JOBNAME: ASHA-AJA PAGE: 1 SESS: 4 OUTPUT: Mon Dec 16 16:51:39 2002
/balt6/io-aja/io-aja/io0202/io0069-02a

AQ: A

The Relative Importance of Amplitude, Temporal, and Spectral Cues for Cochlear Implant Processor Design

Robert V. Shannon
House Ear Institute, Los Angeles, CA

Speech understanding with cochlear implants has improved steadily over the last 25 years, and the success of implants has provided a powerful tool for understanding speech recognition in general. Comparing speech recognition in normal-hearing listeners and in cochlear-implant listeners has revealed many important lessons about the types of information necessary for good speech recognition—and some of the lessons are surprising. This paper presents a summary of speech perception research over the last 25 years with cochlear-implant and normal-hearing listeners. As long as the speech is audible, the even relatively severe amplitude distortion has only a mild effect on intelligibility. Temporal cues only appear to be useful for speech intelligibility up to about 20 Hz. Whereas temporal information above 20 Hz may contribute to improved quality, it contributes little to speech understanding. In contrast, the quantity and quality of spectral information appears to be a critical factor for speech understanding. Only four spectral "channels" of information can produce good speech understanding, but more channels are required for difficult listening situations. Speech understanding is sensitive to the placement of spectral information along the cochlea. In prosthetic devices, in which the spectral information can be delivered to any cochlear location, it is critical to present spectral information to the normal acoustic tonotopic location for that information. If there is a shift or distortion of 2 to 3 mm between frequency and cochlear place, speech recognition is decreased dramatically.

Further improvements in cochlear implant speech processor design will require a better understanding of the relative roles of ear and brain in speech recognition. The success of cochlear implants demonstrates that not all of the detailed information from the cochlea is necessary for speech recognition. But which speech cues are most important for speech pattern recognition?

Modern cochlear implants have multiple electrodes that are inserted into the scala tympani 20 to 30 mm, with electrodes covering a range corresponding to normal acoustic frequencies of about 500 to 5000 Hz (Greenwood, 1990). Cochlear implant speech processors typically filter speech into multiple frequency bands, and the processed outputs from the frequency bands are presented to the electrodes to recreate the normal tonotopic distribution of information along the cochlea. Within each "channel," the low-frequency temporal envelope is extracted. After mapping the acoustic amplitude range into the smaller electrical range of each electrode, this low-frequency envelope from each band is used to modulate a high-frequency train of electrical pulses presented to the electrodes spaced along the tonotopic dimension of the cochlea.

The audiologist typically measures threshold levels and comfortable loudness levels on each electrode and sets the amplitude mapping function so that the electrical signal is audible and does not produce overly loud sounds. But most other parameters of the implant speech processor are either fixed by the manufacturer or are not ordinarily adjusted by the audiologist. To fit a speech processor optimally on each implant patient, it is important to understand which processor parameters are most important and how errors in their adjustment might affect speech understanding. This paper will review recent research studies on the relative importance of amplitude, temporal, and spectral cues in speech recognition.

Amplitude Cues

It has long been known that speech recognition is highly resistant to amplitude degradation, especially when the full spectral resolving power of the normal ear is available (e.g., Licklider & Pollack, 1948). Recent studies have shown that even with limited spectral resolution, speech recognition is relatively unaffected by amplitude nonlinearities, quantization, and peak and center clipping. As long as the speech information is audible, the relative amplitude cues are of only secondary importance. Fu and Shannon (1998) measured consonant and vowel recognition in normal-hearing (NH) and cochlear-implant (CI) listeners as the amplitude information was altered according to a power law. In a power law mapping, the output amplitude is equal to the input amplitude raised to a power. Power law exponents less than 1 produce compression of amplitude,

Orig. Op.	OPERATOR:	Session	PROOF:	PE's:	AA's:	COMMENTS	ARTNO:
1st disk, 2nd mjr	taylormp	4					0009

whereas exponents greater than 1 produce expansion of amplitude. They found the best phoneme recognition at the exponent that produced the appropriate loudness mapping function (1.0 for NH, 0.2 for CI). Amplitude mapping exponents greater or less than these values produced a drop in vowel and consonant recognition. However, the drop in performance was quite modest, with only a 10% to 15% drop even for exponents that were in error by a factor of 2, that is, a large error in loudness mapping. This pattern of results was observed even when the spectral resolution was reduced to only four broad bands. Almost no reduction in phoneme or speech recognition is observed if amplitude mapping is in error by a factor of 2 when full spectral information is available.

Two recent studies (Loizou, Dorman, Poroy, & Spahr, 2000; Shannon, Fu, Wang, Galvin, & Wygonski, 2001) reduced the number of amplitude steps by quantizing the output of a speech processor across the dynamic range. No matter how detailed the input amplitude information, these experiments limited the number of amplitude steps available to the listener to only 4, 8, 16, or 32 discrete amplitude levels. Both studies observed a decrease in phoneme recognition only when there were fewer than 8 levels of amplitude available. No differences were observed in recognition of speech processed with 8, 16, 32, or 1024 amplitude levels.

Several investigators (Drullman, 1995; Loizou et al., 2000; Shannon et al., 2001; Zeng & Galvin, 1999) eliminated high-amplitude information or low-amplitude information from speech by peak clipping or center clipping, respectively. They observed only modest degradation in phoneme recognition until the clipping exceeded 50% of the entire amplitude range.

In summary, amplitude mapping in a prosthetic device is important to the extent that the information must be audible to the listener. However, once the stimulation is audible, the exact mapping of amplitude is relatively unimportant. Only small decreases in phoneme recognition were observed even for amplitude mappings that were in error by a factor of two.

Temporal Cues

Temporal cues in speech can transmit information on envelope (<50 Hz), periodicity (50–500 Hz), and spectral fine structure (>500 Hz)(Rosen, 1992; Plomp, 1983). Listeners with normal hearing are capable of detecting and discriminating temporal fluctuations up to 300 to 500 Hz, even in the absence of spectral information (Burns & Viemeister, 1976, 1981; Viemeister, 1979), and may be able to use temporal information up to 1500 Hz for specialized tasks like spatial localization and complex pitch. However, recent studies that manipulated the temporal properties of speech (envelope low-pass filtering, stimulation pulse rate, cross-spectral desynchronization, temporal reversal) suggest that temporal information faster than 20 Hz is not important for speech, even when spectral cues are limited. Removal or distortion of all temporal information above 20 Hz has a negligible effect on speech recognition (Arai & Greenberg, 1998; Fu & Shannon, 2000; Saberi & Perrott, 1999; Fu & Galvin, 2001; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Vandali, Whitford, Plant, & Clark, 2000). Classic work on vocoder speech used only envelope information below 20 Hz in each frequency band and was able to reconstruct good quality speech at the receiving end (e.g., Hill, McRae, & McClellan, 1968). Shannon et al. (1995) systematically reduced the low-pass cutoff frequency on the envelope filter in their simulation of implant speech processing. Even for processors with only four spectral channels, they observed no reduction in performance when the cutoff frequency was reduced from 500 Hz to 50 Hz. Only a small reduction in speech recognition performance was observed when the cutoff frequency was reduced to 16 Hz. Drullman, Festen, and Plomp (1994a, 1994b) measured speech recognition in normal-hearing listeners when either slow or fast envelope modulations were systematically removed from speech. They found no decrement in speech recognition as long as envelope fluctuations below 16 Hz were present. Further reductions in envelope frequencies below 16 Hz resulted in a decrement, primarily for consonants.

Plosive bursts in speech produce a strongly synchronized burst of energy across the frequency spectrum, which might provide a cue for grouping of phonetic elements. There is a synchronous burst of energy across frequency for stop consonants and for each glottal pulse in a voiced sound. Several investigators (Arai & Greenberg, 1998; Fu & Galvin, 2001) disrupted this cross-spectral synchrony by introducing a random time delay between adjacent spectral bands. Although this manipulation disrupted the cross-spectral synchrony, speech recognition was almost completely unaffected up to asynchrony durations of more than 200 ms—the duration of a complete syllable.

Saberi and Perrott (1999) recently presented a dramatic illustration of the robustness of speech to temporal distortion. They divided sentences into equal time segments and reversed each segment in time; that is, each segment was played backward in time. Despite this severe temporal manipulation, speech was highly intelligible for reversal of segments as long as 100 ms.

Overall, these experiments demonstrate that speech recognition is highly resistant to temporal distortion. Most manipulations to the temporal structure of speech have only minimal effects on intelligibility, as long as the distortion does not span a time segment of more than 50 to 100 ms. This suggests that the effective temporal "window" for speech is about 50 ms. Speech with complete spectral cues is more resistant to temporal distortion than speech with reduced spectral resolution, but even spectrally reduced speech is highly resistant to temporal distortion up to 50 ms. AQ: B

Spectral Cues

In contrast to amplitude and temporal cues, both the quantity and quality of spectral cues are highly important for speech recognition. At least 4 spectral channels are necessary for speech recognition in quiet (Dorman, Loizou, & Rainey, 1997b; Shannon et al., 1995), more than 4 channels are necessary in noise (Fu, Shannon & Wang, 1998), and at least 16 channels are needed for music (Smith, Delgutte, & Oxenham, 2002).

Orig. Op.	OPERATOR:	Session	PROOF:	PE's:	AA's:	COMMENTS	ARTNO:
1st disk, 2nd mjr	taylormp	4					0009

Cochlear implant listeners increase in speech recognition performance as the number of electrodes is increased (Fishman, Shannon, & Slattery, 1997; Friesen, Shannon, Baskent, & Wang, 2001; Geier & Norton, 1992; Lawson, Wilson, & Finley, 1993). However, implant listeners' performance does not increase as the number of electrodes is increased above 8, whereas normal-hearing listeners continue to improve speech recognition as the number of spectral channels is increased (Fishman et al., 1997; Friesen et al., 2001). It is not clear at the present time what factors limit the performance of implant listeners to 8 effective spectral channels. One hypothesis is that the ability to make use of spectral information can be limited if the spectral information is not presented to the normal cochlear location (Greenwood, 1990).

In one experiment, "holes" were created in the frequency domain by turning off electrodes in cochlear implants (Shannon, Galvin, & Baskent, 2002). These conditions were simulated in normal-hearing listeners with a noise-band vocoder. Holes from 1 to 8 mm in tonotopic extent were created in the apical, middle, or basal part of the speech spectrum. Only apical holes caused a significant decrease in speech recognition and the decrease was similar for implant and normal-hearing listeners. This result suggests that listeners can tolerate a large section of missing spectral information in the middle and basal regions, as long as the rest of the spectrum is intact. Speech recognition is severely reduced if the missing spectral information is in the critical low-frequency region.

Spectral information must be delivered to the "normal" cochlear tonotopic location for speech to be recognized (Shannon, Zeng, & Wygonski, 1998). Speech recognition is significantly diminished if the spectral information is shifted by more than about 3 mm (2/3 octave; Dorman, Loizou, & Rainey, 1997a; Fu & Shannon, 1999), or if the frequency-to-place mapping is linearly expanded or compressed by more than 3 mm (Baskent & Shannon, 2003). These results suggest that tonotopic patterns of information for speech are stored and retrieved in the brain in terms of absolute cochlear location—distortions or shifts in the frequency-to-place mapping can result in substantial decreases in performance.

Can a new frequency-to-cochlear place map be learned? Results of short-term training in normal-hearing listeners suggested that people can adapt relatively quickly to a shift in the frequency-place map (Rosen, Faulkner, & Wilkinson, 1999). In contrast, recent results with cochlear-implant listeners (Fu, Shannon, & Galvin, 2002) suggest that even long-term learning may not be able to compensate for an incorrect frequency-place mapping. Fu et al. shifted the frequency-to-electrode map in three Nucleus-22 implant listeners by 3 mm. Their speech recognition was immediately reduced dramatically and recovered slightly over the first week of experience. However, speech recognition did not recover to its original level even after 3 months of daily use. This result suggests that the frequency-place mapping is highly important and not easily relearned, and so more attention should be paid to this dimension at the time of initial device fitting. This result is also consistent with studies of normal-hearing children listening to spectrally degraded speech (Eisenberg, Shannon, Martinez, Wygonski, & Boothroyd, 2000). Eisenberg et al. measured speech recognition in 5- and 11-year-old children with noise-band vocoders as a function of the number of bands. Results for Eleven-year-old children looked much like those of adults, but 5-year-old children needed more spectral resolution than adults or 11 year olds to achieve the same performance. This result suggests that speech pattern recognition is a complex process that takes between 5 and 11 years to fully mature, even in normal-hearing listeners. Thus, it may take many years to learn new patterns of speech information, especially in adults, in whom cortical plasticity is less flexible than in children.

Cochlear implant speech processors should deliver the highest number of spectral channels of information possible, and fitting procedures need to be developed to ensure the best mapping of frequency-to-place for each individual patient.

Summary

To fit cochlear implants properly and to design the next generation of cochlear implant speech processors, it is important to understand the relative importance of various cues for speech understanding. Recent research on amplitude, temporal, and spectral cues in speech show that speech recognition is only mildly affected by alterations or distortions in the amplitude and temporal domains, but is highly sensitive to manipulations in the spectral domain. Frequency-place manipulations that shift the speech pattern or distort it by more than about 3 mm cause severe degradation in understanding. Understanding the sensitivity of speech recognition to spectral manipulations will ultimately improve the fitting of an implant speech processor to the residual nerve survival in an individual patient.

References

Arai, T., & Greenberg, S. (1998). Speech intelligibility in the presence of cross-channel spectral asynchrony, *Proc. IEEE/ICASSP*, Seattle, WA: (pp. 933–936).

Baskent, D., & Shannon, R. V. (2002) *Speech recognition under conditions of frequency-place compression and expansion.* Manuscript submitted for publication.

Burns, E. M., & Viemeister, N. F. (1976). Nonspectral pitch. *Journal of the Acoustical Society of America, 60,* 863–869.

Burns, E. M., & Viemeister, N. F. (1981). Played-again SAM: Further observations on the pitch of amplitude-modulated noise. *Journal of the Acoustical Society of America, 70,* 1655–1660.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997a). Simulating the effect of cochlear-implant electrode insertion depth on speech understanding. *Journal of the Acoustical Society of America, 102,* 2993–2996.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997b). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America, 102,* 2403–2411.

Orig. Op.	OPERATOR:	Session	PROOF:	PE's:	AA's:	COMMENTS	ARTNO:
1st disk, 2nd mjr	taylormp	4					0009

Drullman, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *Journal of the Acoustical Society of America, 97,* 585–592.

Drullman, R., Festen, J. M., & Plomp, R. (1994a). Effect of temporal envelope smearing on speech perception. *Journal of the Acoustical Society of America, 95,* 1053–1064.

Drullman, R., Festen, J. M., & Plomp, R. (1994b). Effect of reducing slow temporal modulations on speech perception. *Journal of the Acoustical Society of America, 95,* 2670–2680.

Eisenberg, L., Shannon, R. V., Martinez, A. S., Wygonski, J., & Boothroyd, A. (2000). Speech recognition with reduced spectral cues as a function of age. *Journal of the Acoustical Society of America, 107*(5), 2704–2710.

Fishman, K., Shannon, R. V., & Slattery, W. H. (1997). Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. *Journal of Speech and Hearing Research, 40,* 1201–1215.

Friesen, L., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *Journal of the Acoustical Society of America, 110,* 1150–1163.

Fu, Q.-J., & Galvin, J. (2001). Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users. *Journal of the Acoustical Society of America, 109*(3), 1166–1172.

Fu, Q.-J., & Shannon, R. V. (1998). Effects of amplitude nonlinearities on speech recognition by cochlear implant users and normal-hearing listeners. *Journal of the Acoustical Society of America, 104,* 2570–2577.

Fu, Q.-J., & Shannon, R. V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *Journal of the Acoustical Society of America, 105,* 1889–1900.

Fu, Q.-J., & Shannon, R. V. (2000). Effect of stimulation rate on phoneme recognition in cochlear implants, *Journal of the Acoustical Society of America, 107(1),* 589–597.

Fu, Q.-J., Shannon, R. V., & Galvin, J. (2002). Perceptual learning following changes in the frequency-to-electrode assignment with the Nucleus-22 cochlear implant. *Journal of the Acoustical Society of America, 112,* 1664–1674.

Fu, Q.-J., Shannon, R. V., & Wang, X. (1998). Effects of noise and number of channels on vowel and consonant recognition: Acoustic and electric hearing. *Journal of the Acoustical Society of America, 104,* 3586–3596.

Geier, L., & Norton, S. (1992). The effects of limiting the number of Nucleus 22 cochlear implant electrodes programmed on speech perception. *Ear and Hearing, 13,* 340–348.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species: 29 years later. *Journal of the Acoustical Society of America, 87,* 2592–2605.

Hill, F. J., McRae, L. P., & McClellan, R. P. (1968). Speech recognition as a function of channel capacity in a discrete set of channels. *Journal of the Acoustical Society of America, 44,* 13–18.

Lawson, D., Wilson, B., & Finley, C. (1993). New processing strategies for multichannel cochlear prostheses. In J. A. Allum, D. J. Allum-Mecklenburg, F. P. Harris, and R. Probst (Eds.), *Natural and Artificial Control of Hearing and Balance, Progress in Brain Research, 97,* (pp. 313–321). Amsterdam: Elsevier.

Licklider, J. C. R., & Pollack, I. (1948). Effects of differentiation, integration, and infinite peak clipping on the intelligibility of speech. *Journal of the Acoustical Society of America, 20,* 42–51.

Loizou, P. C., Dorman, M., Poroy, O., & Spahr, T. (2000). Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution. *Journal of the Acoustical Society of America, 108,* 2377–2387.

Plomp, R. (1983). The role of modulation in hearing. In R. Klinke & R. Hartmann (Eds.), *Hearing –Physiological Bases and Psychophysics* (pp. 270–276). Berlin: Springer-Verlag.

Rosen, S. (1992). Temporal information in speech and its relevance for cochlear implants. *Philosophical Transactions of the Royal Society of London Series B Biol Sci, 336,* 367–373.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 106,* 3629–3636.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature, 398,* 760.

Shannon, R. V., Fu, Q-J., Wang, X., Galvin, J., & Wygonski, J. (2001). Critical cues for auditory pattern recognition in speech: Implications for cochlear implant speech processor design. In D. J. Breebaart, A. J. M. Houtsma, A. Kohlrausch, V. F. Prijs, & R. Schoonhoven (Eds.), *Physiological and Psychological Bases of Auditory Function: Proceedings of the 12th International Symposium on Hearing* (pp. 500–508). Maastricht, NL: Shaker Publishing BV.

Shannon, R. V., Galvin, J. G. & Baskent, D. (2002). Holes in hearing. *Journal of the Association for Research in Otolaryngology, 3,* 185–199.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270,* 303–304.

Shannon, R. V., Zeng, F.-G., & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *Journal of the Acoustical Society of America, 104,* 2467–2476.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature, 416,* 87–90.

Vandali, A. E., Whitford, L. A., Plant, K. L., & Clark, G. M. (2000). Speech perception as a function of electrical stimulation rate: Using the Nucleus 24 cochlear implant system. *Ear and Hearing, 21,* 608–624.

Viemeister, N. F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *Journal of the Acoustical Society of America, 66,* 1364–1380.

Zeng, F.-G., & Galvin, J. (1999). Amplitude mapping and phoneme recognition in cochlear implant listeners. *Ear and Hearing, 20,* 60–74.

Received
Accepted
First published (online) ???, 2002
http://professionals.asha.org/resources/journals/aja
DOI: 10.1044/1059–0889(2002/013)

Contact author: Robert V. Shannon, Department of Auditory Implants and Perception, House Ear Institute, 2100 W. Third St., Los Angeles, CA 90057. Email: Shannon@hei.org

Key Words: cochlear implants, speech recognition

JOBNAME: AUTHOR QUERIES PAGE: 1 SESS: 1 OUTPUT: Tue Dec 17 11:49:13 2002
/balt6/io-aja/io-aja/io0202/io0069-02a

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES 1

A—Author: Short title OK? If not, please provide and limit to 55 characters

B—Author: McCrae changed to McRae to match reference. Correct?

C—Author: Can this reference be updated?