Santos-Sacchi, J., & Dilger, J. P. (1988). Whole cell currents and mechanical responses of isolated outer hair cells. *Hearing Research, 35,* 143–150.

Santos-Sacchi, J., & Marovitz, W. F. (1980). An evaluation of normal strial capillary transport using the electron-opaque tracers ferritin and iron dextran. *Acta Otolaryngoly, 89,* 12–26.

Santos-Sacchi, J., Song, L., Zheng, J., & Nuttall, A. L. (2006). Control of mammalian cochlear amplification by chloride anions. *Journal of Neuroscience, 26,* 3992–3998.

Song, L., & Santos-Sacchi, J. (2012). Chloride dependent coupling of molecular to cellular mechanics in the outer hair cell of corti's prgan. *AIP Conference Proceedings, 1403,* 179–184.

Zheng, J., Shen, W., He, D. Z., Long, K. B., Madison, L. D., & Dallos, P. (2000). Prestin is the motor protein of cochlear outer hair cells. *Nature, 405,* 149–155.

Speech Perception and Hearing Loss

By Jont B. Allen, Ph.D., Andrea Trevino, and Woojae Han, Ph.D.

Over 150 years after the early research of Alexander Graham Bell, it remains unclear how the auditory system decodes speech, both in individuals who have "normal ears" and those who have "non-normal ears." Recent research has shown that normal ears can decode isolated consonants without error. However, when the inner ear is damaged, such as with sensorineural hearing loss where hair cells and synaptic connections are not properly functioning, speech can be heard but not understood. *In these cases, two seemingly-normal articulated utterances of the same consonant can result in totally different responses. Such specific and consistent confusions uniquely depend on the auditory system's function and the utterance. This presentation will discuss the differences between how the auditory systems of normal ears and non-normal ears receive and decode speech.*

Jont Allen, Ph.D., is an associate professor in the Department of Electrical and Computer Engineering at the University of Illinois. Allen received a B.S. in electrical engineering from the University of Illinois, Urbana-Champaign, and an M.S. and Ph.D. from the University of Pennsylvania. Upon graduation in 1970, Allen joined Bell Laboratories in Murray Hill N.J. From 1997–2002 he was a member of the newly created Research Division of AT&T (formerly Bell) Labs. In 2003, Allen retired from AT&T and joined the Department of Electrical and Computer Engineering at the University of Illinois. Andrea Trevino is a Ph.D. student in Human Speech Recognition group of the Department of Electrical and Computer Engineering at the University of Illinois, Urbana-Champaign. Woojae Han, Ph.D., is a recent graduate of the Department of Electrical and Computer Engineering at the University of Illinois, Urbana-Champaign and current assistant professor at Hallym University in Korea.

Introduction

Existing clinical methods for diagnosing speech disorders in individuals with damaged inner ears seem fundamentally broken. Today when patients go to an audiology clinic, their pure-tone hearing thresholds are first measured. Based on the degree of tonal hearing loss, a hearing aid may be prescribed, which is subsequently adjusted to partially compensate for the pure-tone loss. This may or may not improve the ear's speech loss (Walden & Montgomery, 1975). But since the speech loss is infrequently measured (or worse, the method of measurement is ineffective), the change is not quantified.

Based on the evidence available, it has been shown that speech testing has not been successful in fitting hearing aids (Walden & Montgomery, 1975). This seems counterintuitive since the main purpose of wearing a hearing aid is to improve speech understanding. Due to historically poor understanding of the fundamentals of speech perception, it has proven difficult to resolve this inconsistency. First, researchers may not understand the process of learning speech, which typically takes place in the first one to two years of life. Second, due to middle ear infections, young children can temporary lose their hearing, which can interfere with learning spoken language. It is not until the first year of school when the child is learning how to read that the child's ability to hear consonants is first fully tested.

Children who cannot accurately decode consonants may have increased difficulty with orthography. For example, if an ear cannot hear the distinction between /b/ and /d/ or between /t/ and /f/, the child is likely to misunderstand the importance of the shape of the letter [loop at bottom, closing to the left (d) or right (b), and curl at top (f) or bottom (t)]. The classroom teacher assumes that if a child's hearing is normal, then the child can hear the consonant distinctions. However, this assumption can be wrong and if so, the child's consonant decoding deficiency will go undetected (it will not show up in a pure-tone hearing test). When the child passes a hearing screen it is assumed, incorrectly, that they can decode syllables. What is needed is a targeted consonant discrimination test to predict these reading disorders.

Clinical audiologists can also make the same assumptions about adult speech perception, and research has shown that many of these assumptions can be wrong. The most serious assumption has been that consonants are homogeneous. Research has shown that for "normal ears," confusions systematically depend on the consonant (Phatak & Allen, 2007; Phatak, Lovitt, & Allen, 2008; Singh & Allen, 2012). For "non-normal ears," the errors dramatically increase, again depending on the ear, the noise-level, and, most significantly, the utterance.

If consonants were homogeneous, the confusions, as a function of the noise level, would be the same from one consonant to the next. This is not the case, since consonant confusions are highly dependent on the utterance (Han, 2011; Singh & Allen, 2012). While normal ears give similar confusions for a given utterance as a function of the noise, non-normal ears are idiosyncratic in their error patterns. The idiosyncratic nature of the speech scores implies that they may not be averaged. It is this inappropriate averaging that has led clinicians to believe that speech is not a reliable measure for fitting hearing aids.

In the last few years, the Human Speech Research (HSR) group at the Beckman Institute for Advanced Science and Technology at the University of Illinois, Urbana-Champaign, has determined some key elements in this chain that seem to enlighten responses from both normal and non-normal ears. For our purposes, "normal ears" are defined as those with pure-tone thresholds less than 20 dB-SPL, and "non-normal ears" are defined as having pure-tone thresholds greater than 20 dB-SPL.

Until very recently, it was not understood that the normal ear can detect speech with essentially zero error, down to –10 dB SNR (three times more speech-shaped noise than speech) (Phatak et al., 2008). As the noise increases, the error goes from zero to chance performance over a small signal-to-noise ratio (SNR) range. These new results totally change the understanding of what is happening in normal ears because it means consonant perception is binary (Singh & Allen, 2012).

The focus of this paper is to describe this difference in performance between the normal and non-normal ear at the utterance level. The paper will explain what the HSR group has found, and then predict where this research will go in the next few years. In addition, we will discuss a speech test that teases out such natural occurring idiosyncratic speech confusions, which we argue will eventually be useful for fitting hearing aids.

How Does Speech Perception Fail?

The challenge remains to understand the auditory processing strategy of the auditory cortex, which is wired to non-normal ears. To understand how normal ears decode consonants, the HSR group repeated the classic consonant perception experiments of Fletcher (1922) and Miller and Nicely (1955), among others. This gave us access to important new data and the ability to reassess many widely held assumptions. The first lesson of this research is the "sin of averaging"—while audiology is built on averaging measures, most of the interesting information is lost in these averages. We have shown, for example, that averaging across consonants distorts the measure as does averaging across talkers for a given consonant. We have also found that entropy (a probabilistic measure of consistency) is more robust than the average error.

In 1970–80, a number of studies explored the role of the transitional and burst cues in a consonant-vowel (CV) context. In a review of the literature, Cole and Scott (1974) argued that the burst must play at least a partial role in perception, along with transition and speech energy envelope cues. Explicitly responding to Cole and Scott (1974), Dorman and colleagues (1977) executed an extensive experiment using natural speech made up from nine vowels proceeded by /b, d, g/. The experimental procedure consisted of truncating the consonant burst and the devoiced transition (following the burst) of a CVC, and then splicing these onto a second VC sound, presumably with no transition component (since it had no initial consonant). Their results were presented as a complex set of interactions between the initial consonant (burst and devoiced cue) and the following vowel (i.e., coarticulations).

The same year Blumstein and colleagues (1977) published a related /b, d, g/ study using synthetic speech that also presented a look at the burst and a host of transition cues. They explored the possibility that the acoustic cues were *integrated* (acted as a whole). This study was looking to distinguish the *necessary* from the *sufficient* cues, and first introduced the concept of *conflicting cues* in an attempt to pit one type (burst cues) against the other (transition cues).

While these three key studies highlighted the relative importance of the two main types of acoustic cue, burst and transition, they left unresolved the identity and relative roles of these cues. No masking noise was used in the studies, ruling out any form of information analysis. Masking is key to an information theoretic analysis of any communication channel (Allen, 1994, 1996; Fletcher, 1922; Shannon, 1948). As discussed by Allen (2005), based on the earlier work of Fletcher and Galt (1950), Miller and Nicely (1955), and inspired by Shannon's source-channel model of communication, the HSR group repeated many of the classic experiments (Li & Allen, 2009; Phatak & Allen, 2007; Phatak et al., 2008). The data resulting from our several experiments will be discussed in the remainder of the paper.

Identifying Perceptual Cues

Li and colleagues (2010) first described a method to robustly identify speech cues for a variety of naturally produced CV speech sounds. This method uses a 3-dimensional psychophysical approach using a variety of noise levels, time-truncation, and high and low pass filtering. These experiments made it possible, for the first time, to reliably locate the subset of perceptually relevant cues in time and frequency, while the noise-masking data characterizes the cue's masked threshold (i.e., its strength).

Figure 4 describes the resulting consonant maps. Not surprisingly, the perceptual cues associated with fricative sounds are quite different from the plosives. Timing and bandwidth remain important variables. For the fricative sounds, the lower edge of the swath of frication noise is the perceptual cue.

Briefly summarized in Figure 4, the CV sounds /ta, da/ are defined by a burst at high frequencies, /ka, ga/ are defined by a similar burst in the mid frequencies, and /ba, pa/ were traced back to a wide-band burst. As noise is added, the wide-band burst frequently degenerates into a low frequency burst, resulting in low-level confusions. The recognition of burst-consonants further depends on the delay between the burst and the sonorant onset, defined as the voice onset time (VOT). Consonants /t, k, p/ are voiceless sounds, occurring

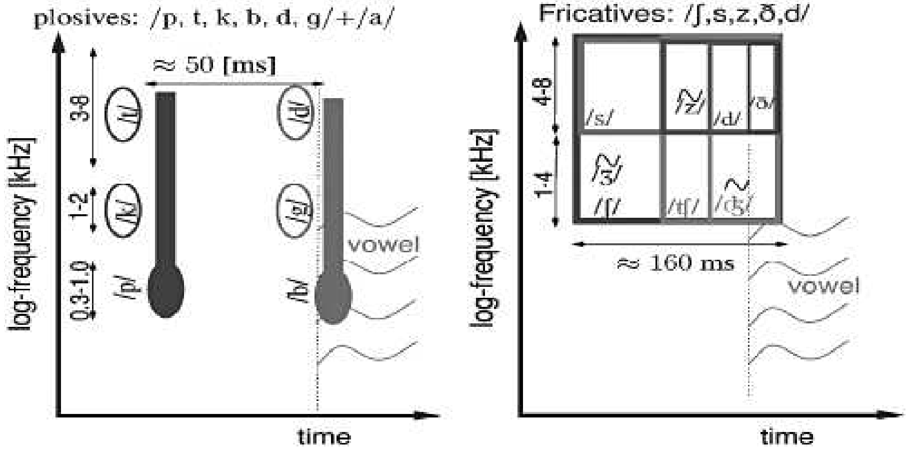


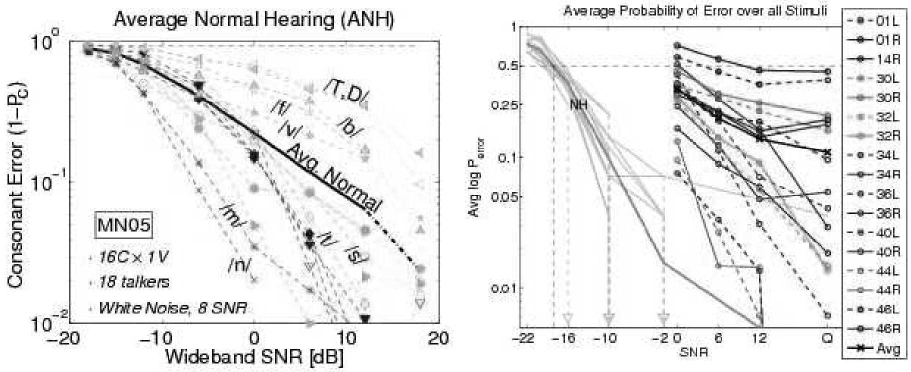
Figure 4. Time-frequency allocation of the plosives and the fricatives. Mapping these regions into perceptual cues required extensive perceptual experiments (Li et al., 2010). Once the sounds have been evaluated, it is possible to prove how the key noise-robust perceptual cues map to acoustic features. In the cases of the three voiced consonants indicated with a tilde (/z, ʒ, ʤ/), the frication noise is modulated at the pitch frequency.

about 50 [ms] before the onset of F0 voicing while /d, g/ have a VOT <20 [ms]. Plosive /b/ may have a negative VOT.

Based on the results of Li and colleagues (2010), this study, along with a host of verification experiments on the ~100 CV utterances in the HSR database (Kapoor & Allen, 2012; Li & Allen, 2011; Régnier & Allen, 2008), we have conclusively demonstrated that these features uniquely label the indicated consonant.

Methods

Isolated CVs were taken from naturally produced speech from 18 talkers. Noise was added to the speech with a range from –26 dB to quiet (Q). Both uniform and speech weighted spectrum level noise was added to the speech. The listener corpus consisted of more than 200 normal and 45 non-normal ears, with 9-16 consonant and 8 vowel sounds. To assure the estimates of the error are reliable, a minimum of 10 trials per utterance and SNR are required (Han, 2011; Phatak, Yoon, Gooler, & Allen, 2009; Singh & Allen, 2012). The difference between these new experiments and their classic counterparts is that the utterances of each consonant are not averaged.



(a) Consonant errors ANH

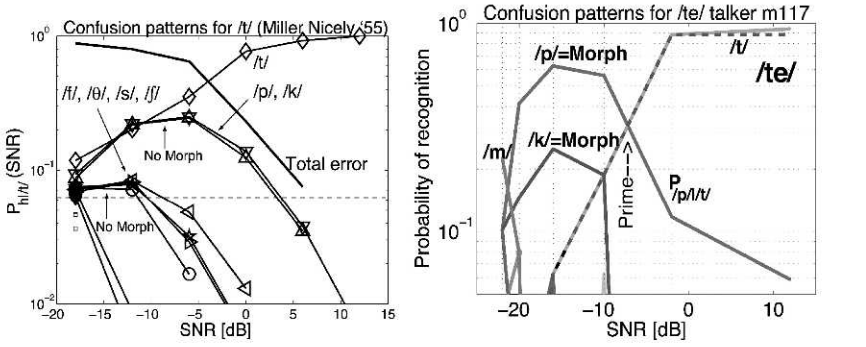
(b) Error for HI ear 112R.

Figure 5. LEFT: Shown here is the average error (log scale) for 16 CV consonants as a function of the relative intensity of constant-spectrum-level masking noise (Phatak et al., 2009). The solid black curve labeled "Avg. Normal" shows the average across all the consonants. Note the large variation in error. RIGHT: This family of curves compares the average consonant error for 14 normal and 17 non-normal ears in speech shaped masking noise. For the non-normal ears, there is a large spread in scores due to the variation in hearing loss as compared to listeners with normal hearing (gray region), all of whom are similar in their average performance.

Results

In Figure 5, the average probability of the error P_e(SNR) is shown (for speech-weighted noise the SNR is the same as the articulation index). On the left (a), the "average normal hearing" (ANH) score P_e(SNR) (black line), along with the score for each heard consonant /h/, given spoken consonant /s/ as a function of the SNR for flat-spectrum masking noise (Phatak et al., 2009). There is a huge variation in scores across the consonants: the SNR corresponding to the 50% point ranges from –12 dB [/m, n/] to +8 dB [/θ, ð/] [shown as /T/ and /D/ in Figure 5)]. Such a large range of scores is not captured by an average. Not shown here, each utterance in the HSR database has a wide range of scores, varying in error from zero to chance depending on the masking noise intensity (Singh & Allen, 2012).

The right panel (b) shows the average scores for the 17 non-normal ears as compared to the average scores of the participants with normal hearing in speech-shaped masking noise. One of the best ears in terms of average error is 36R. Not shown is that his error for /ba/ reaches 100%, while the remaining 13 consonants tested had zero error. Thus, the reported performance is highly distorted, again due to the "sin of averaging."



(c) Average over all /t/s.

(d) Talker m117 /te/ $P_{h|/ta/}(SNR)$

Figure 6. The "sin of averaging" extends to the utterance level. On the left (a) we see confusion patterns for the average score for /ta/ from Miller and Nicely (1955) (white masking noise). On the right (b) we show the confusion patterns for male talker 117 saying /te/ (speech shaped noise). Based on data with the same masking conditions, and as concluded in Figure 5, averaging across utterances removes critical information from the ANH scores. The confusion error is a function of the SNR in dB. As we shall see, this "sin" is much worse for non-normal ears at the utterance level. The arrow at –8 dB and 30% shows the priming point, defined as where a listener reports one of a small set of sounds (Li & Allen, 2011).

A second major conclusion is that when characterizing a listener with hearing loss, one must look at the individual confusions. In Figure 6, confusion patterns (CPs) are compared to SNR. The CP is a graphical display of the confusion probabilities as a function of the intensity of the masking noise relative to the speech. To estimate a CP requires a totally different clinical measure than is being applied today. CPs allow one to visualize the confusions of each sound as a function of the SNR. From the CP it is easy to identify a sound that primes, meaning that it can be heard as one of several sounds with equal probability by changing one's mental bias. In this case the CPs show subject responses that are equal (the curves cross each other), similar to the CP of Figure 6(b) at –8 dB where one naturally primes /p/, /t/, and, to a lesser extent, /k/ (at –10 dB).

When asked, most clinicians report that they do not have the time to make detailed measures. In our opinion, this is more a reflection of old habits than actual time constraints. The confusion sets, and their dependence on the noise, are not predictable without such tests. Utterance confusions and their masked dependence are important because they reveal the mix of underlying perceptual cues being confused with the target sound.

When using an utterance confusion measure, each non-normal ear consistently makes large errors on a small subset of utterances. Furthermore,

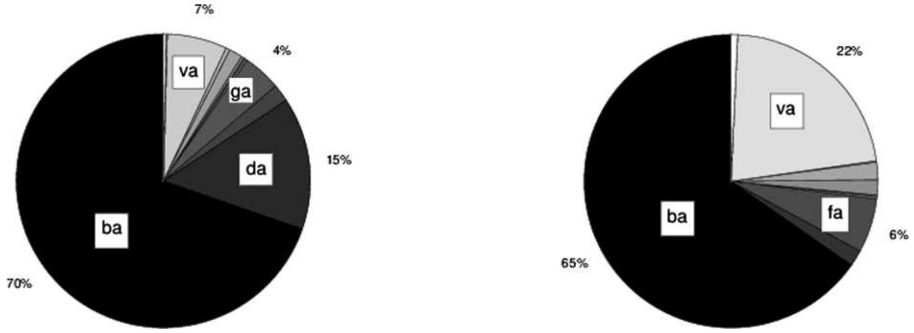


Figure 7. These pie charts show the proportion of confusions for two different /ba/ utterances, as reported by all of the 17 non-normal ears. The most common error for the /ba/ on the left is /da/ and then /va/, while the one on the right is most frequently heard as /va/ and then /fa/. The one on the left is almost never heard as /fa/ and the one on the right almost never as /da/. These two /ba/ sounds are reported correctly by normal ears.

for a given utterance, there are patterns in these errors across listeners with hearing loss. In other words, normally spoken utterances are heard idiosyncratically by non-normal ears, yet with correlated error patterns.

Confusions in Non-Normal Ears

As a direct extension of earlier studies (e.g., Phatak et al., 2009), four experiments were performed (Han 2011), two of which will be reported on here. In Experiment 1 (Exp-1), full-rank confusion matrices for the 16 Miller-Nicely CV sounds were determined at 6 SNR [Q, 12, 6, 0, –6, and –12 dB] for 46 non-normal ears (25 subjects). In Experiment 2 (Exp-2), a subset of 17 ears were remeasured, but with the total number of trials per SNR per consonant raised from 2–8 (Exp-1) to as high as 20 (Exp-2) to statistically verify the reliability of the subjects' responses in doing the task.

Figure 7 shows that listeners with hearing loss are using a common strategy that depends systematically on the utterance. Clearly, if such very different scores for the two /ba/ sounds were to be averaged together (i.e., present clinical practice), the idiosyncratic (i.e., the most important) information about the ears would be lost. As discussed earlier, the average score is a distorted metric due to its high variance a) across consonants, b) across utterances for each consonant, and c) across subjects with hearing loss. Entropy gives a direct measure of consistency and is insensitive to mislabeling errors (e.g., consistently across a voicing error, as in reporting /d/ given /t/). Given the observed increased mislabeling of sounds in non-normal ears, a high-consistency measure (i.e., entropy) seems to be a better measure.

Summary

This article has reviewed some of what the HSR group has recently learned about speech perception of consonants, and how this knowledge might impact understanding of nonlinear (NL) cochlear speech processing. However, the role of outer hair cell (OHC) processing of speech is still poorly understood (Allen, 2008; Allen & Li, 2009). It is now widely accepted that OHCs provide dynamic range and are responsible for much of the NL cochlear speech signal processing, thus the common element that links all the NL data (Allen, Régnier, Phatak, & Li, 2009). OHC dynamics must be understood before any model can hope to succeed in predicting basilar membrane, hair cell, neural tuning, and NL compression. Understanding the OHC's two-way mechanical transduction may be the key to solving the problem of the cochlea's dynamic range and dynamic response (Allen, 2003).

However, the perception of speech by the non-normal ear does not seem to be consistent with the above commonly held view. For example, the large individual differences seem inconsistent with the OHC as the tying link, and seem more likely related to synaptic dead regions (Kujawa & Liberman, 2009). Continued analysis of these confusions will hopefully provide further insights into this important question. The detailed study of how a complex system fails can give deep insights into how the normal system works.

The key open problem here is, "How does the auditory system (e.g., the NL cochlea and the auditory cortex) process human speech?" There are many applications of these results including speech coding, speech recognition in noise, hearing aids, and cochlear implants as well as language acquisition and reading disorders in children. If we can solve the *robust phone decoding problem,* we will fundamentally change the effectiveness of human-machine interactions. For example, the ultimate hearing aid is the hearing aid with built in robust speech feature detection and phone recognition. While researchers have no idea when speech-aware hearing aids will come to be, and the time is undoubtedly many years off, when it happens, it will be a technological revolution of some magnitude.

Acknowledgments

We gratefully acknowledge Phonak and Research in Motion for their generous support of this research. Much of the work was also supported by NIH grant # RDC009277A.

References

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio, 2*, 567–577.

Allen, J. B. (1996). Harvey Fletcher's role in the creation of communication acoustics. *Journal of the Acoustical Society of America, 99*, 1825–1839.

Allen, J. B. (2003). Amplitude compression in hearing aids. In *MIT Encyclopedia of Communication Disorders*, R. Kent (ed.), p. 413–423. Boston: MIT Press.

Allen, J. B. (2005). *Articulation and intelligibility*. LaPorte, CO: Morgan and Claypool.

Allen, J. B. (2008). Nonlinear cochlear signal processing and masking in speech perception. In *Springer Handbook on speech processing and speech communication*, J. Benesty & M. Sondhi (eds.), p. 1–36. Heidelberg, Germany: Springer.

Allen, J. B. & Li, F. (2009). Speech perception and cochlear signal processing. *IEEE Signal Processing Magazine, 26,* 73–77.

Allen, J. B., Régnier, M., Phatak, S., & Li, F. (2009). Nonlinear cochlear signal processing and phoneme perception. In *Proceedings of the 10th mechanics of hearing workshop*, y N. P. Cooper & D. T. Kemp (eds.), p. 93–105. Singapore: World Scientific Publishing Co.

Blumstein, S. E., Stevens, K. N., & Nigro, G. N. (1977). Property detectors for bursts and transitions in speech perceptions. *Journal of the Acoustical Society of America, 61*, 1301–1313.

Cole, R., & Scott, B. (1974). Toward a theory of speech perception. *Psychological Review, 81*, 348–74.

Dorman, M., Studdert-Kennedy, M., & Raphael, L. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionlly equivialent, context-dependent cues. *Perception and Psychophysics, 22*, 109–22.

Fletcher, H. (1922). The nature of speech and its interpretation. *Journal of the Franklin Institute, 193*, 729–747.

Fletcher, H., & Galt, R. (1950). Perception of speech and its relation to telephony. *Journal of the Acoustical Society of America, 22*, 89–151.

Han, W. (2011). Methods for robust characterization of consonant perception in hearing-impaired listeners. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Kapoor, A., & Allen, J. B. (2012). "Perceptual effects of plosive feature modification." *Journal of the Acoustical Society of America, 131*, 478–491.

Kujawa, S., & Liberman, M. (2009). Adding insult to injury: Cochlear nerve degeneration after "temporary" noise-induced hearing loss. *Journal of Neuroscience, 29*, 14077–14085.

Li, F., & Allen, J. B. (2009). Additivity law of frequency integration for consonant identification in white noise. *Journal of the Acoustical Society of America*, *126*, 347–353.

Li, F., & Allen, J. B. (2011). Manipulation of consonants in natural speech. *IEEE Transactions on Audio, Speech, and Language Processing, 19*, 496–504.

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America*, *127*, 2599–2610.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*, 338–352.

Phatak, S., & Allen, J. B. (2007). Consonant and vowel confusions in speech-weighted noise. *Journal of the Acoustical Society of America*, *121,* 2312–2326.

Phatak, S., Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *Journal of the Acoustical Society of America*, *124,* 1220–1233.

Phatak, S., Yoon, Y., Gooler, D. M., & Allen, J. B. (2009). Consonant loss profiles in hearing impaired listeners. *Journal of the Acoustical Society of America*, *126,* 2683–2694.

Régnier, M. S., & Allen, J. B. (2008). A method to identify noise-robust perceptual features: Application for consonant /t/. *Journal of the Acoustical Society of America*, *123,* 2801–2814.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423 (parts I, II), 623–656 (part III).

Singh, R., & Allen, J. B. (2012). Sources of stop consonant errors in low-noise environments. *Journal of the Acoustical Society of America*, *131,* in press.

Walden, B. F., & Montgomery, A. A. (1975). Dimensions of consonant perception in normal and hearing-impaired listeners. *Journal of Speech and Hearing Research, 18,* 444–455.