pulses on a few electrodes. To design improved signal processing for CIs it is important to understand the relation between the number of spectral channels and speech recognition. Shannon et al. (1) demonstrated that a high level of speech recognition was possible with as few as four spectral channels of information. This result was obtained with simple sentence materials and in quiet listening conditions. More spectral information is required for more difficult speech materials and/or more difficult listening conditions. There is at present no quantitative model that specifies the number of spectral channels necessary for speech recognition under different listening conditions.

The present manuscript presents a meta-analysis of speech and music recognition data obtained in previous studies with a variety of test materials and listening conditions. We propose that the relative difficulty of the materials and listening conditions can be expressed as a single numeric quantity that also indicates the effective number of channels of spectral information available to the listener.

Speech recognition with reduced spectral information has been widely studied. Hill et al. (2) measured speech recognition as a function of the spectral resolution, using a sinusoidal vocoder. They divided the speech spectrum into an equal number of frequency bands, evenly distributed in terms of log frequency. The time-varying amplitude envelope was extracted from each analysis band and used to modulate sinusoidal carriers; the frequency of each sinusoid matched the center frequency of the corresponding analysis band. In this way, the spectral information was reduced to a small number of modulated sinusoids – all other spectral information was removed. Hill et al. found that consonant and vowel recognition reached asymptotic performance levels at six to eight channels, even when the temporal envelope for each channel was low-pass filtered below 100 Hz. These results demonstrated that good speech recognition was possible with relatively few spectral channels and relatively slow temporal fluctuations. Somewhat similar results had been found in early work with speech vocoders (3, 4).

SPEECH RECOGNITION AS A FUNCTION OF THE NUMBER OF SPECTRAL CHANNELS

For CI speech processor design, it is important to understand the importance of the number spectral channels for speech recognition. Shannon et al. (1) developed a noise-band vocoder to simulate CI speech processing for normal hearing (NH) listeners. Shannon et al. used processing similar to that of Hill et al., except that noise bands were used as carrier signals rather than sinusoids. Shannon et al. found that as few as four channels of spectral information produced high levels of speech recognition. Dorman and colleagues (5–7) replicated the results of Shannon et al. using processors with both noise and sinusoidal carriers. They also found that excellent speech recognition could be achieved with four to six channels of spectral information and that sinusoidal and noise-

 ISSN 0365-5237 DOI 10.1080/03655230410017562

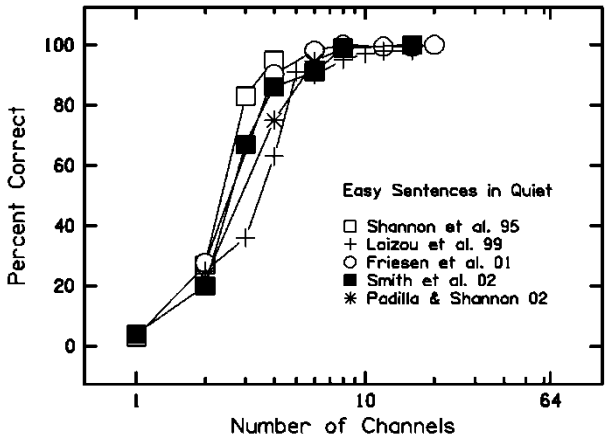

Fig. 1. Percent correct recognition of simple sentences in quiet as a function of the number of spectral channels. The acoustic input frequency range (200 to 6 kHz) was divided into frequency analysis bands that corresponded with equal distances along the basilar membrane. The envelope signal was extracted from each band and used to modulate a carrier band of noise of the same frequency range as the analysis band. The modulated noise bands there then combined and presented to the listeners either in a sound field or through headphones.

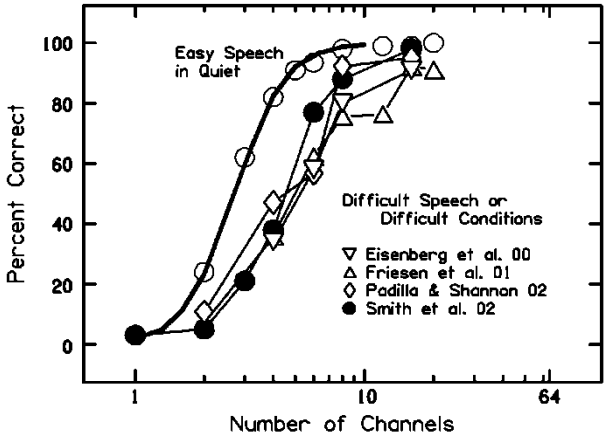

Fig. 2. Percent correct recognition of simple sentences as a function of the number of spectral channels by: 5–7-year-old children (10); in the presence of competing noise (5 dB SNR) (8); in the presence of competing speech (11); non-native listeners who learned English before the age of 5 years (12). In each case, the speech and noise combination was processed by a noise-band vocoder (1) with differing numbers of spectral channels. The thick solid line represents the average data from Fig. 1, fitted by a sigmoid function ($r^2 = 0.99$).

band carriers produced similar results. Several studies have measured sentence recognition as a function of the spectral resolution using noise-band vocoders (Fig. 1). These studies all used relatively simple sentence materials and tested NH listeners in quiet listening conditions. Note that the results are quite consistent across studies, showing that only three to four channels of spectral information were needed to produce sentence recognition better than 50% correct; six spectral channels were sufficient to produce near-perfect sentence recognition. These results confirm that speech recognition does not require fine spectral resolution under ideal listening conditions. The results also demonstrate how CIs are able to provide good speech understanding using only a few electrodes to represent the speech spectral pattern.

The results shown in Fig. 1 were obtained using relatively easy test materials, presented in quiet. When listening to more difficult test materials, or to simple materials presented in difficult listening conditions, more spectral channels are necessary to achieve the same level of recognition (Fig. 2). (The average data from Fig. 1 have been fitted with a simple sigmoid function ($r^2 = 0.99$) and are represented by the thick solid line in Fig. 2.) The results from Friesen et al. (8) were obtained for simple sentence materials presented at a +5 dB signal-to-noise ratio (SNR), and showed that more spectral channels were required in noise to achieve similar performance in quiet. Fu et al. (9) showed a similar shift for phoneme recognition in noise. Results from Eisenberg et al. (10) were obtained in 5–7-year-old children using simple sentences specially designed for children. These results suggest that compared with adult NH listeners young children have not developed sufficient speech pattern recognition skills to tolerate reduced spectral information; thus, children require more spectral channels than adults to obtain similar speech recognition results. Smith et al. (11) tested sentence recognition in the presence of a competing masker sentence, as a function of the spectral resolution. Padilla and Shannon (12) tested speech recognition with bilingual listeners, as a function of spectral resolution, and found that non-native listeners (even those who were fully bilingual) required more spectral channels than native, monolingual listeners. Taken together, these studies demonstrate that, regardless of the source of test difficulty (noise, competing speech, inexperience in either a primary or second language), 5–6 spectral channels were needed to produce speech recognition better than 50% correct; high levels of recognition required more than 10 spectral channels.

Even more spectral channels are required for recognition in more complex tasks or more difficult listening conditions (Fig. 3). (The averaged results from Figs. 1 and 2 were fitted by a simple sigmoid function with fixed slope ($r^2 = 0.98$) and are represented by thick solid lines in Fig. 3.) The data from Padilla and Shannon (12) show sentence recognition in noise (10 dB SNR) by bilingual listeners who learned English as a second language after the age of 18 years. These results show that more spectral channels were required to overcome the combined

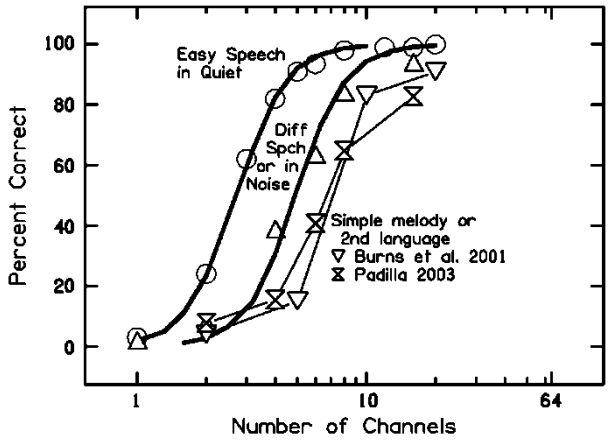

Fig. 3. Percent correct recognition of simple melodies (13) or sentences by non-native listeners who have less experience with English language (20) as a function of the number of spectral bands in a noise-band vocoder. Melodies were simple familiar tunes with the rhythm and timing information removed (isochronous). English sentence recognition in the presence of noise (10 dB SNR) was measured for native Spanish-speaking listeners who learned English after the age of 18 years. The thick solid lines represent the average data from Figs. 1 and 2, each fitted by a sigmoid function with a fixed slope ($r^2 = 0.98$).

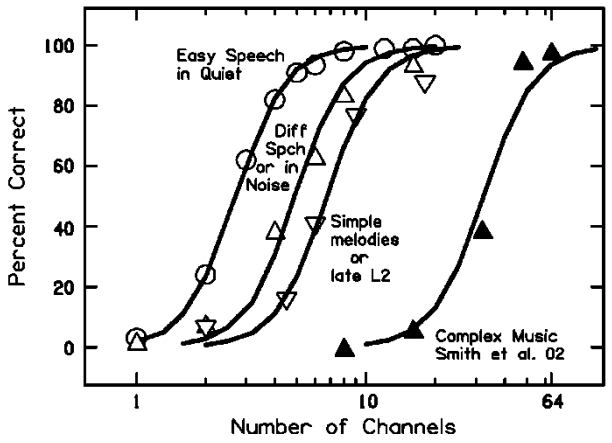

Fig. 4. Percent correct recognition of music masked by competing music (music–music chimeras (11)). Average data from Figs. 1–3 are plotted and fitted with sigmoid functions with the same slope. The thick solid lines represent the average data from Figs. 1–3, each fitted by a sigmoid function with a fixed slope ($r^2 = 0.97$). Note that all functions are well fitted by sigmoid functions with the same slope (60%/doubling).

effects of language inexperience and difficult listening conditions. Data from Burns et al. (13) show musical melody recognition as a function of spectral resolution. In this study, simple, familiar, isochronous (no rhythm cues) melodies were processed through a noise-band vocoder; subjects were asked to identify the melody. These two studies demonstrated that 7–8 spectral channels were required to produce more than 50% correct recognition; more than 15 spectral channels were required for high recognition levels.

When the listening task involves melody recognition in the presence of competing melodies (music–music chimeras from Smith et al. (11)), even more spectral channels are needed (Fig. 4). (The averaged results from Figs. 1–3 were fitted by a simple sigmoid function with fixed slope ($r^2 = 0.97$) and are represented by the thick solid lines in Fig. 4.) In the Smith et al. study, 50% correct recognition of the target melody was not achieved until more than 32 channels of spectral information were provided. While the melodies may have been recognizable with 48 channels, subjects reported that the sound quality was poor, suggesting that good sound quality for music requires more than 48 spectral channels.

Overall, it is apparent that more spectral information is necessary as the listening task becomes more difficult. Simple recognition tasks can be accomplished with only a few spectral channels, while complex perceptual tasks require many channels. Plotting the sentence recognition as a function of the number of channels, we were able to fit the data quite well with simple sigmoid functions (that increase from 0% correct with few channels to 100% correct with many channels). The slope and shape of the functions appear to be constant when plotted against the number of channels (log scale). In Fig. 4, the slope of each curve is 60 percentage points/doubling; i.e. doubling the number of channels improved recognition performance by 60 percentage points. As the difficulty of the test materials or listening conditions increases, the functions simply shift to the right – showing that more channels are required to maintain a given level of performance. Note that the simple sigmoid curves fit the data well for each level of difficulty. The residual error (r^2) in the fits for each curve were better than 0.97 for a sigmoid function with a single free parameter (curve horizontal location). We have previously shown (14) that the number of spectral channels needed for speech recognition could be simply related to a spectral distortion factor in a model proposed by Plomp (15). Consider how the data of Figs. 1–4 might fit within Plomp's framework.

PLOMP'S SPEECH RECEPTION THRESHOLD (SRT) MODEL

Plomp (15) proposed a model of hearing loss in which conductive hearing loss was modeled as an attenuation factor (A) and all other problems were modeled as a distortion factor (D). While attenuation could be compensated for by amplification, the distortion factor resulted in a constant deficit in the SNR required to reach a criterion performance level for sentence recognition. For a wide cross-section of hearing-impaired (HI) listeners, Plomp showed that

D could range from 3 to 5 dB. HI listeners would therefore require a 3–5 dB improvement in SNR to achieve the same performance levels found with NH listeners (for who, by definition, $D = 0$ dB). While a D factor of 3–5 dB may seem small, consider that sentence intelligibility shifts 10–20 percentage points for each dB change in the SNR; 3–5 dB of impairment may reduce sentence recognition by 30–100 percentage points, depending on the SNR.

The effects of Plomp's D factor may be interpreted in terms of the reduced frequency selectivity that often accompanies hearing loss. Even if the loss in sensitivity is compensated by amplification, the loss of spectral resolution persists. A loss of spectral resolution would have the effect of smearing the spectral information, which results in poorer speech recognition (14, 16–18). Shannon and Fu (19) suggested that Plomp's model may also be applied to CI users; the limited number of spectral channels available to a CI listener can be thought of in terms of spectral smearing. The speech recognition of a CI listener who only makes use of six spectral channels may be compared with that of a HI listener with a high D factor – reflecting a broadening of the peripheral analysis filters.

To measure the effect of spectral smearing in terms of Plomp's model, we compared data from two studies (8, 20) that measured sentence recognition by NH subjects listening with a noise-band vocoder (1). The noise-band vocoder simulates the type of signal processing that occurs in a CI and allows the spectral resolution to be varied systematically. In a noise-band vocoder, the signal is first divided into spectral bands via band-pass filtering. The analysis filters correspond to equal distances along the basilar membrane according to Greenwood's (21) formula. The envelope is extracted from each band by half-wave rectification and low-pass filtering at 160 Hz. The envelope from each analysis band is then used to modulate a band of noise whose bandwidth is identical to that of the analysis filter. The two studies (8, 20) measured sentence recognition as a function of the number of spectral channels and as a function of SNR. For each spectral resolution condition, the SRT – the SNR required to produce 50% whole sentence recognition – was normalized to the SRT for unprocessed speech. For NH subjects listening to unprocessed speech, $D = 0$ by definition (15). Also, because NH listeners are estimated to have 64 available spectral channels, the point at which $D = 0$ was assumed to be 64 channels. The D factor for all other spectral resolution conditions was defined as the shift in SRT relative to the unprocessed speech condition.

The results show that D changed linearly as a function of the number of spectral channels (log scale) (Fig. 5). The slope of this function was approximately

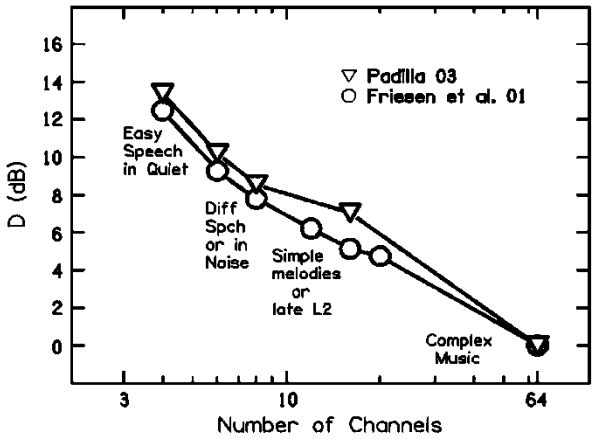

Fig. 5. Increase in the speech reception threshold for simple sentences as a function of the number of spectral channels from two studies (8, 20). Note that the threshold increased 3 dB for each halving of the number of channels.

-10 dB/decade (or, -3 dB/doubling of the number of channels); thus, D increased by 3 dB as the number of spectral bands was reduced by half. Padilla (20) showed that D increased as bilingual listeners' experience with English as a second language decreased. Less familiarity with a second language (increased difficulty with test materials) had the same effect as a loss in spectral resolution.

We suggest that the D factor in Plomp's model can be used to represent not only the degree of spectral distortion or smearing experienced by a listener, but also the degree of a listener's unfamiliarity with the test materials, as well as the degree of complexity of the task itself. In most speech and music recognition tasks, performance improves as the signal is made clearer. The signal can be made clearer by increasing the spectral resolution, or by improving the SNR. People with less language experience (children with their native language; bilinguals with their second language) require a better signal to achieve the same recognition as adults listening to their native language. In terms of a modified Plomp model, reduced language experience is equivalent to increased distortion (D). Similarly, adults require more spectral resolution (decrease in D) for difficult listening tasks to achieve the same level of performance achieved in simple tasks. Using Plomp's terms, the difficulty of a task may be characterized quantitatively as the shift in D factor relative to a task that can be performed at a criterion level (e.g. 50% correct recognition). Because D is linearly related to the log number of spectral bands, D can be used as a metric for both listener experience and task difficulty. The linear relation between D and log number of channels (Fig. 5) illustrates the difficulty HI and CI listeners may face when confronted with complex listening tasks, i.e.

more channels (and lower value of D) are required for more complex listening tasks. As a listener's hearing loss progresses and the D factor increases, music perception, speech recognition in noise, recognition of a second language all become more difficult, requiring more spectral channels than are available. The labels below the curves in Fig. 5 indicate the number of spectral channels necessary to achieve better than 50% recognition in the various conditions of Figs. 1–4.

CONCLUSION

The present meta-analysis quantifies the relation between the number of spectral channels and good recognition performance in difficult listening conditions. Although modern CIs provide as many as 22 electrodes in the cochlea, studies have shown that implant users receive only 4–8 channels of spectral information (8, 22). It is important to understand why implant listeners are not able to make use of all the spectral information provided by the implant. Significant performance improvements may be possible with existing devices if speech-processing strategies can be designed to enhance the spectral information already provided by the implanted electrodes. The D metric could be used to quantify the estimated number of channels received by a CI listener as well as an index for the degree of recognition difficulty permitted by different amounts of spectral resolution. If future speech processing strategies are able to reduce D by only a few dB, CI users' speech recognition will be largely improved.

REFERENCES

1. Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science 1995; 270: 303–4.
2. Hill FJ, McRae LP, McClellan RP. Speech recognition as a function of channel capacity in a discrete set of channels. J Acoust Soc Am 1968; 44: 13–8.
3. Dudley H. The Vocoder. Bell Lab Rec 1939; 17: 122–6.
4. Fletcher H, Galt RH. The perception of speech and its relation to telephony. J Acoust Soc Am 1950; 22: 89–151.
5. Dorman MF, Loizou PC, Rainey D. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. J Acoust Soc Am 1997; 102: 2403–11.
6. Dorman MF, Loizou PC, Fitzke J, Tu Z. The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. J Acoust Soc Am 1998; 104: 3583–5.
7. Loizou PC, Dorman MF, Tu Z. On the number of channels needed to understand speech. J Acoust Soc Am 1999; 106: 2097–103.
8. Friesen L, Shannon RV, Baskent D, Wang X. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J Acoust Soc Am 2001; 110: 1150–63.
9. Fu QJ, Shannon RV, Wang X. Effects of noise and spectral resolution on vowel and consonant recognition: acoustic and electric hearing. J Acoust Soc Am 1998; 104: 3586–96.
10. Eisenberg L, Shannon RV, Martinez AS, Wygonski J, Boothroyd A. Speech recognition with reduced spectral cues as a function of age. J Acoust Soc Am 2000; 107: 2704–10.
11. Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. Nature 2002; 416: 87–90.
12. Padilla M, Shannon RV. Could a lack of experience with a second language be modeled as a hearing loss? J Acoust Soc Am 2002; 112: 2385.
13. Burns EM, Sanborn ES, Shannon RV, Fu QJ. Perception of familiar melodies by implant users. Proceedings of the Conference on Implantable Auditory Prostheses, Pacific Grove, CA; 19–24 August 2001: 81.
14. Boothroyd A, Mulhearn B, Gong J, Ostroff J. Effects of spectral smearing on phoneme and word recognition. J Acoust Soc Am 1996; 100: 1807–18.
15. Plomp R. A signal-to-noise ratio model for the speech reception threshold of the hearing impaired. J Speech Lang Hear Res 1986; 29: 146–54.
16. Shannon RV, Zen FG, Wygonski J. Speech recognition with altered spectral distribution of envelope cues. J Acoust Soc Am 1998; 104: 2467–76.
17. ter Keurs M, Festen JM, Plomp R. Effect of spectral envelope smearing on speech reception 1. J Acoust Soc Am 1992; 91: 2872–80.
18. ter Keurs M, Festen JM, Plomp R. Effect of spectral envelope smearing on speech reception 2. J Acoust Soc Am 1993; 93: 1547–52.
19. Shannon RV, Fu QJ. Contributions of spectral resolution to speech recognition in noise. J Acoust Soc Am 1999; 105: 1238.
20. Padilla M. English phoneme and word recognition by nonnative English speakers as a function of spectral resolution and English experience [PhD dissertation]. Los Angeles: University of Southern California; 2003.
21. Greenwood DD. A cochlear frequency-position function for several species – 29 years later. J Acoust Soc Am 1990; 87: 2592–605.
22. Fishman K, Shannon RV, Slattery WH. Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. J Speech Hear Res 1997; 40: 1201–15.

Address for correspondence:
Robert V. Shannon, PhD
House Ear Institute
2100 W. Third St.
Los Angeles
CA 90057
USA
Tel: +1 213 353 7020
Fax: +1 213 413 0950
E-mail: Shannon@hei.org