Enrique A. Lopez-Poveda
Alan R. Palmer
Ray Meddis
Editors

The Neurophysiological Bases of Auditory Perception

Springer

The Neurophysiological Bases of Auditory Perception

Enrique A. Lopez-Poveda • Alan R. Palmer
Ray Meddis
Editors

The Neurophysiological Bases of Auditory Perception

Springer

Chapter 42
Identification of Perceptual Cues for Consonant Sounds and the Influence of Sensorineural Hearing Loss on Speech Perception

Feipeng Li and Jont B. Allen

Abstract A common problem for people with hearing loss is that they can hear the noisy speech, with the assistance of hearing aids, but still they cannot understand it. To explain why, the following two questions need to be addressed: (1) What are the perceptual cues making up speech sounds? (2) What are the impacts of different types of hearing loss on speech perception? For the first question, a systematic psychoacoustic method is developed to explore the perceptual cues of consonant sounds. Without making any assumptions about the cues to be identified, it measures the contribution of each subcomponent to speech perception by time truncating, high/low-pass filtering, or masking the speech with white noise. In addition, AI-gram, a tool that simulates auditory peripheral processing, is developed to show the audible components of a speech sound on the basilar membrane. For the second question, speech perception experiments are used to determine the difficult sounds for the hearing-impaired listeners. In a case study, an elderly subject (AS) with moderate to severe sloping hearing loss, trained in linguistics, volunteered for the pilot study. Results show that AS cannot hear /ka/ and /ga/ with her left ear, because of a cochlear dead region from 2 to 3.5 kHz, where we show that the perceptual cues for /ka/ and /ga/ are located. In contrast, her right ear can hear these two sounds with low accuracy. NAR-L improves the average score by 10%, but it has no effect on the two inaudible consonants.

Keywords Hearing loss • Perceptual cue • Consonant

F. Li (✉)
ECE Department, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA
e-mail: fli2@illinois.edu

E.A. Lopez-Poveda et al. (eds.), *The Neurophysiological Bases of Auditory Perception*,
DOI 10.1007/978-1-4419-5686-6_42, © Springer Science+Business Media, LLC 2010

42.1 Introduction

People with hearing loss often complain about the difficulty of hearing speech in noisy environments. Depending on the type and degree of hearing loss, a hearing-impaired (HI) listener may require a more favorable signal-to-noise ratio (SNR) than the normal-hearing (NH) listeners, to achieve the same level of performance for speech perception. The gap between the HI and NH listeners usually is unevenly distributed across different speech sounds. In other words, an HI listener may have serious problems with certain consonants, yet show no problem at all with other consonants.

State-of-the-art hearing aids have little effect in diminishing the gap, because they amplify everything, including noise, instead of enhancing the speech sounds that the HI listeners have difficulty with. As a consequence, many HI listeners now can hear the speech, with the assistance of hearing aids, but still they cannot understand it. To explain this, the following two questions need to be addressed. First, what are the perceptual cues of speech sounds? Second, what is the impact of hearing loss on speech perception?

Past study on hearing-impaired speech perception has two major limitations. First, most researches have been focused on the use of pure tone audiometry. For example, Zurek and Delhorne (1987) showed that the difficulties in consonant reception can be fully accounted on average, by the shift in the pure tone threshold for a group of listeners with mild-to-moderate hearing loss. Second, the impact of hearing loss on speech perception can only be assessed for speech on average, not for individual sound, due to the lack of accurate information on speech cues. Speech banana, a tool of qualitative assessment used in audiological clinic, is based on the pure tone audiogram and the formant data of vowels and consonants, which accounts for only part of the perceptual cues for speech sounds.

It is well known that most sensorineural hearing loss can be attributed to the malfunctioning of outer hair cells (OHCs) and inner hair cells (IHCs) within the cochlea. Damage to the OHCs reduces the active vibration of the cell body that occurs at the frequency of the incoming signal, resulting in an elevated detection threshold. Damage to the IHCs reduces the efficiency of mechanical-to-electrical transduction, also results in an elevated detection threshold. The audiometry configuration is not a good indicator of the physiological nature of the hearing loss (Moore et al. 2000), specifically, subjects with OHC loss and IHC loss may show the same amount of shifting in hearing threshold, yet the impacts of the two types of hearing loss on speech perception are very different. In the past decade, Moore and his students developed a threshold equalized noise (TEN) test (Moore et al. 2000) and a psychoacoustic tuning curve (PTC) test (Moore and Alcántara 2001) for the diagnosis of cochlear dead regions, an extreme case of IHC loss, which provides a psychoacoustic way of partitioning the two types of hearing loss.

In this chapter, we investigate the impact of cochlear dead regions on consonant identification. Based on the analysis of a large amount of data, it is hypothesized that speech sounds are encoded by some time-frequency energy onsets called acoustic cues.

Perceptual cues (events), the representation of acoustic cues on the basilar membrane (BM), are the basic units for speech perception. The HI listeners have difficulty understanding noisy speech because they cannot hear the weak events, missing due to the hearing loss and the masking effect introduced by the noise. The research work can be divided into two parts: identification of events for consonant sounds and influence of hearing loss on event reception.

42.2 Identification of Perceptual Cues

Speech sounds are characterized by three properties: time, frequency, and amplitude (intensity). Event identification involves isolating the cues along the three dimensions.

42.2.1 Modeling Speech Reception

In very general terms, the role of cochlear is to decompose the sound wave received at the eardrum through an array of overlapping narrow-band critical filters along the BM, with the base and apex of BM being tuned to the high-frequency (20,000 Hz) and low-frequency (20 Hz) respectively. Once a speech sound reaches the cochlea, it is represented by some time-varying energy patterns across the BM. Some of the energy patterns contribute to speech recognition, others do not. The purpose of event identification is to isolate the specific parts of the psychoacoustic representation that are critical for speech perception.

To better understand how speech sounds are represented on the BM, AI-gram (Lobdell 2006), a what-you-see-is-what-you-hear, WISIWYH (/wisiwai/) tool that simulates the auditory peripheral processing, is applied for the visualization of the speech sound. Given a speech sound, AI-gram provides an approximate image of the effective components that are audible to the central auditory system. However, it does not implicate which component is critical for speech recognition. To find out which component entertains the most significance for speech recognition, it is necessary to correlate the results of the speech perception experiments to the AI-grams.

42.2.2 Principle of 3D Approach

In this research, we developed a 3D approach (Li et al. 2009) to explore the perceptual cues of consonants from natural speech. To isolate the events along time, frequency, and amplitude, speech sounds are truncated in time, high/low-pass filtered, or masked with white noise, before being presented to normal-hearing listeners. Suppose an acoustic cue that is critical for speech perception has been removed or masked, it would degrade the speech sound and reduce the recognition score significantly.

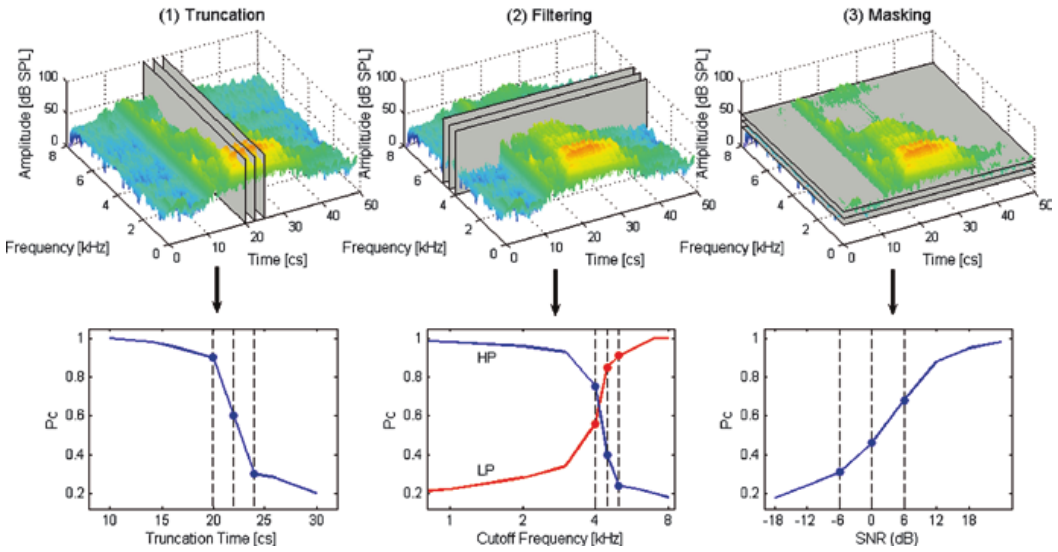

Fig. 42.1 3D approach for the identification of acoustic cues. The three plots on the *top row* illustrate how the speech sound is processed. The three plots on the *bottom row* depict the correspondent recognition scores for modified speech

For a particular consonant sound, the 3D approach (refer to Fig. 42.1) requires three experiments to measure the weight of subcomponent to speech perception. The first experiment determines the contribution of various time intervals by truncating the consonant into multiple segments of 5, 10, or 20 ms per frame depending on the duration of the sound. The second experiment divides the fullband into multiple bands of equal length on the BM and measures the importance of different frequency bands by using high-pass/low-pass filtered speech as the stimuli. Once the coordinates of the event are identified, the third experiment assesses the event strength by masking the speech at various signal-to-noise ratios.

42.2.3 Data Interpretation

The direct results of speech perception tests take the form of confusion patterns (CPs), which display the probabilities of all possible responses, including the target sound and the competing sounds, as a function of time, frequency, and amplitude. Since the probabilities of responses are not additive (Allen 1994; Fletcher and Galt 1950), in other words, the total contribution of multiple cues cannot be assessed by summing up the recognition scores of individual cues. Following the same idea of additivity of AI from multiple articulation bands, we defined a set of heuristic importance functions for the evaluation of contribution from different speech components.

Let $e_{\mathrm{L}}(f_k)$ and $e_{\mathrm{H}}(f_k)$ denote the low-pass and high-pass errors at the kth cutoff frequency f_{k}. The *frequency importance function* (FIF) is defined as

$$D_L(f_k) = \log_{e_0} e_L(f_k) - \log_{e_0} e_L(f_{k-1})$$

for the low-pass filtering case, or

$$D_H(f_k) = \log_{e_0} e_H(f_{k+1}) - \log_{e_0} e_H(f_k)$$

for the high-pass filtering case. The cumulative FIF is defined as the merge of the low-pass and high-pass values

$$D_F(f_k) = \max [D_L(f_k) \cup D_H(f_k)].$$

Generally, D_F should have its maximum value around the intersection point of the low-pass error e_L and high-pass error e_H that divide the full band into two parts of equal information.

The *event strength function* (ESF) is defined as

$$D_S(\text{snr}_k) = \log_{e0} e_S(\text{snr}_k) - \log_{e0} e_S(\text{snr}_{k-1})$$

where $e_S(\text{snr}_k)$ is the probability of error under the kth SNR condition. When a speech sound is masked by noise, the recognition score usually keeps unchanged until the noise hits a certain level snr_k (in dB) and has the event masked. Then, the recognition error e_S will increase dramatically and create a peak on the ESF at the corresponding position. Usually, the higher the peak, the more important the event is to the perception of the sound; the lower snr_k the more robust the sound is to noise.

The *time importance function* (TIF) is defined as the product of the probability of correctness, i.e., $1 - e_T(t_k)$ and the instantaneous AI $a(t_k)$ (Lobdell 2006)

$$D_T(t_k) = [1 - e_T(t_k)] \cdot a(t_k),$$

where $a(t_k)$ can be simply regarded as the energy of the speech signal at time t_k.

42.2.4 Perceptual Cues of Stop Consonants

In this section, we demonstrate how the events of consonants are identified by applying the 3D approach. To facilitate the integration of multiple information sources, the AI-gram and the importance functions are aligned in time or frequency, and depicted in a compact figure form, as shown by Fig. 42.2a.

Analysis reveals that the event of /ka/ is a mid-frequency burst around 1.6 kHz and articulated 50–70 ms before the vowel, as highlighted by the rectangular box in panel (a) of Fig. 42.2a. The event is strong enough to resist white noise at 0 dB SNR. Since the three importance functions all have a single sharp peak, which clearly shows where the event is, the process of event identification is straight forward. The TIF (panel (b) of Fig. 42.2a) has a distinct peak at $t = 165$ ms, and it

drops to zero after that. As soon as the burst is truncated, the perceptual score for /ka/ drops drastically. It is perceived as /pa/ thereafter. The FIF (panel (d) of Fig. 42.2A) has a peak around 1.6 kHz, where the intersection point of the high-pass and low-pass recognition scores is located, indicating that the mid-frequency burst is critical for the perception of /ka/. The above analysis is verified by the

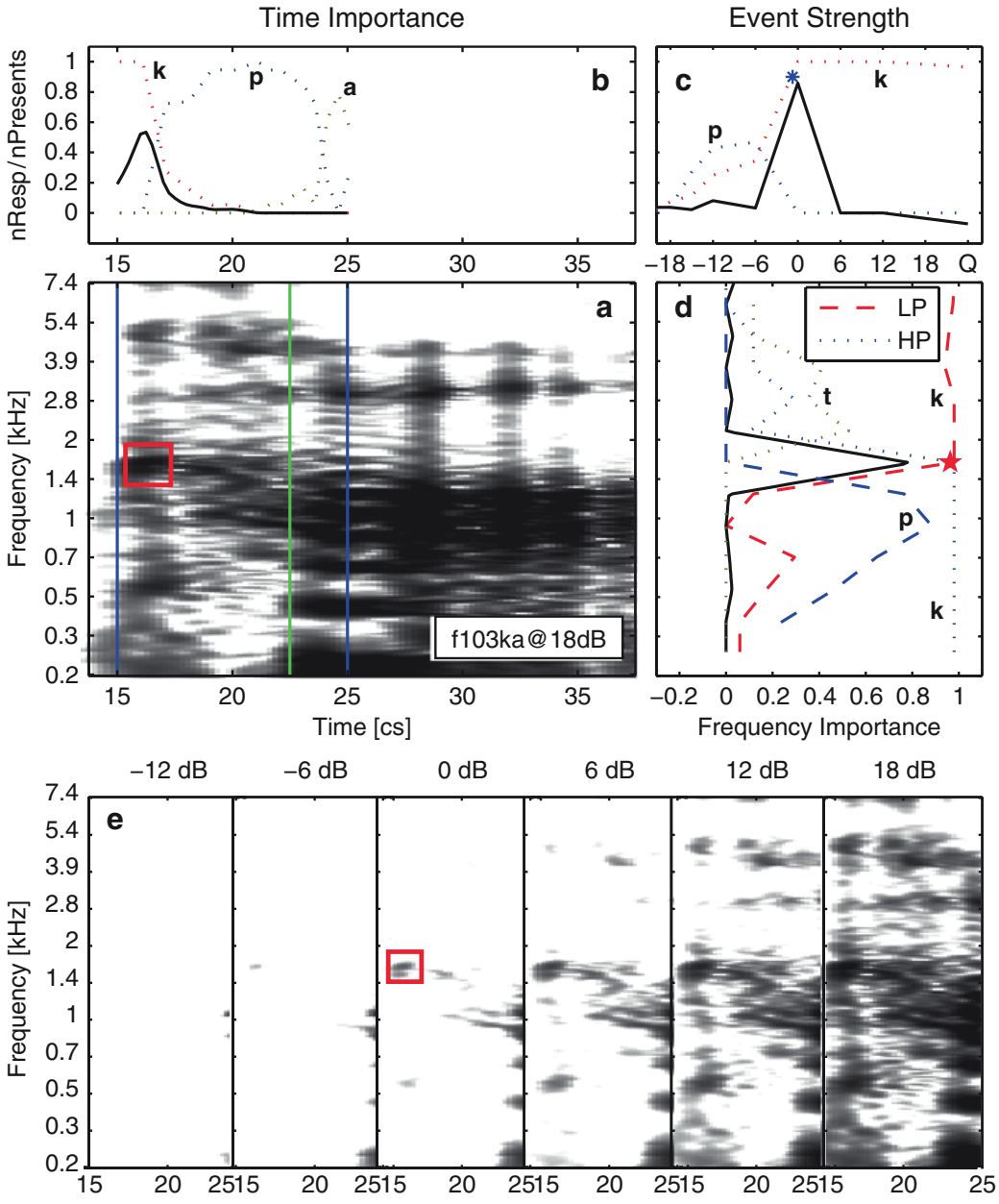

Fig. 42.2 Events of /ka/ by talker f103 (*left*) and /ga/ by talker m111 (*right*). (**a**) Identified events highlighted by *rectangular boxes* on the AI-gram. (**b**) TIF (*solid*) and CPs (*dotted*) as a function of truncation time. (**c**) ESF (*solid*) and CPs (*dotted*) as a function of signal-to-noise ratio. (**d**) FIF (*solid*) and CPs (*dotted*) as a function of cutoff frequency. (**e**) AI-grams of the consonant part at −12, −6, 0, 6, 12, 18 dB SNR

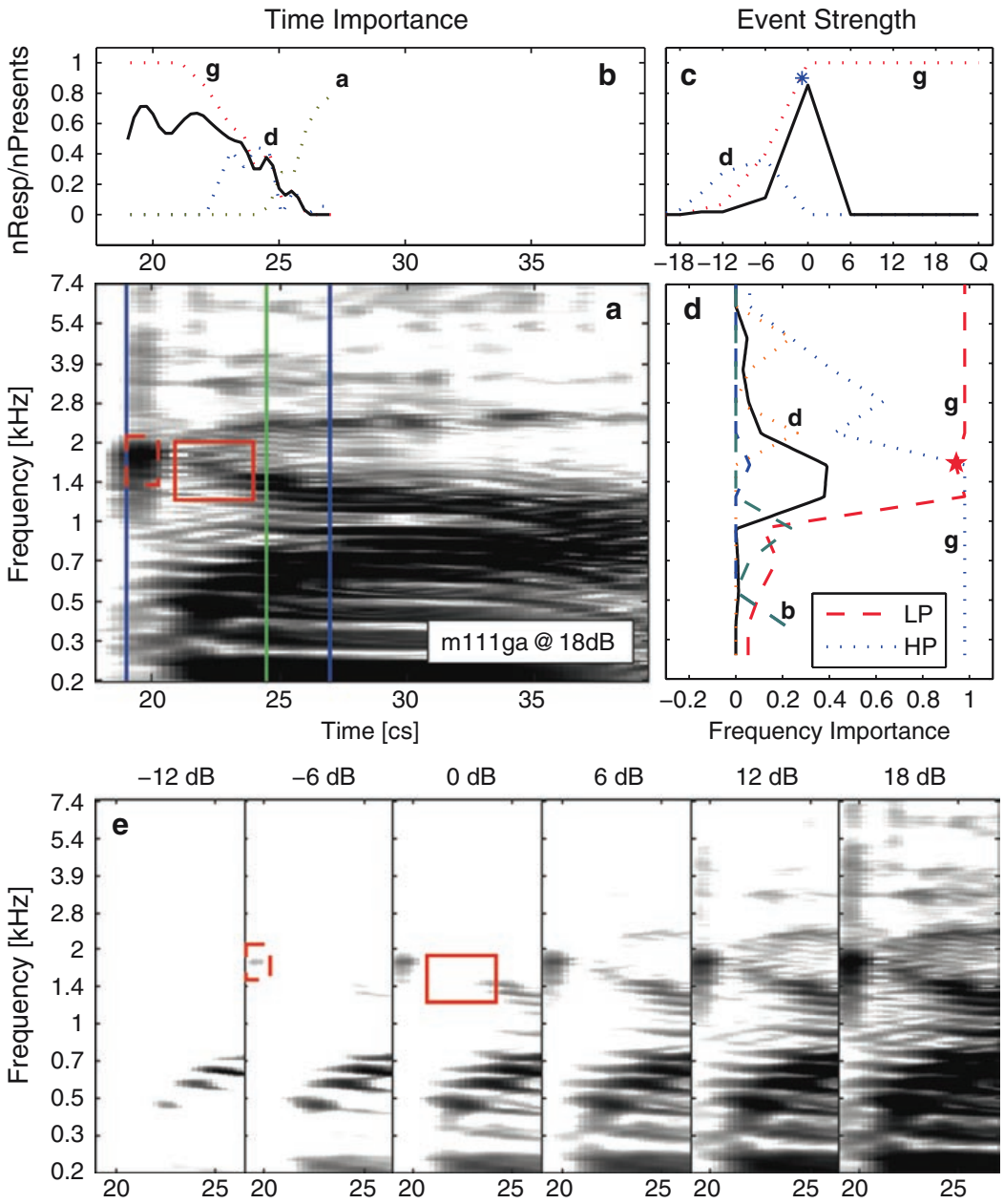

Fig. 42.2 (continued)

AI-grams under different SNRs (panel (e) of Fig. 42.2a), which shows that the /ka/ event becomes barely intelligible at 0 dB SNR. Correspondingly, ESF (panel (c) of Fig. 42.2a) has a sharp peak at the same SNR, where the recognition score of /ka/ begins to drop dramatically.

The events of /ga/ include a burst and a transition between 1 and 2 kHz as highlighted in panel (a) of Fig. 42.2b. The TIF in panel (b) of Fig. 42.2b shows two peaks at around $t=190$ ms and $t=230$ ms, where the burst and transition are located.

When the initial burst is truncated, the recognition score of /ga/ begins to drop, and the confusion with /da/ rapidly increases. The frequency importance function (panel (d) of Fig. 42.2B) has a peak at the intersection point of the high-pass and low-pass curves, which falls within the frequency range of the burst and the transition. Both the high-pass and low-pass scores show sharp decreases when the frequency range of 1–2 kHz is removed, indicating that the burst and the transition region are the perceptual cues for /ga/. This analysis is supported by the ESF and the AI-grams under various SNRs. When the F2 transition, the dominant cue, becomes masked, the recognition score of /ga/ from the masking experiment drops quickly and leaves a sharp peak on the ESF curve.

Through the same method, we also identified the events of other stop consonants preceding vowel /a/. Except for the two bilabial consonants, /pa/ and /ba/, which have a noise spike at the beginning, the stop consonants are characterized by a burst, caused by the sudden release of pressure in the oral cavity, and a transition in the second formant created by the jaw movement thereafter. Specifically, the events of /pa/ include a formant transition at around 1–1.4 kHz and a wide band noise spike at the beginning. /ba/ is characterized by a low frequency burst around 0.4 kHz and a transition around 1.2 kHz and a wide band noise spike in the range of 0.3–4.5 kHz. The event of /ta/ is a high frequency burst above 4 kHz. The /da/ event is a high frequency burst similar to the case of /ta/ in unvoiced sounds at around 4 kHz and an additional transition region at around 1.5 kHz. Generally, the noise spikes are much weaker than the bursts and transitions. The former usually disappear at 12 dB SNR, while the latter are still audible at 0 dB SNR.

42.3 Influence of Hearing Loss on Speech Perception

Depending on the nature of the hearing loss, an HI listener may have difficulty with a certain speech sounds. Finding the inaudible sounds that the HI listener cannot hear is the first step of solving the problem. With the identified events of consonants, the reason why an HI listener cannot hear certain sounds could be explained by integrating the speech events and the configuration of hearing loss.

42.3.1 Diagnosis of Hearing Loss

The degree of hearing loss is measured by the traditional pure tone audiometry. In addition to that, TEN test (Moore et al. 2000; Moore 2004) and PTC test (Moore and Alcántara 2001) are combined for the diagnosis of cochlear dead regions. Due to the time issue, the PTC test is used only to verify the existence of a cochlear dead region at certain frequencies suggested by the TEN tests. The procedures of the tests are controlled by a Matlab program.

42.3.2 Quantification of Consonant Loss

A speech perception experiment (SL07), using 16 nonsense CVs as the stimuli, is employed to collect the CPs across consonant sounds. There are two versions of the experiments: SL07a and SL07b. The only difference between the two versions is that the speech sounds are amplified by NAL-R in the latter in order to compensate for the hearing loss of the HI listener. NH Listeners only take the first test, while the HI listeners are instructed to take both of the tests. The results of each HI ear are compared to those of the average normal-hearing (ANH) listeners to determine the distribution of consonant loss. The detail of experiment SL07 is given below.

42.3.2.1 Speech Stimuli

Sixteen nonsense CVs: /p, t, k, f, T, s, S, b, d, g, v, D, z, Z, m, n/ + /a/ chosen from the LDC-2005 S22 corpus, were used as the common test material for both the HI and ANH listeners. The speech sounds were sampled at 16,000 Hz. Each CV has only six talkers, half male and half female. The speech sounds were presented through an ER-2 insert earphone to one ear of the subjects at the listener's most comfortable level. All experiments were conducted in a sound-proof booth.

42.3.2.2 Conditions

Besides the quiet condition, speech sound were masked at five different signal-to-noise ratios [−12, −6, 0, 6, 12] using speech-shaped noise. In SL07b, the noisy speech was amplified with NAL-R before being presented to the HI listeners.

42.3.2.3 Procedure

A mandatory practice session was given to each subject at the beginning of the experiment. Speech tokens were randomized across the talkers, conditions, and consonants. Following each presentation, subjects responded to the stimuli by clicking on the button labeled with the CV that he/she heard. In case the speech was completely masked by the noise, or the processed token did not sound like any of the 16 consonants, the subject was instructed to click a "Noise Only" button. To prevent fatigue, the subjects were asked to take a break whenever they feel tired. Subjects were allowed to play each token for up-to three times. A PC-based Matlab program was created for the control of the procedure.

42.4 Results

An elderly female subject (AS), mentally agile, trained in linguistics, volunteered for the pilot study. The subject has been wearing a pair of ITE hearing aids for more than 2 years.

42.4.1 Hearing Loss

Pure tone audiometry (Fig. 42.3) shows that AS has a bilateral moderate sloping hearing loss. Her left ear and right ear have similar configurations of hearing threshold with the PTA values being equal to 40 and 42 dB HL respectively.

Results of TEN tests (Fig. 42.4) suggest that subject AS may have a cochlear dead region around 2–3.5 kHz in the left ear, as the absolute hearing threshold and TEN-masked hearing threshold have a gap of more than 10 dB SPL at the frequencies of 2 and 3 kHz. This result is confirmed by the PTC test, in that the tuning curve of 2 kHz barely has a tip, while the turning curve at 3 kHz has a tip displaced in frequency. In contrast, her right ear has no cochlear dead region in the mid-frequency range, as suggested by the results of the TEN and PTC tests. The right ear may have a cochlear dead region around 8 kHz, since the absolute hearing threshold is greater than 120 dB SPL. PTC tests at 1 and 2 kHz show tuning curves of normal shape.

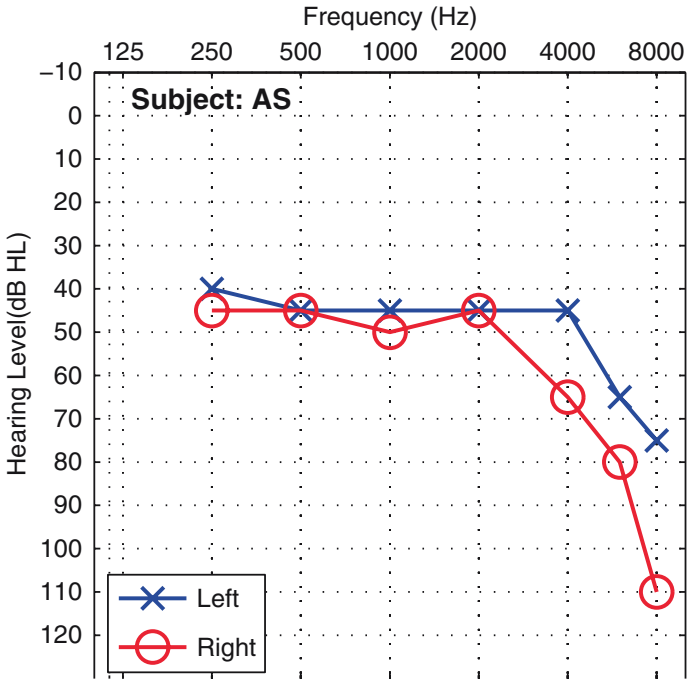

Fig. 42.3 Pure tone hearing threshold of AS

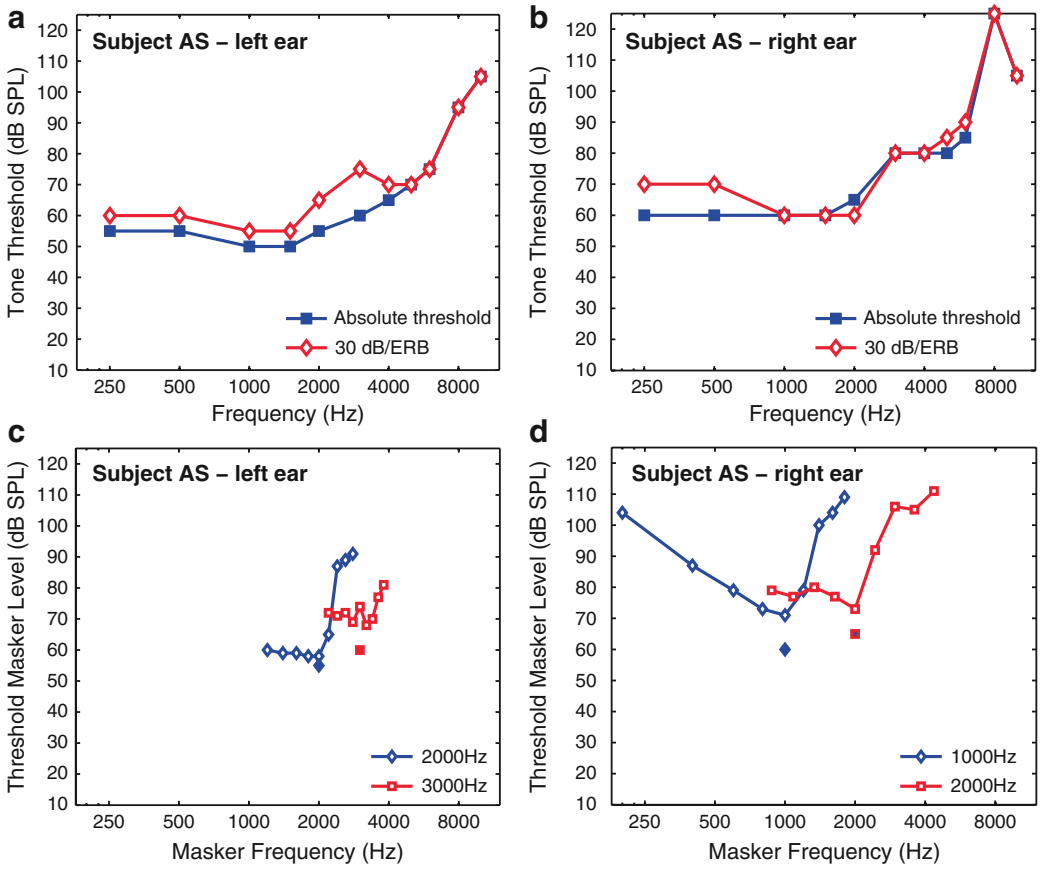

Fig. 42.4 Results of TEN and PTC tests for subject AS. *Upper panels* show the absolute and TEN-masked hearing thresholds in dB SPL. *Lower panels* show the PTCs for the frequencies of interest

42.4.2 Consonant Identification

The results of experiment SL07 indicate that cochlear dead regions may have a significant impact on the perception of speech sounds. Figure 42.5 depicts the recognition scores of subject AS for six stop consonants /pa, ta, ka, ba, da, ga/. Due to the cochlear dead region from 2 to 3.5 kHz, where the perceptual cues for /ka/ and /ga/ are located (Fig. 42.5), subject cannot hear these two sounds in the left ear. In contrast, her right ear can hear these two sounds with low accuracy, despite that the two ears have close configuration of hearing loss in terms of pure tone audiometry.

Confusion analysis shows that more than 80% of the /ka/s are misinterpreted as /ta/, while about 60% of the /ga/s are reported as /da/. A noteworthy fact revealed by the perceptual data is that NAL-R does not always help for the HI listeners. It increases the grand percent correctness (Pc) by about 10% for both ears in quiet, but it has no effect on the perception of the two inaudible consonants /ka/ and /ga/ for the left ear. Sometimes, it even degrades the speech sounds, for example, /ka/ for AS-R at 6, 12, and 18 dB SPL, due to unknown reasons.

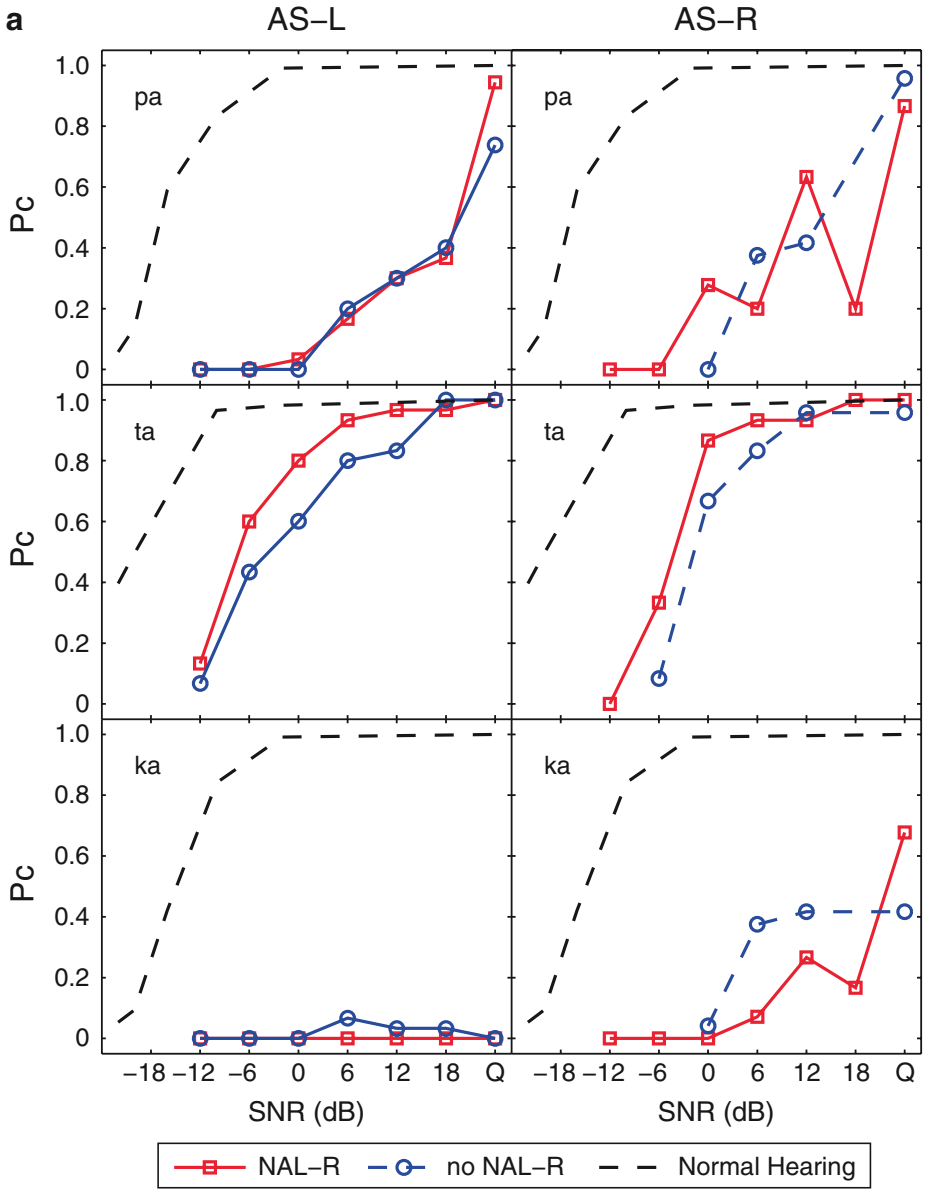

Fig. 42.5 Probability of correctness (Pc) for unvoiced stops /pa, ta, ka/ and voiced stops /ba, da, ga/ for subject AS. In both subfigures, panels on the *left column* show the data of the left ear (denoted by AS-L), and panels on the *right column* show the data of the right ear (denoted by AS-R). Perceptual data w/out NAL-R are depicted by the *square-masked* and *circle-masked curves* respectively. The recognition scores of ANH listeners are depicted by the *unmasked dashed curves* for comparison

42.5 Discussion

The pilot study presented above, in which the mentally healthy volunteer subject has similar configuration of audiometry for both ears with the left ear has a big cochlear dead region around 2–3.5 kHz, is an excellent case for the investigation of HI speech perception. The results support our hypothesis that the HI listeners have

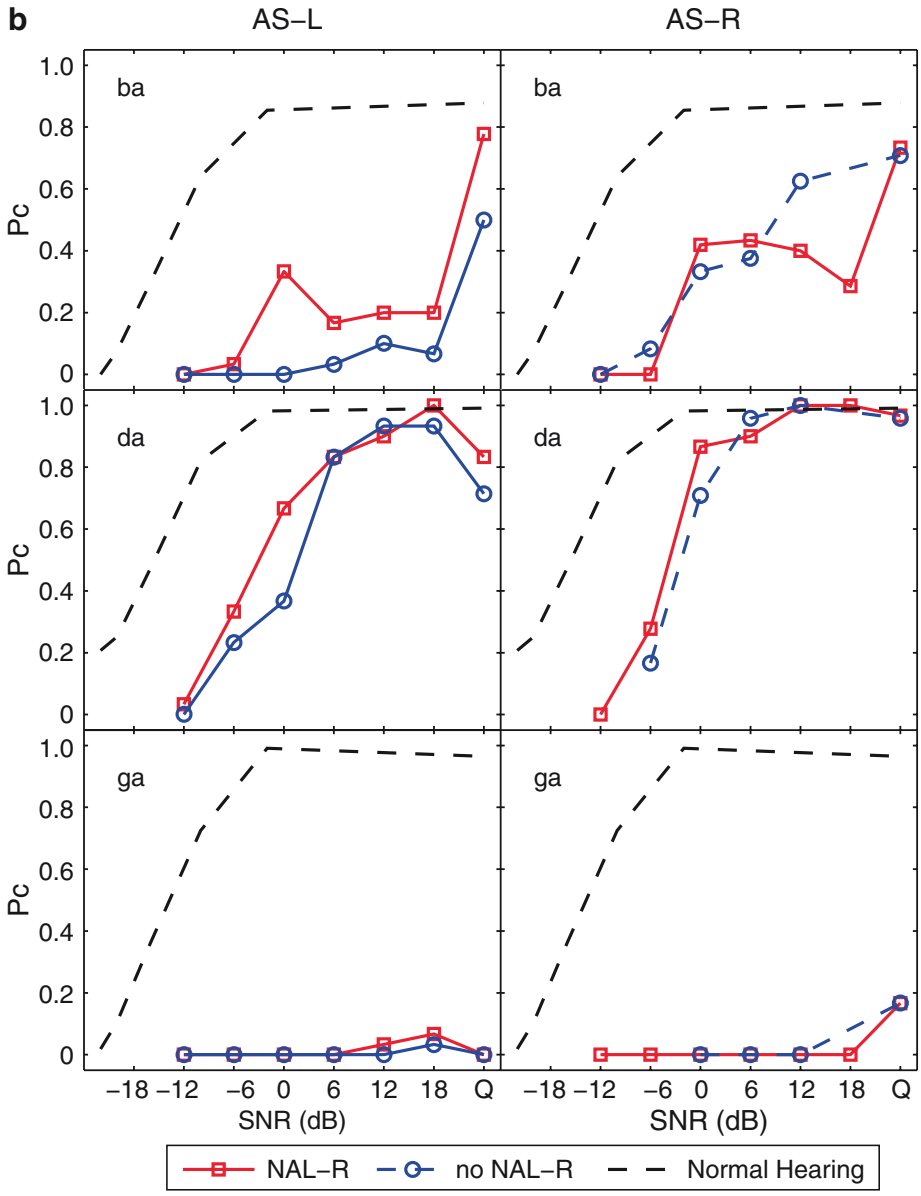

Fig. 42.5 (continued)

problem understanding speech because they cannot hear certain sounds whose events are missing because of their hearing loss or the masking effect introduced by the noise.

Other useful findings of the pilot study include: (1) Pure tone audiometry, and therefore the use of speech banana in audiological clinic, is a poor way of evaluating the problem of HI speech perception, especially for those who have cochlear dead regions; (2) cochlear dead regions have considerable impact on consonant identification and the effectiveness of NAL-R amplification; (3) speech enhancement can be regarded as a problem of optimization. As a static compensation scheme, NAL-R

improves the recognition scores of certain sounds, but it hurts the perception of other sounds. An adaptive amplification scheme that accounts for the HI listener's hearing loss as well as the speech characteristics may produce a better performance.

References

Allen JB (1994) How do humans process and recognize speech? IEEE Trans Speech Audio Process 2(4):567–577

Fletcher H, Galt R (1950) The perception of speech and its relation to telephony. J Acoust Soc Am 22:89–151

Li F, Menon A, Allen JB (2009) Perceptual cues in natural speech for 6 stop consonants. J Acoust Soc Am 127

Lobdell BE (2006) Information theoretic tool for investigating speech perception. MS Thesis, University of Illinois at Urbana-Champaign, Urbana, IL

Moore BCJ (2004) Dead regions in the cochlea: conceptual foundations, diagnosis, and clinical applications. Ear Hear 25(2):98–116

Moore BCJ, Alcántara JI (2001) The use of psychophysical tuning curves to explore dead regions in the cochlea. Ear Hear 22(4):268–278

Moore BCJ, Huss M, Vickers DA, Glasberg BR, Alcántara JI (2000) A test for the diagnosis of dead regions in the cochlea. Br J Audiol 34:205–224

Zurek PM, Delhorne LA (1987) Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. J Acoust Soc Am 82(5):1548–1559