# Sources of decoding errors of the perceptual cues, in normal and hearing impaired ears[a]

JONT B. ALLEN[1,*] AND WOOJAE HAN[1]

[1] *University of Illinois, Urbana-Champaign, IL, USA*

[2] *Hallym University, Korea*

After many decades of work it is not understood how the average normal-hearing (NH) ears, or significantly hearing-impaired (HI) ears, decode consonants. We wish to discover the strategy HI persons use to recognize consonants in a consonant-vowel (CV) context. To understand how NH ears decode consonants, we have repeated the classic consonant perception experiments of Fletcher, French and Steinberg, G.A. Miller, Furui, and others. This has given us access to the raw data (e.g., to allow for ANOVA testing) and the ability to verify many widely held (typically *wrong*) assumptions. The first lesson of this research is the *sin* of averaging: While audiology is built on average measures, most of the interesting information is lost in these averages. It has been shown, for example, that averaging across consonants is a grievous error, as is averaging across talkers for a given consonant. It will be shown how an average entropy measure (a measure of dispersion in probability) has higher utility than the average error.

## INTRODUCTION

A fundamental problem in auditory science is the perceptual basis of speech, that is, phoneme decoding. How the ear decodes basic speech sounds is important for both hearing-aid and cochlear-implant signal processing, both in quiet and in noise. The object of our studies are three-fold (We are at the out-set of objective 3, objectives 1 and 2 being mostly complete):

1. We have isolated the acoustic cues in $>100$ consonant-vowel (CV) utterances.
2. We have measured the full-rank confusions in $\approx 50$ hearing-impaired (HI) ears.
3. We are attempting to relate the measured HI confusions to the NH cues.

*Objective 1):* An *acoustic cue* is defined as the time-frequency features of the acoustic signal which are decoded by the auditory system for representing the consonant-vowel (CV) combination (Cole and Scott, 1974). The acoustic cues used by the average normal-hearing (ANH) ear are made up of at least four different cues (Li and Allen, 2009): a) onset bursts, b) low-frequency "edges," c) durations, and d) F0 modulation.

[a] Page numbers starting with A refer to the ISAAR 2011 proceedings numbering.

*Corresponding author: jontalle@illinois.edu

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. C. Dalsgaard, and T. Poulsen. ISBN: 978-87-990013-3-0. The Danavox Jubilee Foundation, 2012.

The timing of the onset burst is relative to the onset of voicing of the vowel. A low-frequency *edge* is defined as the lowest frequency of the fricative region (Li and Allen, 2011).

*Objective 2):* We shall see that individual differences are the rule in HI confusions. No two ears are the same.

*Objective 3):* Our underlying hypothesis is that the consonant loss experienced by the HI ear is due to degradations in the cochlea, that cause a specific loss of detectability of specific classes (e.g., onset-burst, F0 detection) of consonant cues. Based on the HI data obtained, the most likely character of the consonant loss is *cochlear dead regions*, e.g., regions where the synapse is poorly connected to the auditory nerve (Allen *et al.*, 2009).

We hypothesize that when one or more of these cues is diminished in the ANH ear, certain consonants are confused with others in a predictable way. This hypothesis seems in agreement with our present findings, however the precise relationships are yet to be determined. It is significant that a) there are large individual differences, that appear to be b) uncorrelated to the audiograms, and that c) the HI ears are consistent in their judgments.

Questions being addressed in our publications include the following (many papers are still under review, as identified in Table 1):

1. What is the the phone error rate in NH and HI ears? (Phatak and Allen, 2007)
2. What is the source of this error (which consonants and confusions vs. SNRs)? (Singh and Allen, 2012)
3. What are the invariant acoustic cues used by NH and HI ears to identify consonants? (Li *et al.*, 2010)
4. Is audibility of an acoustic cue sufficient (it is necessary), and how may this be measured? (Li and Allen, 2011)
5. How does the HI ear differ from the NH ear in detecting invariant acoustic cues? (Han, 2011)
6. When does enhancing the SNR of a missed cue improve the robustness to noise of a consonant (Kapoor and Allen, 2012)?
7. What is the impact of NAL-R amplification on consonant perception (Phatak *et al.*, 2009; Han, 2011)?
8. How can we clinically quantify and diagnose the HI ear using speech (Han, 2011)?

Additional questions for future research include:

1. How do invariant acoustic cues depend on the following vowel?

2. Can we fit a hearing aid using consonant confusion profiles?

## HISTORICAL STUDIES

The first speech studies were done in by Lord Rayleigh (1908) following the telephone's commercialization. Within a few years, Western-Electric's George Campbell (1910) developed the electrical wave filter to high and lowpass speech signals, as well as probabilistic models of speech perception such as the *confusion matrix method* of analysis. With these tools established, Harvey Fletcher (1921) extended these with related studies. He soon discovered that by breaking the speech into bands having equal scores, he could formulate a rule relating the errors in each band to the wide-band error. This method became known as the *articulation index* (AI). Even today it is not clear why the AI is well correlated to the average speech score (Singh and Allen, 2012). Today we know that Fletcher's 1921 AI formulation is similar to Claude Shannon's theory of information (1948) (Allen, 2004).

### Contemporary studies

In 1970-80 a number of papers explored the role of the transitional and burst cues in consonant-vowel 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 *et al.* (1977) executed an extensive experiment, using natural speech consisting of nine vowels, preceded 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 having 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 *et al.* (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 publications highlighted the relative importance of the two main types of acoustic cue, burst and transition, they left unresolved their identity, or even their relative roles. In these three studies, no such masking noise was used, ruling out any form of information analysis. Masking is the classical key element basic to an information theoretic analysis of any communication channel (Fletcher, 1922; Shannon, 1948; Allen, 1994, 1996). 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, we repeated many of the classic experiments (Phatak and Allen, 2007; Phatak *et al.*, 2008; Li and Allen, 2009). A table summarizing the speech experiments done at UIUC between 2003-2011 is shown in Table 1.[1]

| Year | Experiment | Student &Allen | Details | Publications |
|---|---|---|---|---|
| 2004 | MN04(MN64) | Phatak | MN14 | Phatak and Allen (2007) |
| 2005 | MN16R | Phatak, Lovitt | MN55R | Phatak *et al.* (2008) |
| 2005 | HIMCL05 | Yoon, Phatak | 10 HI ears | Phatak *et al.* (2009) |
| 2006 | HINALR05 | Yoon *et al.* | 10 HI ears | Yoon *et al.* (2011) |
| 2006 | Verification | Regnier | /ta/ | Régnier and Allen (2008) |
| 2006 | CV06-s/w | Phatak/Regnier | 8C+9V SWN/WN | – |
| 2007 | CV06 | Pan | CV06 | – |
| 2007 | HL07 | Li | Hi/Lo pass | Li and Allen (2009) |
| 2008 | TR08 | Li | Furui86 | ASSP |
| 2009 | 3DDS | Li | plosives | Alen and Li (2009); Li *et al.* (2010); Li and Allen (2011) |
| 2009 | Verification | Kapoor/Cvengros | burst mods | Kapoor and Allen (2012) |
| 2009 | MN64 NZ-Error | Singh | PA07 | Submitted JASA |
| 2010 | HI-MCL10 1,2,3 | Han | 46 HI ears @MCL | Submitted EH |
| 2011 | 3DDS | Li | Fricatives | Submitted JASA |
| 2011 | HI-NAL11 4 | Han | 17 HI ears w NALR | Thesis Ch. 3 |

**Table 1:** Table of HSR experiments performed at UIUC from 2004-2011

## Methods

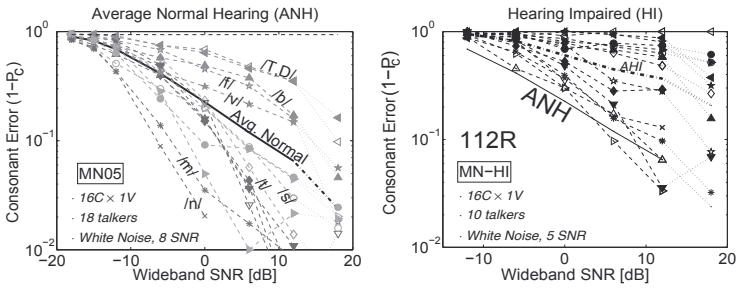
Isolated CVs were taken from real speech, with up to 20 talkers. Noise was added to the speech with a range of between 4-8 SNRs, from -26 to quiet (Q). The speech was high- and lowpass-filtered with up to 10 high/lowpass cutoff frequencies. Both white and speech-weighted additive noise was used. The listener corpus consisted of more than 200 NH subjects, 45 HI ears, up to 18 consonants and 8 vowels, and always maintaining a high source entropy (e.g., 4 bits) to eliminate guessing. To assure the estimates of the error are reliable, a minimum of 20 trials per consonant and SNR are required.

In Fig. 1 the average probability of the error $P_e(SNR)$ is shown (for speech-weighted noise the SNR *is* the articulation index). In Fig. 2 *confusion patterns* (CPs) are displayed vs. SNR.

## RESULTS

From Fig.1 we see the ANH score $P_e(SNR)$ (black line), along with the score for each heard consonant $h$ given spoken consonant $s$ [i.e., $P_{h|s}(SNR)$], as a function of the SNR. What is most obvious is the large variation in scores: the SNR corresponding

[1] `http://hear.beckman.illinois.edu/wiki/Main/Publications`




(a) Consonant errors ANH (b) Error for HI ear 112R.

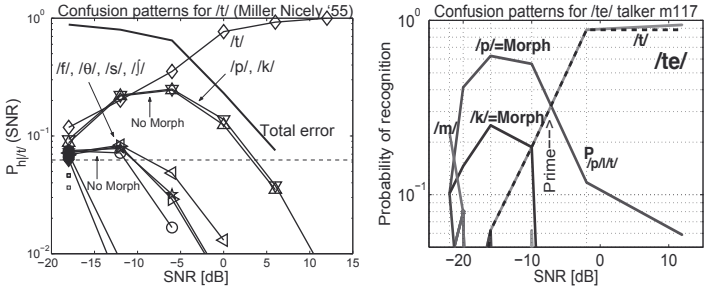
**Fig. 1:** Due to the large variation across consonants, the average error [e.g., $P_e(SNR) \equiv 1 - P_c(SNR)$, black line] fails to characterize speech loss. This *sin of averaging* results from (a) averaging across the natural variance across consonants (left: NH listeners), (b) across consonants for individual HI listeners (right). HI data for 112R from Phatak *et al.* (2009).

to the 50% point ranges from −12 dB [/m, n/ to +8 dB /θ, ð/ (shown as /T/ and /D/ in the figure)]. Such a large range of scores is not well captured by an average. The same is true for HI ear 112R shown in Fig. 1(b): The average score (black dashed curve) does not meaningfully represent the consonant scores. Although not shown, every consonant in our database has a wide range of scores, varying from zero error on most cases, to chance, over a wide range of SNRs (Singh and Allen, 2012).

CPs allow one to determine the precise nature of the confusions of each sound as a function of the SNR. The confusion set, and their dependence on SNR, are not predictable without running masking experiments. These confusions, and their masked dependence, are important because they reveal the mix of underlying perceptual cues. From the CP it is easy to identify a sound that *primes*, meaning that it can be heard as one of several sounds, by changing one's mental bias. In this case the confusion patterns show subject responses that are equal (the curves cross each other), similar to the CP of Fig. 2(b) at −8 dB, where one naturally primes /p/, /t/ and to a lesser extent /k/ (at −15 dB).

### Identifying perceptual cues

Li *et al.* (2010) first described the 3DDS method, used to identify speech cues for a variety of real speech sounds. This method uses extensive psychophysical experimental on-CV speech-by-noise masking at a variety of SNRs, along with time-truncation and high- and lowpass 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 feature's masked threshold (i.e., its strength). In Fig. 3 the speech was displayed by an *AIgram* (Régnier and




(a) Average over all /t/s. (b) Talker m117 /te/ $P_{h|/ta/}(SNR)$

**Fig. 2:** The *sin of averaging* extends down to the utterance level. On the left (a) we see CPs for the average score across /ta/ from Miller and Nicely (1955), while on the right (b) we see the CPs for a specific /te/. As in Fig. 1, one must conclude that averaging across utterances removes critical information from the ANH scores. As we shall see, this sin is much worse for HI ears, at the utterance level. *Priming* is reporting the sound one is thinking of, typically from a small group of sounds (Li and Allen, 2011).

Allen, 2008). The AIgram resolves acoustic features than are not easily visualized in the traditional spectrogram due to its fixed frequency resolution. First the AIgram is normalized to the noise floor. This is similar to the cochlea which dynamically adapts to the noise floor due to outer-hair-cell (OHC) nonlinear (NL) processing (Allen, 2003; Allen *et al.*, 2009). Second, unlike a fixed-bandwidth spectrogram, the AIgram uses a cochlear filter bank, with bandwidths given by Fletcher critical bands (ERBs) (Allen, 1996). Finally the intensity scale in the plot is proportional to the signal-to-noise ratio, in dB, in each critical band, as in AI-band densities $AI_k(SNR)$ for the $k$th band (Li *et al.*, 2010; Li and Allen, 2011). At the present time the AIgram is linear as it contains no on-frequency neural masking, nor forward and upward spread of neural masking. As a result the AIgram shows details in the speech that are not actually audible. Much work remains to be done on time-domain NL cochlear models of speech.

A summary of the audible sound cues at the threshold of masking are shown in the AIgram, as exampled in the lower-left panel for each of the six consonants in Fig. 3.

### Plosives

In Fig. 3 there are six sets of 4 panels, as described in the caption. Each of the six sets corresponds to a specific consonant, labeled by a character string that defines the gender (m,f), subject ID, consonant, and SNR for the display. For example, in the upper-left 4 panels we see the analysis of /ta/ for female talker 105 (`f105ta0dB`) at 0 dB. Along the top are unvoiced plosives /t/, /k/, and /p/ while along the bottom are voiced plosives /d/, /g/, and /b/. Data from the same talker were not always available

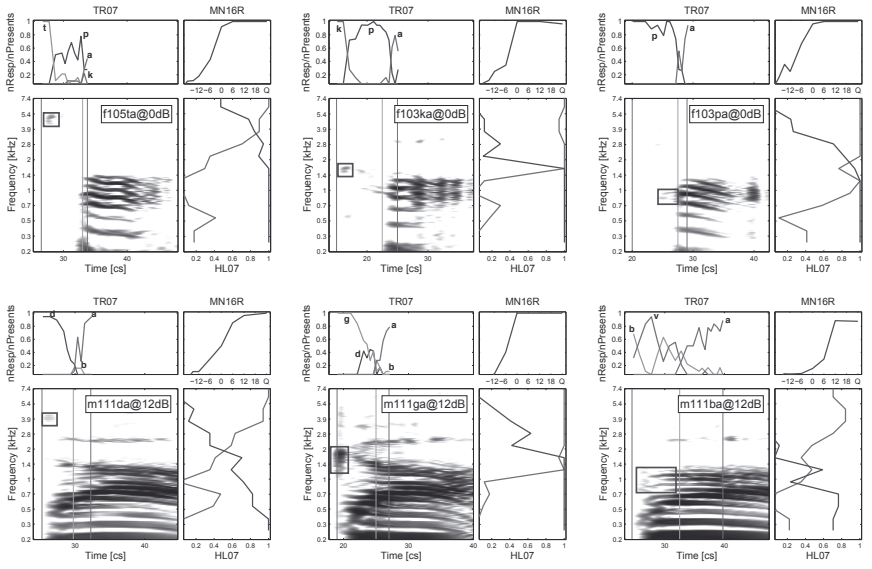



**Fig. 3:** Identification of cues by time, frequency, and intensity bisection using the 3-dimensional deep search (3DDS) methods, as shown here. Along the top we have unvoiced consonants /t/, /k/, and /p/, while along the bottom, the corresponding voiced consonants /d/, /g/, and /b/. Each of the six sounds consists of 4 sub-panels. For example, for /t/, upper-left, shows four panels consisting of the time-truncation confusions (upper-left), the score vs. SNR (upper-right), the AIgram (lower-left), and the score as a function of low and highpass filtering (lower-right). This last panel is rotate by 90 degrees with the score along the abscissa and the frequency along the ordinate, to line-up with the AIgram frequency axis.

in the LDC database (Fousek *et al.*, 1974), so different talkers are sometimes used for this analysis.

Three different modifications have been made to the speech: The *first* was the reported experiments (MN64, MN16R) (Phatak and Allen, 2007; Phatak *et al.*, 2008) where each CV sound was subjected to a variable signal-to-noise ratio, from $-12$ dB SNR to quiet, and the average score was measured by 23 NH listeners.

*Next* each sound was time-truncated from the onset in 10-ms steps (Exps. TR07, TR08) (Furui, 1986), and played back in random order to 14 listeners. Noise was added to the truncated sound at 12 dB SNR to remove any low-level artifacts. The results of this *truncation experiment* are presented in the top upper-left panel (labeled as TR07). Each curve is the probability $P_{h|s}(t_k)$, where $h$ is the *heard* (reported) sound

as a function of the *spoken* sound $s$ at a truncation time $t_k$, labeled with the identified consonant.

Finally each CV sample was high- and lowpass filtered to a variable cutoff frequency (Li and Allen, 2009, Exps. HL05 & HL07), as indicated on the frequency axis. These HL07 data are rotated by 90 degrees so that the frequency axis lines up with that of the AIgram on the far left.

One may learn to identify perceptual cues from the 3DDS display (Li *et al.*, 2010). For example, the feature that labels the sound is indicated by the blue rectangle in the AIgram (lower-left panel) of each of the six sounds. When this burst is time-truncated (the TR07 experiment), the /t/ morphs to /p/. The term *morphs* means that one sound can be primed, i.e., is heard as several different sounds. When masking noise is added to the sound, such that it masks the boxed region, the percept of /t/ is lost. When the high- and lowpass filters remove the frequency of the /t/ burst, again the consonant is lost. Thus the three experiments are in agreement, and collectively they uniquely identify the location of the acoustic cue responsible for /t/. This generalizes to the other plosive consonants shown (i.e., voiced /k/, /p/, and unvoiced /d/, /g/, /b/), fricatives, as well as consonants followed by other vowels (not shown).

Looking at specific examples in the individual 3DDS plots is helpful. From the top-left 4 panels we see that /t/ is defined by a 4-5.4 kHz burst of energy, ≈10 cs (100 ms) before the vowel, whereas /k/ is defined as a 1.4-2 kHz burst, also ≈10 cs before the vowel. The consonant /p/ shows up as a burst of energy between 0.7-1 kHz, sticking out in front of the vowel, but connected. The three voiced sounds /d/, /g/, and /b/ have similar frequencies but onset with the vowel. The case of /b/ is not obvious, and the low score seem to reflect this weak burst. Many of the sounds in our consonant database (≈100 consonants) were analyzed using this 3DDS method, and gave similar results.

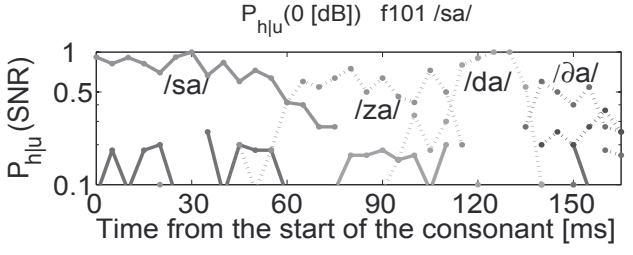



**Fig. 4:** Frication sound female 101 saying /sa/ (Exp. TR07). As the sound is truncated from the onset, the /s/ is heard as /z/, then /d/ and finally /ð/. Each time the conversion happens at about a factor of two in frication duration.

### Fricative sounds

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, a swath of bandwidth of fixed duration and intensity is used to indicate the sound.

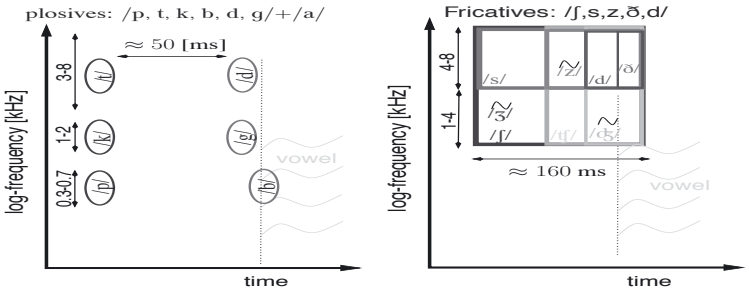



**Fig. 5:** Time-frequency allocation of the plosives and the fricatives. Mapping these regions into perceptual cues requires extensive perceptual experiments. Once the sounds have been evaluated, it is possible to prove how the key noise-robust perceptual cues map to acoustic features. The three consonants with the tilde over them (/z,ʒ,ʤ/), indicating that they are modulated at the pitch frequency, are voiced.

Using a time-truncation experiment similar to Furui (1986), as reported in Régnier and Allen (2008), we see the importance of duration to these consonants. In Fig. 4, a /sa/, spoken by female talker 101 and presented at 0 dB, was truncated in 10-ms steps. After about 60 ms of truncation from the onset of the sound, our pool of subjects reported /za/ instead of /sa/. After 30 additional ms of truncation, /d/ was heard. Finally at the shortest duration /ða/ was reported. A related experimental result found ʃa → ʧ→ ʤ→ d. At the end of this chain is the plosive. Thus the fricatives and the voiced-plosives seem to form a natural continuum, in the limit of very-short duration sounds.

The 3DDS results for the plosive consonants are summarized in the left half of Fig. 5, and for the fricatives in the right half of the figure. A small subset of acoustic cues define perceptual cues. Figure 5 is a modified version of the graphic by Alen and Li (2009), detailing the various *acoustic cues* for CV sounds, specifically with the vowel /a/, that were established to be perceptual cues, using a method denoted the *three-dimensional deep search* (3DDS) (Li *et al.*, 2010). Briefly summarized, 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 many low-level confusions. The recognition of burst-consonants 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 about 50 ms before the onset of F0 voicing while /d, g/ have a VOT <20 ms. Plosive /b/ may have a negative VOT.

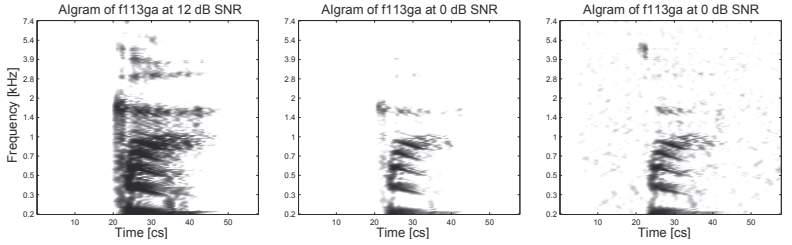



**Fig. 6:** On the left we see an AIgram of the original sound f113ga at 12 dB SNR, and in the middle, at 0 dB. The sound is identified 100% of the time, at and above 0 dB, 90% at $-6$ dB, and 30% at $-12$ dB. On the right is an AIgram of the sound after modification by the STFT method, where the mid-frequency burst at [20 cs, 1.5 kHz] was removed, along with remnants of the pre-vocalic burst, and 12 dB of gain was applied at 20 cs between 3.9-5.4 kHz, amplifying the low-level burst of energy, unmasked at 12 dB (left panel), as seen in the right panel. These two modifications resulted in the sound being reported as /da/.

### Verification methods

To further verify all these results we have developed a method to modify the speech sounds using *short-time Fourier transform* (STFT) methods (Allen, 1977; Allen and Rabiner, 1977), to attenuate and amplify these bursts of energy. These studies have confirmed that the narrow-band bursts of energy shown in Fig. 3 are both necessary and sufficient to robustly label the plosive consonants (Li and Allen, 2011). Above the feature's masked threshold, the score is independent of SNR (Régnier and Allen, 2008; Singh and Allen, 2012).

Verification methods using STFT modifications are exampled in Fig. 6. On the left is the unmodified sound at 12 dB SNR, and in the middle again the unmodified sound at 0 dB SNR. For the right panel the /g/ perceptual cue at 1.4-2 kHz has been removed and the /d/ perceptual cue between 4-5.5 kHz has been enhanced. Following the two modifications, noise was added at 0 dB. The two modifications resulted in the morph /ga/ $\rightarrow$ /da/.

### Summary

Based on such 3DDS results along with the verification experiments on the $\approx$100 CV in our database, we are confident that these bursts of energy label the identity of these consonants.

## CONFUSIONS IN HEARING IMPAIRED EARS

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

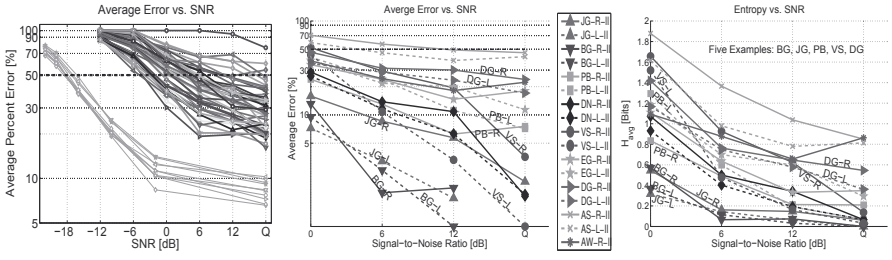



**Fig. 7: Left:** Average consonant error for 46 HI ears of Exp. I (Solid colored lines) and 10 NH ears (gray lines). **Middle:** Average consonant errors for the 17 HI ears of Exp. II (solid colored lines), as function of signal-to-noise ratio (SNR) using speech-weighted noise. **Right:** Average entropy for Exp. II.

The average error as a function of SNR for the 46 ears from Exp-I is shown on the left most panel of Fig. 7. The intersection of the thick horizontal dashed line at the 50% error point and the plotted average error line for each ear, marks the *consonant recognition threshold* (CRT) in dB. The data for 10 NH ears are superimposed as solid gray lines for comparison. NH ears have a similar and uniform CRT of $-18$ to $-16$ dB (a 2-dB range), while the CRT of HI ears are spread out between $-5$ to +28 dB (a 33-dB range). Three out of 46 ears had greater than 50% error in quiet (i.e., no definable CRT).

The data for the 17 ears (Exp. II) are mostly from the $<0$ dB CRT region, thus the mean error is much smaller (1% or so) compared to Exp. I, where the mean error is 15%. The minimum error for Exp. II is much lower because two high-error consonants [θ,ð Fig. 1(a)] were removed.

As discussed earlier the average score is a crude metric due to its high variance (i) across consonants, (ii) across utterances for each consonant, and (iii) across HI subjects, across both consonants and utterances. Entropy (Fig. 7, right) 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 HI ears, a high-consistency measure (i.e., entropy) seems like a better measure.

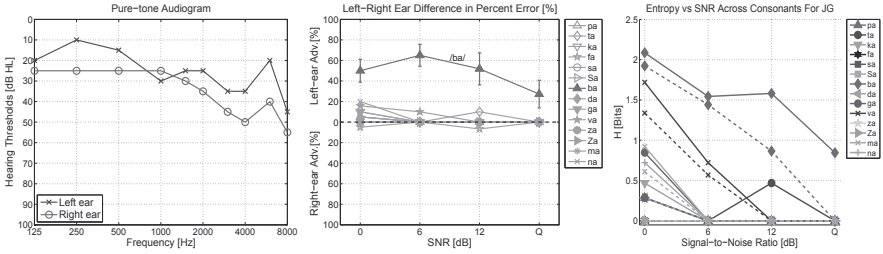



**Fig. 8:** Subject JG (HI36) has similar audiograms in the two ears, but a dramatic difference in the scores for /b/, of more than 50% difference between the two *consonant loss profiles* (ΔCLPs). On the right is the entropy for each consonant vs. SNR (dashed=left ear, solid=right ear).

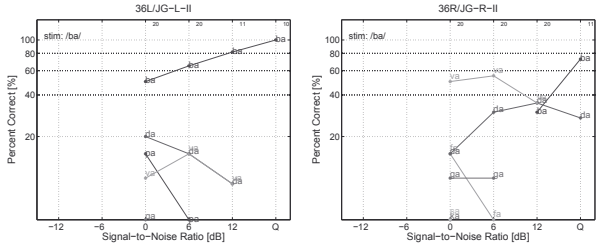



**Fig. 9:** Here we show the /ba/ confusion patterns $P_{h|s}(SNR)$ for subject JG (HI36). On the right we see that /b/ is confused with /v/ and /d/, even in quiet, while on the left the error is zero in quiet.

### Comparison between the audiogram and confusion patterns

The observation that HI ears can exhibit large individual differences in their average consonant loss given similar *pure tone average* (PTA) (Phatak *et al.*, 2009), is further supported by the data of Fig. 8. Subject JG (HI36) has (left panel) 10-20 dB better thresholds in the left ear (blue-x) and (right panel) has a large left-ear advantage for /ba/. In the middle panel is ΔCLP(SNR), defined as the difference in consonant scores between the ears, as a function of SNR. The left-ear advantage for /ba/ peaks at 6 dB SNR at 60%. Other than /b/, subject JG heard most consonants similarly in both ears (less than 20% difference), and with no difference in /pa/, whose burst spectrum has energy in the same frequency range of .3–2 kHz with /ba/. The results for HI36 in Exp. I, collapsed over SNR, showed little difference in consonant loss between

left and right ears. However in Exp. II a left-ear advantage in the /ba/ syllable was clearly indicated. This illustrates the utility of the 20 trials/condition for Exp. II, which allowed us to determine the loss as a function of SNR.

Subject HI30/DG (Fig. 10) has a 30-dB right-ear advantage for /za/, and has a distinct left-ear advantage for syllables /va, sa, fa/ and a 60% left ear advantage for /va/, at 12 dB.

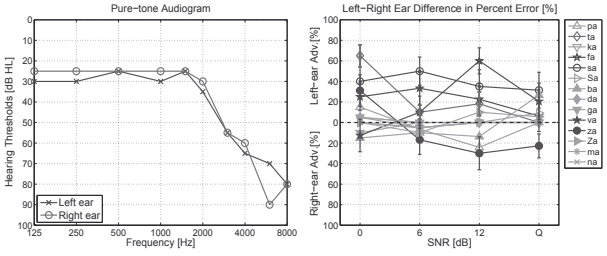



**Fig. 10:** Subject 30/DG L/R PTA (left) along with the difference in the confusions vs SNR (ΔCPL) on the right. While the two HLs are virtually identical, the scores are highly biased toward the left ear (left-ear advantage). These data are from Exp. II where the number of trials was up to 20 per consonant per SNR.

## Summary

This article has reviewed some of what we have recently learned about speech perception of consonants, and how this knowledge might impact our understanding of NL cochlear speech processing. The application of NL OHC processing in speech is still an under-developed application area (Allen, 2008; Alen and Li, 2009) Many new ideas and methods for testing and analysis have been suggested and evaluated. The jury is out.

It is now widely accepted that outer hair cells (OHCs) provide dynamic range and are responsible for much of the NL cochlear speech signal processing, thus the common element that link all the NL data (Allen *et al.*, 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 outer hair cell's two-way mechanical transduction is viewed as the key to solving the problem of the cochlea's dynamic range and dynamic response (Allen, 2003).

However, the perception of speech by the HI 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. Continued analysis of these confusions will hopefully provide further key 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 speech HI perception results provided here may provide further insight into normal speech perception.

The key open problem here is "How does the auditory system (e.g., the NL cochlea and the auditory cortex) processes human speech?" There are many applications of these results including speech coding, speech recognition in noise, hearing aids, 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 we 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 for their generous support of this research. Much of the work was supported by an NIH grant #RDC009277A.

## REFERENCES

Allen, J.B. (**1977**). "Short time spectral analysis, synthesis, and modification by discrete Fourier transform," IEEE T. Acoust. Speech, **25**, 235-238.

Allen, J.B., and Rabiner, L.R. (**1977**). "A unified approach to short-time Fourier analysis and synthesis," Proc. IEEE, **65**, 1558-1564.

Allen, J.B. (**1994**). "How do humans process and recognize speech?" IEEE T. Speech Audio P., **2**, 567-577.

Allen, J.B. (**1996**). "Harvey Fletcher's role in the creation of communication acoustics," J. Acoust. Soc. Am., **99**, 1825-1839.

Allen, J.B. (**2003**). "Amplitude compression in hearing aids," in *MIT Encyclopedia of Communication Disorders*. Edited by R. Kent (MIT Press, MIT, Boston, MA), Chapter IV, pp. 413-423.

Allen, J.B. (**2004**). "The articulation index is a Shannon channel capacity," in *Auditory signal processing: physiology, psychoacoustics, and models*. Edited by D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet (Springer Verlag, New York, NY), Chapter Speech, pp. 314-320.

Allen, J.B. (**2005**). *Articulation and Intelligibility* (Morgan and Claypool, 3401 Buckskin Trail, LaPorte, CO 80535), ISBN: 1598290088.

Allen, J.B. (**2008**). "Nonlinear cochlear signal processing and masking in speech perception," in *Springer Handbook on speech processing and speech communication*. Edited by J. Benesty and M. Sondhi (Springer, Heidelberg Germany), Chapter 3, pp. 1-36.

Allen, J.B., and Li, F. (**2009**). "Speech perception and cochlear signal processing," IEEE Signal Proc. Mag., **26**, 73-77.

Allen, J.B., Régnier, M., Phatak, S., and Li, F. (**2009**). "Nonlinear cochlear signal processing and phoneme perception", in *Proceedings of the 10th Mechanics of Hearing Workshop*. Edited by N.P. Cooper and D.T. Kemp (World Scientific Publishing Co., Singapore), pp. 93-105.

Blumstein, S.E., Stevens, K.N., and Nigro, G.N. (**1977**). "Property detectors for bursts and transitions in speech perceptions," J. Acoust. Soc. Am., **61**, 1301-1313.

Campbell, G.A. (**1910**). "Telephonic intelligibility," Phil. Mag., **19**, 152-159.

Cole, R., and Scott, B. (**1974**). "Toward a theory of speech perception," Psychol. Rev., **81**, 348-374.

Dorman, M., Studdert-Kennedy, M., and Raphael, L. (**1977**). "Stop-consonant recognition: Release bursts and formant transitions as functionlly equivialent, contextdependent cues,", Percept. Psychophys., **22**, 109-122.

Fletcher, H. (**1921**). "An empirical theory of telephone quality," AT&T Internal Memorandum, **101**.

Fletcher, H. (**1922**). "The nature of speech and its interpretation," J. Franklin Inst., **193**, 729-747.

Fletcher, H., and Galt, R. (**1950**). "Perception of speech and its relation to telephony," J. Acoust. Soc. Am., **22**, 89-151.

Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (**2004**). "New nonsense syllables database – analyses and preliminary ASR experiments," Proceedings of International Conference on Spoken-Language Processing (ICSLP).

Furui, S. (**1986**). "On the role of spectral transition for speech perception," J. Acoust. Soc. Am., **80**, 1016-1025.

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., and Allen, J.B. (**2012**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am., **131**, 478-491.

Li, F., and Allen, J.B. (**2009**). "Additivity law of frequency integration for consonant identification in white noise," J. Acoust. Soc. Am., **126**, 347-353.

Li, F., Menon, A., and Allen, J.B. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am., **127**, 2599-2610.

Li, F., and Allen, J.B. (**2011**). "Manipulation of consonants in natural speech," IEEE T. Audio Speech, **19**, 496-504.

Miller, G.A., and Nicely, P.E. (**1955**). "An analysis of perceptual confsions among some english consonants," J. Acoust. Soc. Am., **27**, 338-352.

Phatak, S., and Allen, J.B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am., **121**, 2312-2326.

Phatak, S., Lovitt, A., and Allen, J.B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am., **124**, 1220-1233.

Phatak, S.A., Yoon, Y., Gooler, D.M., and Allen, J.B. (**2009**). "Consonant loss profiles in hearing impaired listeners," J. Acoust. Soc. Am., **126**, 2683-2694.

Rayleigh, L. (**1908**). "Acoustical notes – viii", Philos. Mag., **16**, 235-246.
Regnier, M.S., and Allen, J.B. (**2008**). "A method to identify noise-robust perceptual features: application for consonant /t/," J. Acoust. Soc. Am., **123**, 2801-2814.
Shannon, C.E. (**1948**). "The mathematical theory of communication," AT&T Tech. J., **27**, 379-423 (parts I, II), 623-656 (part III).
Singh, R., and Allen, J.B. (**2012**). "The influence of stop consonants' perceptual features on the Articulation Index model," J. Acoust. Soc. Am., **131**, 3051-3068.
Yoon, Y., Allen, J., and Gooler, D. (**2012**). "Relationship between consonant recognition in noise and hearing threshold," J. Speech Lang. Hear. Res., **55**, 460-473.