NONLINEAR COCHLEAR SIGNAL PROCESSING AND PHONEME PERCEPTION

JONT B. ALLEN AND MARION RÉGNIER AND SANDEEP PHATAK AND FEIPENG LI

University of IL,
Urbana IL, jontalle@uiuc.edu

The most important communication signal is human speech. It is helpful to think of speech communication in terms of Claude Shannon's information theory channel model. When thus viewed, it immediately becomes clear that the most complex part of speech communication channel is in auditory system (the receiver). In my opinion, even after years of work, relatively little is know about how the human auditory system decodes speech. Given cochlear damaged, speech scores are greatly reduced, even with tiny amounts of noise. The exact reasons for this SNR-loss presently remain unclear, but I speculate that the source of this must be cochlear outer hair cell temporal processing, not central processing. Specifically, "temporal edge enhancement" of the speech signal and forward masking could easily be modified in such ears, leading to SNR-Loss. What ever the reason, SNR-Loss is the key problem that needs to be fully researched.

Keywords: speech; consonants; vowels; SNR-loss

1. 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. To address these issues, we need a theory of speech perception. Other than Claude Shannon's theory of information, depicted in Fig. 1, such theories are limited.

Consonant speech sounds are typically described in terms of production concepts, such as voicing, manner and place [1], however these categories tell us very little about the perception of speech, and nothing about the effect of masking noise on the received signal. It has not proved to be possible to generalize from copious examples, or the problem would have proved to be easy. Easy is not a word we may associate with this decoding problem.

It is clear from decades of research that the state of the cochlea is an important variable in speech perception studies. For example, *auditory masking* is critical to our understanding of speech and music processing. Furthermore, once the organ is damaged, our ability to process speech in noise is seriously impaired. The reasons for this impairment are not known, but it seems possible, or even likely, that such impairments are related to outer hair cell (OHC) processing in the cochlea.

The goal of this paper is to outline a theory of speech processing and to isolate the features in speech. We would like to answer questions as *What separates /t/ from /d/ or /t/ from /k/ and /p/?* and *Can we quantify the role of NL cochlear processing in this classification task?* We shall show that across-frequency onsets define the plosive consonants, while bandwidth and duration define fricative consonants. Finally we shall speculate on the role of the OHC processing in speech perception.

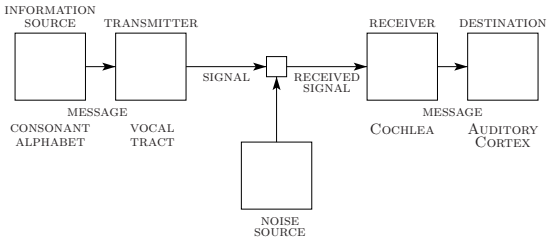

Fig. 1. Shannon's model of information transmission.

2. Key studies

The first speech studies were done in by Rayleigh (1908) [2], following telephone commercialization. Within a few years, Western-Electric's George Campbell (1910) [3] developed the electrical wave filter to high and low-pass speech signals, as well as probabilistic models of speech perception such as the *confusion matrix method* of analysis. With these tools established, by 1921 Harvey Fletcher was deeply into similar studies [4]. Fletcher 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 method*. Today we now know that it is closely related to information theory, introduced many years later by Claude Shannon (1948) [5], as summarized in Fig. 1.

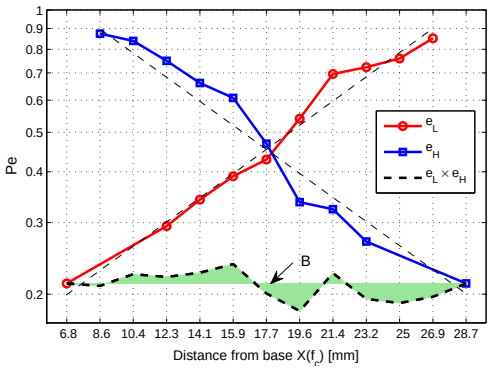

Fig. 2. This figure shows the core concept behind the *articulation index*. Speech is high and low-pass filtered to cutoff f_c, and two errors $e_L(f_c)$ and $e_H(f_c)$ are plotted on a log-error scale and cross near 1.5 kHz, at 17.7 mm along the cochlea, the abscissa scale. Each curve has been fit with a linear regressions having equal slopes but with opposite signs. Thus the product of the high and low-bands obeys Eq. 1.

In Fig. 2 we see a recent version of Fletcher's results, measured in my lab. Two complimentary filters are used, a high and low-pass with cutoff frequency f_c, with a 12 dB SNR masker. As the cutoff is varied, the average speech phone error is determined, as shown in the figure, where the probability of error $P_e(f_c)$ is on a log scale, as a function of position along the cochlea $X(f_c)$. While this data has been collected in our lab in 2006, it is similar in many ways to Fletcher's 1921 results. Fletcher demonstrated that the product of the low and high band errors is a constant, namely that $e_{total} = e_L(f_c) \times e_H(f_c)$ is a constant. The dashed line

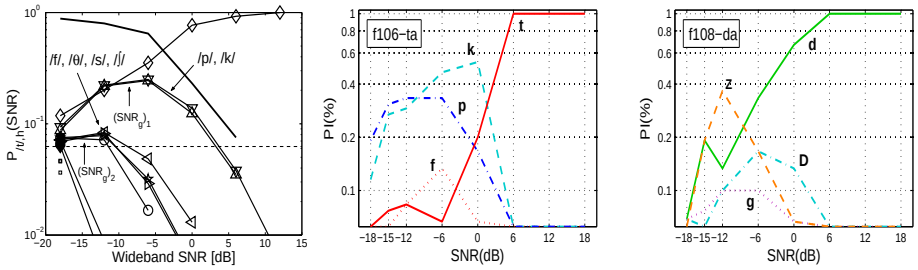

Fig. 3. Confusion patterns are defined by a row of the articulation matrix $\mathcal{A}(h|s)$, defined as the probability of hear sound h after speaking sound s, as a function of the signal to noise ratio (SNR), in dB. In each panel, the curve that rises to 1 is the diagonal element of the matrix, namely the score of /t/ given /t/ for the left and middle curves, and /d/ for the curves on the right. For the left panel, at -6 dB SNR, the *average* confusions with /p/ and /k/ are nearly equal to the score of /t/. Other confusions appear at even lower SNR, with /f/ for example. The other two panels are CPs for a specific /t/ (middle) and /d/ (right), with white noise maskers. Each of these sounds *morphs* (heard as another consonant) below 6 dB SNR. For example, for the middle panel at 0 dB, /k/ is reported 55% of the time while /t/ and /p/ are reported 20% and 18% of the time.

along the bottom, having an average error of 21%, has small fluctuations, labeled B (Bias), shaded in green. The average phone error $s = 1 - e_{total}$, may be computed from the articulation confusion matrix $\mathcal{A}_{h|s}(SNR)$, defined as the probability of hearing sound h after speaking sound s. As an example, confusions for the case of $s \equiv$ `/t/` are shown in Fig. 3, left.

The *Confusion patterns* (CP) shown in Fig. 3 allow one to determine the precise nature of the confusions of each sound. The confusion set, and their dependence on SNR is not predictable without running a masking experiment. These confusions, and their masked dependence, are important because they reveal the mix of underlying perceptual features, or *events*.

It is easy to create a sound that *primes*, meaning that it can be heard as any of several sounds, depending on one's state of mind. In this case the confusion patterns show subject responses that are equal (the curves cross each other), similar to the CP of Fig. 3 (left) at -10 dB, where one naturally primes /p/, /t/ and /k/, and on the right at -9 dB where /z/ and /d/ prime.

Fletcher found that the log-errors $e_L(f_c, SNR)$ and $e_H(f_c, SNR)$ are linear on a cochlea place scale $X(f_c)$ [6]. The implication is the total error may be generalized

$$e_{total} = e_1 e_2 \cdots e_{20}, \tag{1}$$

where e_k is the error contributing to the speech score due to cochlear band, indexed by integer k. This relationship is the key to the *articulation index* method, as reviewed in many places [6, 7, 8, 9, 10], and summarized in Fig. 4.

Along the top of the figure the response-measure is shown, such as the output of the cochlea or the phone scores. Just below the block diagram the mathematical model measure is displayed. For example, the output of the cochlea defines the signal to noise ratio in cochlear bands, as specified by the band articulation index AI_k for band k. Along the very bottom the figure indicates which are physical Φ measures and which are psychophysical Ψ discrete objects. The critical transition

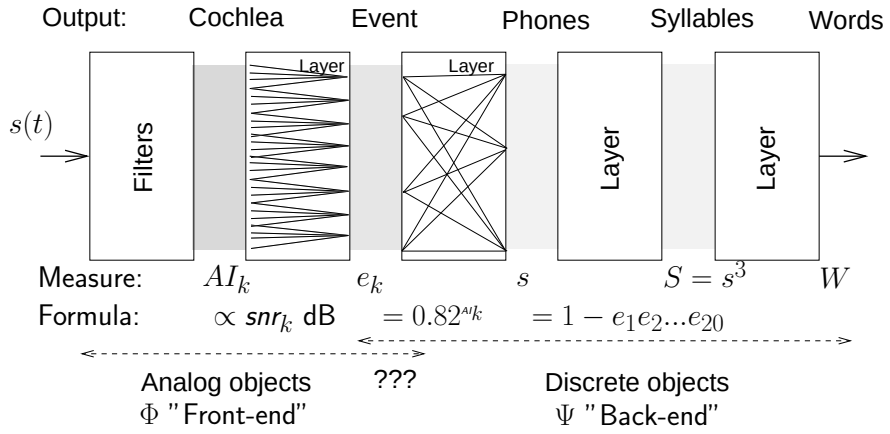

Fig. 4. Outline for the theory of speech perception. The output of the cochlea defines the signal to noise ratio in cochlear bands, as specified by the band articulation index AI_k for band k. The *event band error* is given by e_k. The maximum entropy average phone error is defined as $s(AI)$, also known as the *nonsense phone error*. The consonant-vowel (CV) syllable score S_{cv} is the square of the phone error while the CVC error is the cube (i.e., $S_{cvc} = s^3$).

from Φ to Ψ is presumed to happen at the event level [6]. Once a speech event is quantal, central processing is assumed to be error free.

2.1. *Identifying events*

Two methods have been established for precisely identifying *events* (perceptual features). The first method is outlined in a recent paper by Régnier and Allen (2008) [11]. Rather than reviewing this method here, since it is so recently published, we present a second, perhaps more general method, as yet unpublished.

2.1.1. *Speech-Plosive events*

In Fig. 5 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 was not available in our database, so three different talkers have been used in this analysis.

Each sound was first time-truncated from the onset to a given time, in 10 ms steps [12], and played back in random order to 14 listeners, who indicated what they heard by clicking an icon on the screen. Noise was added to the truncated sound at an SNR of 12 dB. The results of the truncation experiment (TR07) are presented in the upper-left panel with the title TR07 at the top. Each curve is the probability of what was reported, and is labeled with the identified consonant. In a second experiment (MN16R) the same sound was subjected to a variable signal to noise ratio, from -12 dB SNR to quiet, and the average score was measured across

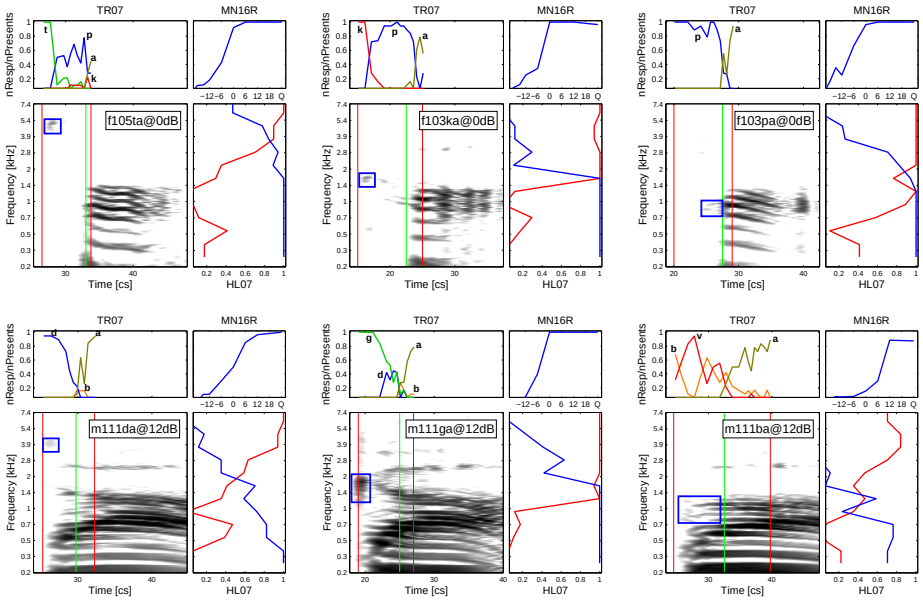

Fig. 5. Identification of features by time and frequency bisection. Along the top we have unvoiced consonants /t/, /k/ and /p/, while along the bottom, the corresponding voiced consonants /d/, /g/ and /b/. Each of the six sounds consists of 4 sub-panels. For example, for /t/ in the upper left wee see four panels consisting of the time-truncation confusions (upper-left), the score vs. SNR (upper-right), the AI-gram (lower-left) and the score as a function of low and high-pass filtering. There are six such groups, one for each of the six consonants displayed.

the 23 listeners [13]. Finally the same sample was high and low-pass filtered to a variable cutoff frequency (experiment HL07), as indicated on the frequency axis.

A summary of the audible sound features at the threshold of masking, are shown by the AI-gram [11], as exampled in the lower-left panel of Fig. 5. This plot is similar to a spectrogram, but differs in several important ways. First the AI-gram is normalized to the noise floor. This is similar to the cochlea which dynamically adapts to the noise floor due to OHC NL processing [14, 15], as discussed in Section 3. Second, unlike a fixed-bandwidth spectrogram, the AI-gram uses a cochlear filter bank, with bandwidths given by Fletcher critical bands (ERBs) [8]. 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 $AI_k(SNR)$ described in Fig. 4.

At the present time the AI-gram is imperfect in that it contains no forward masking, no upward-spread of masking, and no neural masking components. Much work remains to be done on time-domain NL cochlear models of speech.

One may identify the speech event from these displays. For example, the feature that labels the sound (e.g., /t/) is indicated by the red-square in the lower-left panel of each of the six sounds (e.g., next to the descriptor of the sound (e.g. `105ta@0dB`) there is a red box showing the burst of energy that defines the /t/ sound). When this burst is truncated, as in the TR07 experiment, the /t/ morphs to /p/. When masking noise is added to the sound, such that it masks the boxed region, the

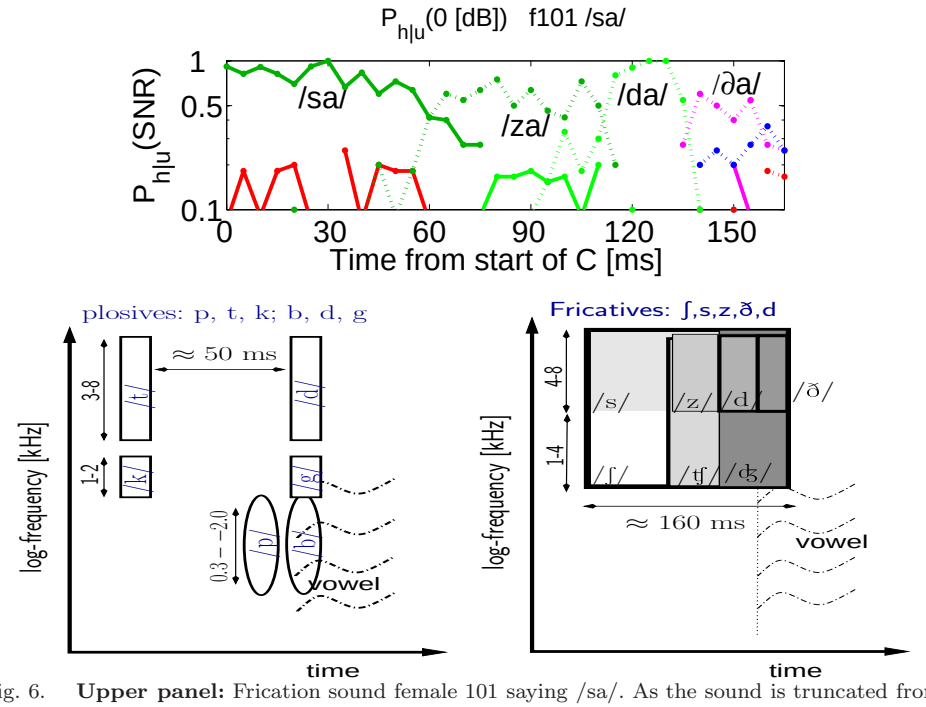

Fig. 6. **Upper panel:** Frication sound female 101 saying /sa/. As the sound is truncated from the onset, the /s/ is heard as /z/, then /d/ and finally /ð/. Each time the conversion happens at about a factor of two in frication duration. **Lower two panels:** Structure of the plosives and the fricatives, in terms of time-frequency allocation. Mapping these regions into events requires extensive perceptual experiments. But once the sounds have been evaluated, it is possible to prove where the key noise-robust events live in perceptual space.

percept of /t/ is lost. When the high and low-pass filters remove the frequency of the /t/ burst, again the consonant is lost. Thus the three experiments are in agreement, and they uniquely isolate the location of the event responsible for /t/. This nicely generalizes to the other plosive consonants shown (i.e. voiced /t/, /k/, /p/, and unvoiced /d/, /g/, /b/).

From such data we see that /t/ is labeled by a 4 kHz burst of energy ≈50ms before the vowel, whereas /k/ is defined as a 1-2 kHz burst, also 50 ms or so before the vowel. A burst of energy leading the vowel at 0.3-2 kHz defines /p/. The three voiced sounds /d/, /g/ and /b/ have similar frequencies but onset with the vowel .

The two high-frequency sounds (top and bottom left) are /t/ and /d/, each produced with the tongue tip on the roof of the mouth slightly behind the teeth. The two mid-frequency sounds, /k/ and /g/ are produced with the back of the tongue, labeled in the frequency domain as bursts between 1-2 kHz, for the examples shown. Finally low-frequency /p/ and /b/ are produced with the release of the lips. These two sounds produce a low frequency 0.2-2 kHz burst.

We have analyzed all the sounds in our consonant database, and similar results have been found. Thus we are confident that these tags of energy label the identity

of these consonant. The distributions of the burst frequencies, durations and delays to [1] voicing needs more study, as does the relationship between tongue place and burst frequency.

2.2. *Fricative sounds*

Not surprisingly, the events associated with fricative sounds are quite different from the plosives. Obviously 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, as shown in Fig. 6.

Using a time-truncation experiment similar to Furui (1986) [12], as disclosed by Régnier and Allen (2008) [11], we see the importance of duration to these consonants. In the top panel of Fig. 6, 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.

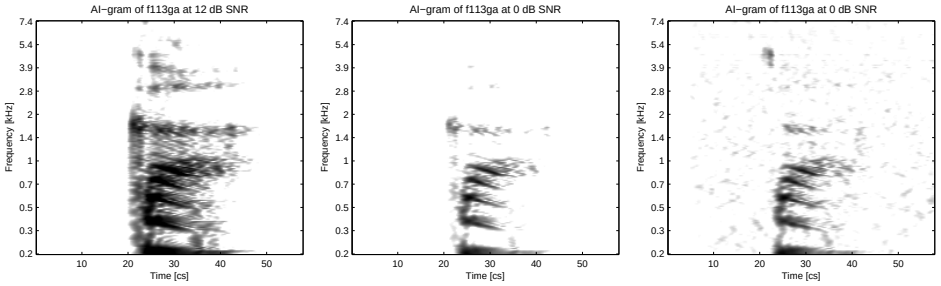

Fig. 7. On the left we see an AI-gram of the original sound f113ga at 12 dB SNR, and in the middle, at 0 dB. The score is 100% at and above 0 dB, and 90% at -6 dB. At -12 dB it is 30%. On the right is an AI-gram of the sound after modification by the STFT method, where the mid-frequency burst at [20 cs, 1.5 kHz] was removed, along with remnants of the pre-vocalic burst, and a 12 dB of gain was applied at 20 cs between 3.9-5.4 kHz, creating a burst of energy seen in the right panel. These two changes resulted in the sound being reported as /da/.

2.3. *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 [16, 17], to attenuate and amplify these Φ bursts of energy underlying the Ψ speech-events. These unpublished studies have confirmed that the narrow band bursts of energy shown in Fig. 5 are both necessary and sufficient to robustly label the plosive consonants. Above the feature's masked threshold, the event is independent of SNR [11].

Two STFT modifications are exampled in Fig. 7. For this case the /g/ event has been removed and the /d/ event enhanced, resulting in the morph /ga/ → /da/.

3. Nonlinear cochlear speech processing

The discussion next focuses on NL cochlear processing. Understanding and modeling NL OHC processing seems key to many speech processing applications. It seems under-appreciated that NL OHC processing (i.e., dynamic masking) is largely responsible for *forward masking* (post-stimulus masking), which results in very large effects over long time scales. For example OHC effects (FM/USM/2TS) can be as large as 50 dB, with a FM "latency" (return to base line) of up to 200 ms [18, 19, 15]. *Forward masking* (FM) and NL OHC *signal onset enhancement* are important to the detection and identification of perceptual features of a speech signal. In contrast, some studies have concluded that forward masking is not related to OHC processing [20, 21], so the topic remains controversial.

3.1. *Function of the Inner Ear in speech perception*

One key goal of cochlear modeling is to refine our understanding of how speech signals are processed [15]. The two main roles of the cochlea are to separate the input acoustic signal into overlapping frequency bands, and to compress the large acoustic intensity range into the much smaller mechanical and electrical dynamic range of the inner hair cell, synapse and neuron. This is a basic issue of signal, noise and information processing by the ear. The eye plays a similar role as a peripheral organ. It breaks the light image into rod and cone sized pixels, as it compresses the dynamic range of the visual signal. Based on the intensity JND, the corresponding visual dynamic range is about 9 to 10 orders of magnitude of intensity [22, 23], while the ear has about 11 to 12. The visual and auditory stimulus has a relatively high information rate compared to the low bandwidth of neural channels. The eye and the ear must cope with this problem by reducing the stimulus to a large number of low bandwidth signals. It is the job of the cortex to piece these pixelated signals back together, to reconstruct the world as we hear and see it.

Thus in general terms, the role of the cochlea is to convert sound at the eardrum into ≈30,000 neural pulse patterns in the human auditory ($VIII^{th}$) nerve. After being filtered by the cochlea, a low-level pure tone has a narrow spread of excitation which excites the cilia of about 40 contiguous inner hair cells [24, 8, 25]. The IHC excitation signal is narrow band with a center frequency that depends on the inner hair cell's location along the basilar membrane.

The prevailing and popular "cochlear amplifier" (CA) view is that the OHC provide *cochlear sensitivity* and *frequency selectivity* [25, 26, 27, 28]. The alternative view, argued here, is that the OHC compresses the excitation to the inner hair cell, thereby providing dynamic range expansion [29, 15].

There are key differences between these two views. The CA view deemphasizes the role of the OHC in providing dynamic range control (the OHC's role is to improve sensitivity and selectivity), and assumes that the NL effects result from OHC saturation. Such a simple model fails many comparisons to neural data. The *NL-compression* view places the dynamic range problem as the top priority. It

assumes that the sole purpose of the OHC nonlinearity is to provide dynamic range compression, and that the OHC plays no role in either sensitivity or selectivity, which are treated as important, but independent issues.[a]

3.2. *The dynamic range problem*

The question of how the large (120 dB) dynamic range of the auditory system is attained has been a long standing problem which remains fundamentally incomplete. Based on a simple noise analysis of the IHC membrane voltage, one may prove that the dynamic range of the IHC must be less than 65 dB [30]. In fact it is widely accepted that IHC dynamic range is less than 50 dB. The obvious question arises: *How can the basic cochlear detectors (the IHCs) have a dynamic range of less than 50 dB (a factor of* 0.3×10^2*), and yet the auditory system has a dynamic range of up to 120 dB (a factor of* 10^6*)?* This discrepancy in dynamic range forms a basic paradox. This would seem to be the necessary condition for the dynamic range compression, provided by OHC processing, that we are looking for. Indirect evidence has shown that this increased dynamic range results from mechanical NL signal compression provided by outer hair cells. This dynamic range compression shows up in auditory psychophysics and in cochlear physiology in many ways.

For example, *recruitment*, the most common symptom of neurosensory hearing loss, is best characterized as the loss of dynamic range [31, 8, 32, 14]. Recruitment results from outer hair cell damage [33]. To successfully design hearing aids that deal with the problem of recruitment, we need models that improve our understanding of *how* the cochlea achieves its dynamic range. Given the observations shown here on speech events, we need to extend our primitive understanding of *wide-dynamic range compression* [14] into the time domain.

As a second example, explaining the proportionality between the neural threshold in dB and the linear membrane voltage, is also key. Sewell (1984) [34] has demonstrated that as the EP voltage driving the hair cells changes, the neural gain in dB at CF changes proportionally by 1 dB/mv. It is not yet known why the dB gain is proportional to the voltage, however Sewell's observation might explain why cochlear forward masking decays exponentially in dB with time, after a strong excitation. Sewell's result implies an exponential decay with time of the neural sensitivity, *in decibels*. In other words, the log-decibel sensitivity should decay linearly in time. If $v_m(t) \propto e^{-t/\tau_m}$, then the dB value with change by a factor of $1/\mathrm{e} = 1/2.7 \approx 8.7$ [dB/dB] in τ [ms]. Given Sewell's result, a plot of the dB change in forward masking on a log scale would be linear. Just this relationship has been demonstrated by Duifhuis (1973) [18], who shows a slope of ≈30 dB in 100 ms. From this one would predict an OHC recovery time constant of 30/100=0.3 [dB/ms]. Thus these estimates define a release time constant for the OHC of the right form (linear in log-dB).

[a]Of course other views besides these two are possible.

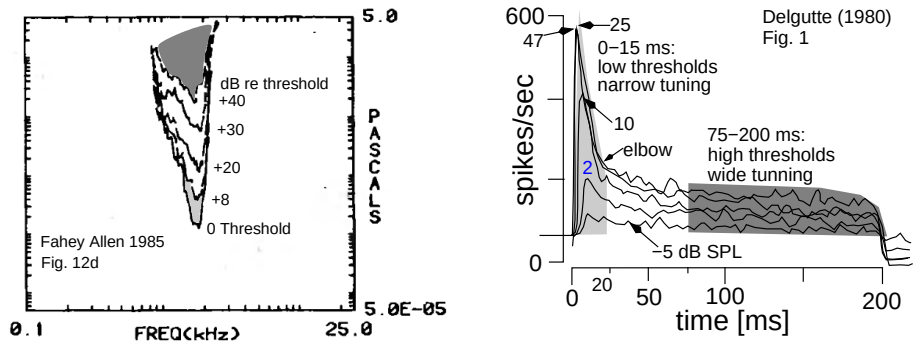

Fig. 8. On the left we see a typical 2 kHz tuning curve in various states of two-tone suppression, due to a low-frequency masker. These tuning curves are similar to what one would expect from driving the neuron with noise at various levels above threshold. On the right we see a PSTH to a 180 ms CF tone. The first 20 ms (to the point labeled elbow) shows a strong adaptation effect, due to the onset. At least some of this onset must be due to the high sensitivity of the neuron when a burst comes on, and reflects the time-course for it to reach a new state of tuning. It appears that this time is longer that one would expect from basilar membrane time constants, which are more like 100 μs (1/2 cycle of BM response). This effect will strongly enhance mid to high frequency transient onsets, such as those seen in speech with /t/, /k/, /g/, etc.

In Fig. 8 we relate the NL upward spread of masking seen in a typical 2 kHz neural tuning curve [35] to the neural rate-based *peristimulus time histogram* (PSTH) [19]. If we assume some level of noise is present, then the cochlear sensitivity must adapt to the level of noise. By way of an example, assume that the noise level is +20 dB re threshold (0 dB in Fig. 8). If a /t/ or /k/ burst then drives the nerve fiber, the rate will jump up, as shown in the right panel (adapted from Delgutte (1980) [19]). According to the PSTH response, the neural sensitivity will remain in the low-threshold, high-rate state, for up to 20 ms (elbow). Other data in [19] shows that this adaptation can forward mask (rate-suppress) a response up to 50 ms.

In summary, onsets will be enhanced by OHC processing, due to the overshoot seen in the auditory nerve PSTH functions of Fig. 8. In the hearing impaired ear, such enhancements would be gone, and this extra kick of response would not be available in those ears.

Summary: This article has reviewed some of what we have recently learned about speech perception of consonants, and how this might impact our understanding of NL cochlea speech processing of speech.

The application of NL OHC processing in speech processing is still an underdeveloped application area [15]. 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 learning 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 fea-

ture detection and phone recognition. While we have no idea when this will come to be, and it is undoubtedly many years off, when it happens there will be a technology revolution that will change human communications. The speech perception results shown here are relevant to this application.

Outer hair cells provide intensity dynamic range control and are responsible for the NL cochlear processing of speech. The OHCs are the one common element that link all the NL data previously observed, and a missing piece of the puzzle that most needs to be understood before any model can hope to succeed in predicting basilar membrane, hair cell, and 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. The OHCs membrane's τ_m RC recover time constant (defined as the τ_m with an OHC cilia admittance of zero) is a determining factor in the cochlear response recover time.

Acknowledgements

We gratefully acknowledge Phonak and the ECE department of UIUC for their generous support of this research.

Comments and Discussion

van Schaik (and others): The adaptation shown in Figure 1 of Delgutte (1980) in response to tone bursts is generally attributed to adaptation of neurotransmitter release in the synaptic cleft (see for instance the Meddis IHC model). If this is correct, surely the adaptation has nothing to do with a change in tuning bandwidth nor does it allow one to estimate the time constant of the change in bandwidth?

The adaptation at the IHC does indeed enhance onsets, but it is not clear to me if it necessarily follows that hearing impaired ears would not have this adaptation. If the mechanism for neurotransmitter release is normal in a hearing impaired ear, then the same onset enhancement would be available.

Allen's Response: I agree (See section 3, page 102) that the neural adaptation model you quote (Hewitt and Meddis [21]) is the widely held view. It is exactly for the reason you quote, that neural adaption would not lead to SNR-Loss, that I am suggesting there must be another mechanism. How else can we explain the widely observed SNR-Loss? The adaptation model has other serious flaws:

- It is widely believed that hearing impaired ears loose a natural robustness to noise, and effect sometimes called *SNR-Loss*. This cannot be accounted for by neurotransmitter adaptation, since there is no reason to believe that the synapse would be modified in the HI ears (as you point out in your question). Rather it is the OHC (e.g., their cilia) that are different.
- Forward masking data shows a "return to baseline" of 200 ms, with a linear in log-dB (a double log). How could such properties come from a simple synapse and its adaptation?
- Sewell found a 1 db/mv dependence between the EP and the tip of the tuning curve, which seems indicative of a mechanical induced transformation,

and it seem highly inconsistent with the adaptation model.

Thus we need something more than such a simple neurotransmitter adaptation to account for these other observations.

In my view, we are a long way from fully understanding nonlinear processing in the cochlea. Many people make a sweeping assumption that the CA explains all the the things they don't understand. Here are a few things that I believe: The first spike at the speech onset is highly significant (Heil). The role of the OHC is for signal dynamic range control, and as I have said many times, the evidence for the OHC's role in providing significant power gain (i.e., cycle-by-cycle) on the BM is quite limited. The dynamics of OHC processing are still not fully understood. As of yet, there are no forward masking data measured on the BM. Please reconsider my explanation of how Sewell's 1 dB/mv might come about [15].

References

1. Olive, J., Greenwood, A., and Coleman, J., 1993. Acoustics of American English Speech: A dynamic approach. Springer-Verlag, New York, Berlin, Heidelberg.
2. Rayleigh, Lord, 1908. Acoustical notes – viii. Philosophical Magazine, 16(6), 235–46.
3. Campbell, G. A., 1910. Telephonic intelligibility. Phil. Mag., 19(6), 152–9.
4. Fletcher, H., 1921. An empirical theory of telephone quality. AT&T Internal Memorandum, 101(6).
5. Shannon, C. E., 1948. The mathematical theory of communication. Bell System Tech. Jol., 27, 379–423 (parts I, II), 623–56 (part III).
6. Allen, J. B., 2005. Articulation and Intelligibility. Morgan and Claypool, 3401 Buckskin Trail, LaPorte, CO 80535. ISBN: 1598290088.
7. Allen, J. B., 1994. How do humans process and recognize speech? IEEE Transactions on speech and audio, 2(4), 567–77.
8. Allen, J. B., 1996. Harvey Fletcher's role in the creation of communication acoustics. J. Acoust. Soc. Am., 99(4), 1825–39.
9. Allen, J. B., 2005. Consonant recognition and the articulation index. J. Acoust. Soc. Am., 117(4), 2212–23.
10. Phatak, S. and Allen, J. B., 2007. Consonant and vowel confusions in speech-weighted noise. J. Acoust. Soc. Am., 121(4), 2312–26.
11. Regnier, M. S. and Allen, J. B., 2007. A method to identify noise-robust perceptual features: application for consonant /t/. J. Acoust. Soc. Am., 123(5), 2801–14.
12. Furui S., 1986. On the role of spectral transition for speech perception. J. Acoust. Soc. Am., 80(4), 1016–25.
13. Phatak, S., Lovitt, A., and Allen, J. B., 2008. Consonant confusions in white noise. J. Acoust. Soc. Am., 124(2), 1220–33.
14. Allen, J. B., 2003. Amplitude compression in hearing aids. In R. Kent, editor, MIT Encyclopedia of Communication Disorders, chapter Part IV, 413–23. MIT Press, MIT, Boston Ma.
15. Allen, J. B., 2007. Nonlinear cochlear signal processing and masking in speech perception. In Jacob Benesty and Mohan Sondhi, editors, Springer Handbook on speech processing and speech communication, chapter 3, 1–36. Springer, Heidelberg Germany.
16. Allen, J. B., 1977. Short time spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Trans. Acoust. Speech and Sig. Processing, 25, 235–38.
17. Allen J. B., and Rabiner, L. R., 1977. A unified approach to short-time Fourier analysis and synthesis. Proc. IEEE, 65(11), 1558–64.

18. Duifhuis, H., 1973. Consequences of peripheral frequency selectivity for nonsimultaneous masking. J. Acoust. Soc. Am., 54(6), 1472–88.
19. Delgutte, B., 1980. Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. J. Acoust. Soc. Am., 63(3), 843–57.
20. Relkin, E.M., and Turner, C.W., 1988. A reexamination of forward masking in the auditory nerve. J. Acoust. Soc. Am., 84(2), 584–91.
21. Hewitt, M.J. and Meddis, R., 1991. An evaluation of eight computer models of mammalian inner hair-cell function. J. Acoust. Soc. Am., 90(2), 904–17.
22. Hecht, S., 1934. Vision II. The nature of the photoreceptor process. In C. Murchison, editor, Handbook of General Experimental Psychology. Clark University Press, Worcester, MA.
23. Gescheider, G.A., 1997. Psychophysics: The Fundamentals, 3d edition. Lawrence Erlbaum Associates, Mahwah, NJ; London.
24. Allen, J. B. and Neely, S.T., 1992. Micromechanical models of the cochlea. Physics Today, 45(7), 40–47.
25. Dallos, P., 1996. Cochlear neurobiology. In P. Dallos, A.N. Popper, and R.R. Fay, editors, The cochlea, 186–257, Springer, New York.
26. Narayan, S.S., Temchin, A.N., Recio, A. and Ruggero, M.A., 1998. Frequency tuning of basilar membrane and auditory nerve fibers in the same cochleae. Science, 282, 1882–4.
27. deBoer, E. , 1996. Mechanics of the cochlea: modeling efforts. In The cochlea, 258–317 Springer, New York.
28. Geisler, D. C., 1998. From Sound to Synapse: Physiology of the Mammalian Ear. Oxford University Press.
29. Duifhuis, H., 1992. Cochlear modelling and physiology. In M. E. H. Schouten, editor, The Auditory Processing of Speech, 15–27, Mouton de Gruyter, Berlin.
30. Allen, J. B., 2001. Nonlinear cochlear signal processing. In A.F. Jahn and J. Santos-Sacchi, editors, Physiology of the Ear, Second Edition, chapter 19, 393–442. Singular Thomson Learning, 401 West A Street, Suite 325 San Diego, CA 92101.
31. Steinberg, J.C., and Gardner, M.B., 1937. Dependence of hearing impairment on sound intensity. Journal of the Acoustical Society of America, 9, 11–23.
32. Neely, S. T. and Allen, J. B., 1997. Relation between the rate of growth of loudness and the intensity DL. In W. Jesteadt and et al., editors, Modeling Sensorineural Hearing Loss, 213–222. Lawrence Erlbaum Assoc., Mahwah, NJ.
33. Carver, W.F., 1978. Loudness balance procedures. In Jack Katz, editor, Handbook of Clinical Audiology, 2^d edition, chapter 15, 164–178. Williams and Wilkins, Baltimore MD.
34. Sewell, W.F., 1984. The effects of furosemide on the endocochlear potential and auditory-nerve fiber tuning curves in cats. Hearing Res., 14, 305–14.
35. Fahey, P. F., and J. B., Allen, 1985. Nonlinear phenomena as observed in the ear canal, and at the auditory nerve. J. Acoust. Soc. Am., 77(2), 599–612.