PERCEPTUAL CUES OF CONSONANT SOUNDS AND IMPACT OF SENSORINEURAL HEARING LOSS ON SPEECH PERCEPTION

BY

FEIPENG LI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Doctoral Committee:

Associate Professor Jont Allen, Chair
Professor Douglas Jones
Associate Professor Mark Hasegawa-Johnson
Associate Professor Robert Wickesberg
Assistant Professor Michael Heinz, Purdue University

ABSTRACT

This research investigates the impact of various types of cochlear hearing loss and masking noise on the perception of basic speech sounds based on the information of identified speech cues. A psychoacoustic method, named three-dimensional deep search (3DDS), is developed to identify the perceptual cues of consonant sounds in natural speech. Unlike the conventional method of synthetic speech, which requires a prior hypothesis about the acoustic cues to generate the speech stimuli, the 3DDS measures the contribution of each subcomponent to speech perception as a function of time, frequency and intensity, without making any tacit assumptions about the speech cues to be identified. Using the 3DDS, we discovered that natural speech often contains conflicting cues that are characteristic of confusable sounds. For instance, a normal /ka/, dominated by a mid-frequency burst at 1–2 kHz, may also have an inaudible /ta/ burst above 3 kHz that promotes the /ka/→/ta/ confusion under noisy environments. Removal of the /ka/ burst may turn the sound into a solid /ta/.

More than a dozen hearing-impaired ears were tested on consonant identification in noise. While the deterioration in performance for flat mild-to-moderate hearing loss can be well predicted by the loss of audibility, subjects with other types of hearing loss often show patterns of difficult sounds that can hardly be explained by the shift of hearing threshold. A subject with almost identical binaural hearing loss is nearly deaf to /ka/ in one ear due to a mid-frequency cochlear dead region. Among the 18 /ka/s produced by different talkers, the subject can only hear one /ka/ at an accuracy of 80% and three other /ka/s at 20–40%. Most /ka/s are highly confused with /ta/ because the subject is listening to the conflicting /ta/ burst in the high-frequency. The /ka/→/ta/ confusion is significantly reduced when the conflicting cue is removed. NAR-L improves the average score by 10%, but it may degrade a few consonants under certain circumstances.

To my father and mother

ACKNOWLEDGMENTS

There are two people without whom this thesis would have been impossible: my supervisor Jont Allen, who initiated the project, and committee member Bob Wickesberg, who guided me through the literature review. I want to say thank you to Jont for taking me as a student when I was in trouble and for having faith in me when I was slow. I'll always remember the famous quote of Mahatma Gandhi, cited by Jont from time to time in our HSR group meeting, "First they ignore you, then they ridicule you, then they fight you, then you win." It was an honor to be a member of this group.

Three years ago while I was still not sure about what to do for my PhD thesis, both Bryce Lobdell and I took Bob's class. We met every Thursday morning at the Starbucks on Green Street and read a classic paper on speech perception. Those meetings were really pleasant and helpful! Every now and then when I had questions about my research and many other things, I came to Bob for advice. I want to thank Bob for the coffee and the invaluable, wise advice.

I would like to thank committee member Mark Hasegawa-Johnson for many insightful comments on my research. I also want to thank Douglas Jones and Mike Heinz, whom I feel privileged to have on my committee.

A number of people have contributed substantially to this work. Bryce Lobdell created the powerful AI-gram for the visualization of speech perception. Woojae Han and Riya Singh collected some of the hearing impaired data. Anjali Menon worked closely with me in analyzing the perceptual data for speech cue identification. Andreas Trevino, Len Pan and Roger Serwy helped proofread portions of the manuscript.

Special thanks go to the volunteer subjects; without their contribution, this thesis would not exist.

This research has been support by Etymotic Research, Phonak and in part by NIH Grant RDC009277A.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Approximately 15% of the general population (33% of individuals over the age of 65 and 50% over 75) have a hearing impairment (HI) that negatively impacts their speech communication skills [1, 2]. Once they stop hearing, they stop talking, feelings of isolation develop, and depression becomes common. Thus it is a medically significant concern for the elderly who are hearing impaired to remain verbal.

1.1 Problem Statement

HI listeners may have difficulty understanding noisy speech because they cannot hear certain sounds for which the characteristic speech cues are missing, due to their hearing loss and the masking effect introduced by noise [3]. During the past three years, we have tested over 40 HI ears for their recognition performances on consonants using 16 initial consonants /p, t, k, f, s, θ, ʃ, b, d, g, v, z, ð, ʒ, m, n/ preceding vowel /a/ as the test stimuli. In [4, 5], 16 HI subjects with mild to moderate hearing loss (30 dB HL $<$ PTA $<$ 50 dB HL), a hearing loss believed to involve no cochlear dead regions, participated in the experiment. Most listeners showed good performance on /d, g, k, m, p/ but had difficulty with /b, s, ʒ, z, v/. Recently we repeated the same test on an elderly subject (AS) with moderate hearing loss. Due to a *cochlear dead region* (an extreme case of IHC loss [6]) from 2 to 3 kHz, AS cannot hear /ka/ and /ga/ with her left ear. In contrast, her right ear can identify /ka/ and /ga/ (with low accuracy), despite the fact that the two ears have almost identical hearing thresholds. Confusion analysis reveals that more than 80% of the /ka/s perceived by the left ear are misinterpreted as /ta/, while about 60% of the /ga/s are reported as /da/.

In addition, hearing loss may reduce the HI listeners' ability to focus on one talker in a noisy environment. The human auditory system groups together acoustic elements

related to a target speaker, allowing the listener to hear a single voice. The cocktail-party effect [7] is defined as a normal hearing listener's ability to isolate target speech from the noisy background. For the HI listeners, the acoustic signals are corrupted with babble and then distorted by their hearing loss, which makes it difficult to group the acoustic elements and identify the speech source [8]. The impact of noise on speech perception is tremendous. Most of our HI subjects showed difficulty hearing speech at 12 or 6 dB SNR in speech-shaped noise, a noise level that causes little trouble for normal hearing people [9].

State-of-the-art hearing aids have low functionality in noisy speech because they amplify the entire signal without taking into account the specific features of the speech sounds. Over the past years, various single-channel noise-reduction techniques have been proposed to increase the SNR [10, 11]. For example, Time-Frequency Gain Manipulation [12] improves the total SNR by assigning larger gains to the time-frequency components with less noise and lower gains to those with more noise. Since the manipulation is based on the distribution of random noise rather than on prior knowledge about speech spectra, none of these methods have been shown effective in improving speech intelligibility [13]. As a consequence, many HI listeners can hear the amplified noisy speech, but still cannot understand it. To help those people, it is necessary to know more about speech perception.

1.2 Human Speech Perception

Speech perception is a complex process that involves multiple stages of signal processing. Once the acoustic signal reaches the human cochlear, it is decomposed into many critical bands on the basilar membrane. The cochlear nucleus then encodes the temporal and frequency information in a way that is meaningful to the central auditory system.

A major goal of speech perception research is to determine how the speech information is represented across the various stages. The research methods can be classified into three major types: psychophysical, computational and neurophysiological. The psychophysical approach [14–16], initiated by Harvey Fletcher and his colleagues in the 1920s, involves presenting subjects with speech stimuli and measuring their conscious responses, without touching the intermediate speech decoding process within

the auditory system. The computational models [17] are created for the simulation of speech perception behavior observed in psychoacoustic tests. The neurophysiological approach [18–21] measures the detailed information of single-unit neuron response to trace the representation of speech signal through the subsequent stages of auditory processing.

After about 100 years of work, very little is known about how the ear decodes basic speech sounds. This is, in part, because it is not ethical to record in the human auditory nerve, and its not practical to do extensive speech psychophysics in non-human animals.

1.2.1 Theory and models

Speech perception depends on the analysis of the continuous acoustic signal into a sequence of discrete phonetic segments [22]. How is the discrete phonetic percept related to the continuous signal? A widely accepted argument is that perception of phonetic segments is based on the context-dependent cues, which are interpreted phonetically in different ways depending on the nature of the context [23]. Other people believe that the properties of speech can be uniquely and invariantly specified from the acoustic signal itself and that these properties are closely related to the distinctive features. Speech perception is a process of decoding the sound into a representation of distinctive features [16, 24, 25]. In order to account for the transform from the variable speech signal to the linguistic units, a lot of theory and models of speech perception have been proposed in the literature.

Auditory Scene Analysis: In a cocktail party [7], a normal hearing listener must isolate the target speech from the noisy background in order to communicate with his conversation partner. Since the acoustic signals are corrupted with babble speech and various nonvocal party sounds, what makes the human auditory system group together the acoustic elements related to the target speech so that the listener hears a single voice? Based on the fundamental principle of Gestalt perception that the human brain tends to order our experience in a manner that is regular, orderly, symmetric, and simple, Bregman proposed a theory of auditory scene analysis [26–28], according to which speech organization consists of two stages: primitive grouping and schematic grouping. The first stage groups speech sound by the fundamental principles of Gestalt

perception, such as proximity in frequency, similarity in change, common fate, etc. The second stage applies learned knowledge to correct the mistakes made in the first stage. The combination of the two mechanisms allows the primitive principles to err without harm to the final output.

Motor Theory: Due to the effect of coarticulation, which greatly affects the muscle contractions of articulators (e.g., tongue, lips, and vocal folds), the mapping between the phonemes and the acoustic signals is quite complex. In contrast, the perceived phonemes and features seem to have a simpler (i.e., more nearly one-to-one) relationship to articulation than to acoustics. In the 1960s Liberman [29, 30] formulated the motor theory (MT), which soon became the dominant account of human speech perception in the following decades. The motor theory claims that the objective of speech perception is articulatory events rather than acoustic or auditory events. Speech perception depends on a special speech module unique to humans and innately organized that recovers the neuromotor commands to the articulators, also referred to as intended gestures, from the speech signal. It follows from the motor theory that mammals and birds, without the special speech module, should not be able to recognized speech. However, results on speech perception in nonhumans demonstrate that speech perception is not special. Birds and animals [31–33] exhibit aspects of speech perceptual performance that cannot be explained by the motor theory.

Direct Realist Theory: Fowler [34], a colleague of Liberman at the Haskins Labs, modified the MT and developed the direct realist theory (DRT). Unlike the MT, which requires a special module for the recovery of intended gesture from the acoustic signal, DRT claims that speech perception is based on the actual gesture that structures the acoustic signal, which helps in addressing why birds and animals can recognize speech. However, it does not explain how the mapping from speech signals onto the vocal tract shape that produced them can be realized. On the other hand, there are researches indicating that the inverse problem is intractable without prior knowledge about the shape of the vocal tract.

COHORT Theory: In the 1970s, Marslen-Wilson and Tyler conducted a series of studies on word identification. The psychological data shows that words in context are recognized within 200 ms, on average, from the onset of the words. At that time the

sensory information is usually insufficient by itself to identify the word being heard. Contextual constraints such as phonological, morphological, syntactic, and semantic information must also play a role in the process of word perception. The fact that the sensory and contextual constraints converge on their target in such a short time rules out the possibility of a strict serial processing of the information. Based on the observations, they proposed a COHORT [35] theory of spoken word recognition, in which the process of spoken word recognition breaks down into the three basic functions of access, selection and integration. Access concerns the mapping of the speech input onto the representations of lexical form, selection concerns the discrimination of the best-fitting match to this input, and integration covers the mapping of syntactic and semantic information at the lexical level onto higher levels of processing. Early in the auditory presentation of a word, those words known to the listener that conform to the sensory information received so far become active and form a list of candidates called "word-initial cohort." As more sensory and contextual information becomes available, some of the words belonging to the "word-initial cohort" are eliminated because they are inconsistent with the new information. The process of elimination continues until at one point only one word is left, which is called the "recognition point." In the first version of COHORT theory, contextual information was allowed to interact with the sensory information very early in the process of candidate elimination. This was changed in the revised version [36], in which the effect of context is very limited until the recognition point, which basically converted the COHORT theory from a both bottom-up and top-down structure into a bottom-up only structure.

TRACE model: McClelland and Elman [17] simulated the process of human speech perception by applying an artificial neural network, named the TRACE model. The model consists of a large number of units organized into three levels: the feature, phoneme, and word levels. The input to the model was a series of vectors representing the features of the mock utterance, specifically, consonantal, vocalic, diffuseness, acuteness, voicing, power, and amplitude of noise. Information processing takes place through the excitatory and inhibitory interactions between the simple processing units. Connections between levels operate in both directions. Connections between units and nodes in the same level are inhibitory. The TRACE model shares many common arguments

with the old version of COHORT theory. For example, it assumes both bottom-up and top-down processing in speech perception. Additionally, it allows contextual information to interact with the information coming by the presented word itself to achieve a word identification. Using the TRACE model, McClelland and Elman successfully imitated some phenomena of human speech perception, for instance, categorical speech perception.

Fuzzy-Logical model: Motivated by the various new findings that greatly challenged the MT and DRT, a number of speech researchers proposed several alternative accounts of speech perception, among which the fuzzy-logic theory of speech perception developed by Massaro [37, 38] is the most influential. Unlike the MT, which claims that speech perception is a special process, the fuzzy-logic theory suggests that speech perception is based on the general ability of the perceiver to make use of multiple imperfect acoustic cues to categorize complex stimuli. It is a part of the same mechanisms of audition and perceptual learning that have evolved in humans or human ancestors to handle other classes of environmental sounds. No special speech decoder is required for the explanation of speech perception. In particular, it is assumed that speech sounds are stored in the human brain as various prototypes (sequence of perceptual units). Speech recognition involves comparison of the sensory representation of the input sound to the exemplary representations of speech sounds in the memory. To demonstrate how the fuzzy logic theory works, Massaro actually built a computer model, named Fuzzy-Logical Model of Perception (FLMP), in which speech perception is modeled as a three-stage (evaluation, integration, and decision) categorization process. The first stage analyzes the features of the input signal; the second stage integrates the features from multiple sources and compares the sensory representation to the various prototypes; the third stage chooses the best fitting prototype for the sound being heard.

Fletcher-Allen model: During the 1920s to 1940s, Fletcher and his colleagues at the Bell Labs conducted a series of speech perception experiments using nonsense syllables as the stimuli, and they discovered some statistical relationships between the articulation scores of phonemes and syllables. Specifically, the phone error rate of full-band signal is equal, on average, to the product of the error rates from the subband signals. The articulation score of a nonsense syllable is equal to the product of the scores

of the individual phonemes that constitute the syllables. Based on the psychological data, Allen proposed a hypothetical cascade model of speech perception as depicted in Fig. 1.1. As a data-driven structure, the Fletcher-Allen model takes a bottom-up form. No feedback is assumed between layers in this over-simplified model of human speech perception.

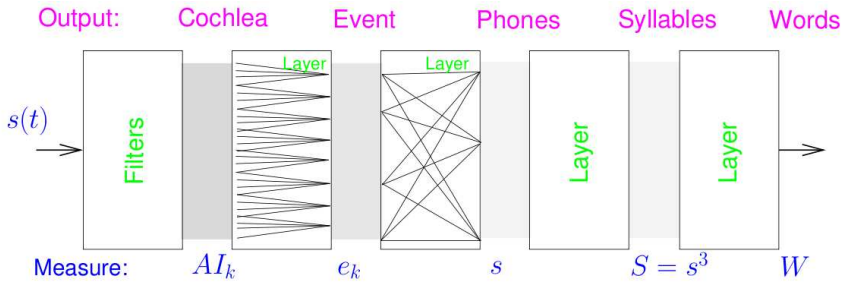


Figure 1.1: Fletcher-Allen model of speech perception. The words along the top describe the physical correlate of the measure. The first layer, the cochlea, determines the signal-to-noise ratio in about 2800 overlapping critical band channels. The next layer extracts perceptual cues (events) from the speech in a local manner. The events are integrated across the entire tonotopic axis, and then syllables and words are identified. From [71].

Distinctive Feature model: Motivated by the theory of inherent distinctive features detected from the languages of the world, K. Stevens [39] believes that distinctive features can be used as the kernel for the analysis of speech communication, including production and perception. Since the properties of speech can be uniquely and invariantly specified from the acoustic signal itself, and since these properties are closely related to the distinctive features, speech perception is a process of decoding the sound into a representation of distinctive features [16, 25]. Just before he retired, K. Stevens further extended this idea and created a model for lexical access based on acoustic landmarks and distinctive features [39]. The lexical-access model he proposed consists of three layers: acoustic cues (called landmarks)→distinctive feature→words. Accordingly, speech perception follows three steps. First, the acoustic cues that provide information about relevant articulatory states and movements are extracted from the input signal. Then the acoustic cues are integrated by the human brain to uncover the distinctive features intended by the speaker. Last, the representation of features is matched against the word, which is also specified in features.

1.2.2 Speech cues and features

Acoustic Cue: Speech sounds are encoded by some time-frequency energy patterns called acoustic cues. Given a speech sound, the human auditory system first detects the events from the speech signal, and then finds the closest match to the patterns. Failure in detecting the cues may cause serious errors in speech perception. As a matter of fact, many difficult problems of speech processing are more or less associated with the lack of information about speech cues. Finding the cues for speech perception is crucial for speech study.

The first search for acoustic cues dates back to 1940s, when Potter, Kopp, and Green at Bell Labs started a project called visible speech [40], with an aim of helping the hearing-impaired people understand speech through eyes rather than ears. For the first time the spectrograph was used for the analysis of speech sounds, with the frequency range being limited to 3.2 kHz, the upper limit of telephony [41]. Five normal hearing and one hearing-impaired listeners participated in the project. After taking a series of lectures on the spectrographs of isolated syllables and continuous speech, all the subjects were successfully trained to read speech by simply looking at the spectrographs. A shortcoming of this pioneering work is that the acoustic cues are identified from the spectrographs by visual inspection. Sophisticated analysis of speech cues requires a quantitative method in which the acoustic properties of a speech sound can be measured accurately.

In the early 1950s, Liberman, Delattre, Cooper, and other researchers at the Haskins Laboratories initiated a series of landmark studies on the acoustic cues of stop consonants using synthetic speech. *Playback*, a system that can synthesize speech from spectrographs, was created for the purpose. Based on the spectrographs of real speech, it was postulated that burst and consonant-vowel transition, which correspond to the "movement" from the locus to the steady state of the vowel, are the two critical variables for the percept of stop consonants. Two experiments were carried out to verify the hypothesis. In the first experiment [42], the effects of the burst and transition frequencies on the percept of unvoiced stop consonants were investigated by a set of CVs synthesized from 12 burst $\times 7$ F2 frequencies (refer to Fig. 1.2). Results showed that most listeners heard /t/ when the burst frequency was greater than F2 in the vowel;

when the burst frequency was close to the F2 frequency, most listeners reported /k/; the rest were identified as /p/.

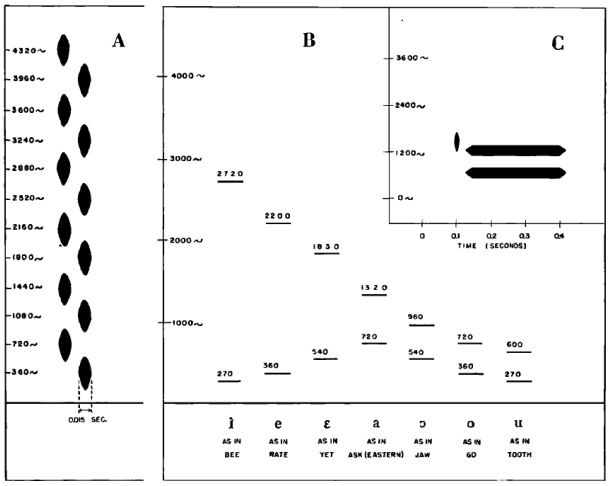


Figure 1.2: Stimulus patterns used in determining the effect of burst position and consonant-vowel transition frequency on the percept of the unvoiced stop consonants. (A) Frequency positions of the twelve bursts of noise. (B) Frequency positions of the formants of the two-formant vowels with which the bursts were paired. (C) An example of the 84 "syllables" formed by pairing a burst of noise and a two-formant vowel. From [42].

In a following experiment [43], they dropped the burst and examined the effect of transition only. The speech sounds were synthesized from two formants, a fixed F1 and an F2 of various transition types, including rising, constant, and falling, as depicted in Fig. 1.3. Results indicated that stimuli with rising transition were identified as /b/, those with F2 emanating from 1.8 kHz were associated with /d/, and those with a falling transition were reported as /g/.

The work of Liberman et al. has had a big impact on speech study. Ever since then, speech synthesis has become a standard method for feature analysis. The same technique was applied in the search for acoustic correlate for stops [44], fricatives [45,46], and nasals [47–49].

Coarticulation: In conversational speech, the position of the articulator for one sound is often assimilated with the movement of articulators for neighboring sounds. As a consequence, the speech cues to successive units of speech may overlap in time. This phenomenon is called coarticulation. In his 1952 study of acoustic cues [42], Liberman discovered that the burst frequency of the stop sound is dependent on the place of the

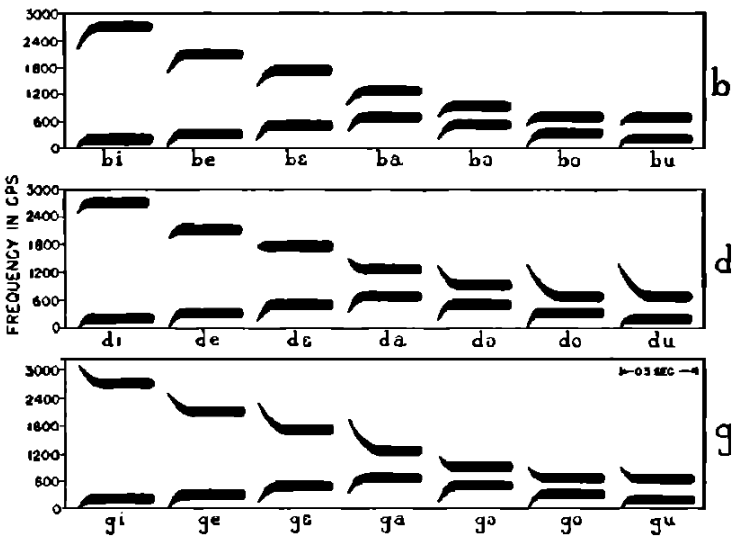


Figure 1.3: Synthetic two-formant speech stimuli used in determining the relationship between the pattern of F2 transition and the percept of place of stop articulation. From [43].

following vowel. Take the alveolar /g/ for example: the burst frequency of /ga/ is about 1kHz. In contrast, the burst frequency of /gi/ is around 4 kHz. Other studies on CVCs indicate that the formant motion into and out of the stops maintains a symmetry. Most of the time, the transitions into the stops mirror the transitions out of them [50]. The variability of the formants immediately after the burst suggests articulatory movement during the closure in anticipation of the next vowel. Because of the overlap of successive phonemes, some people believe that it is counterproductive to try to divide the speech stream up into separate phoneme units in advance of identifying the units [34, 51].

Context Effect: Speech perception is a complex process in that the integration of acoustic cues is governed by the high-level language components, such as lexical, morphological, syntactic, and semantic constraints. In a classical experiment [52], Miller investigated the effect of grammatical restriction on word intelligibility. Sentences containing five key words were presented under the conditions of various signal-to-noise ratios. To measure the effect of syntactic constraints, the key words from the sentences were shuffled in sequence and compared to the normal cases. Results indicated that the removal of coherent sentential context significantly reduced the number of words perceived correctly.

In another study [53], Warren played three sentences:

It was found that the (h)eel was on the shoe

It was found that the (p)eel was on the orange

It was found that the (m)eel was on the table

with the phoneme in the brackets being replaced by a cough-like sound. Although the listeners noticed the prominent intrusion of the sound replacing the speech, they were unable to detect that the phoneme was missing. The sentences were perceived as if they were perfectly intact. This is known as the phonemic restoration effect.

Savin and Bever [54] presented a sequence of nonsense syllables to the subjects and asked them to respond as soon as they heard the target, a whole syllable (e.g., /baeb/) or a phoneme (e.g., /ae/) in a syllable. The listeners responded more slowly to phonemes than to syllables, indicating that the bottom-up structure (phonetic segments, phonemes, syllables) is not always correct.

In a recent study, the influence of linguistic background on speech perception was investigated by Kazanina et al. [55], who compared the perceptual spaces, as reflected in early auditory brain responses, of Russian and Korean speakers by using magnetoencephalographic brain recordings. Results demonstrated that a speaker's perceptual space is shaped not only by bottom-up analysis of the distribution of the sounds in his language but also by more abstract analysis of the functional significance of those sounds.

Temporal Cue: Apart from the aforementioned acoustic cues, which act as the primary source of information, there exist other types of speech cues as revealed by the following studies. In 1981 Remez showed that traditional acoustic cues for phonetic segments such as burst and transitions are not required for speech perception [56]. A three-tone sinusoidal replica, called Sine-Wave speech, of a naturally produced utterance that simulates the time-varying properties of the three formants is sufficient to support perception of the linguistic message.

To assess the contribution of temporal cues on speech recognition, Drullman et al. [57, 58] investigated the extent to which speech intelligibility depends on the details in the temporal envelope using Vocoder speech [59]. The speech signal was decomposed into several bands with the fine structure of frequency detail being destroyed. Only the temporal envelope in each band was preserved. Results show that listeners can

only partially understand speech in quiet when the amplitude fluctuations are limited to 2 Hz; the performance improves as broader frequency bands are used. For envelope cutoff frequencies above 4 Hz, speech intelligibility is independent of the processing bandwidth. Phoneme identification with nonsense syllables shows that consonants are more affected by temporal smearing than vowels. Stops appear to suffer most, due to their short duration. Using a similar technique, Shannon et al. [60] showed that high speech recognition performance can be achieved with only three time-varying bands of noise representing the complex spectral patterns of speech.

In a recent study, Zeng investigated the relative contribution of amplitude modulation (AM) and frequency modulation (FM) on three speech perception tasks [61]: (1) speech recognition with a competing voice; (2) speaker identification; (3) Mandarin tone recognition. Comparison of the results of AM and FM processed stimuli with the original unprocessed stimuli suggests that AM and FM provide independent yet complementary contributions to support robust speech recognition under realistic listening situations.

Visual Cue: Not only audio input, but also visual input contributes to speech perception. In a classical study, McGurk et al. conducted a speech perception test with audio-visual information [62]. It was observed that most listeners heard /da/ when the sound /ba/ was presented with a synchronous video clip showing the lip movement of /ga/. Clearly the visual information also plays a role in the speech perception.

Articulatory Feature: In articulatory phonetics, a subfield of linguistics, the articulatory features are created to characterize how humans produce speech sounds, such as vowels and consonants [63]. Place and manner form the two major types of articulatory features. The place of articulation—labial, dental, palatal, velar, for instance—describes the position where the obstruction occurs in the vocal tract, while the manner of articulation—stops, fricatives, approximant, affricative, for instance—describes how the speech organs (lips, teeth, tongue) get involved in the speech production.

Distinctive Feature: In linguistics, distinctive features are the most basic units of phonological structure. The inherent distinctive features detected in the languages of the world amount to twelve binary oppositions [22]: (1) vocalic/non-vocalic, (2) consonantal/non-consonantal, (3) interrupted/continuant, (4) checked/unchecked, (5) strident/mellow, (6) voiced/unvoiced, (7) compact/diffuse, (8) grave/acute, (9) flat/plain,

(10) sharp/plain, (11) tense/lax, (12) nasal/oral. The main advantage of the distinctive features is that they are sufficient for defining any phoneme of most language. Speech sounds of the same feature should have quantitatively the same articulatory, acoustic, and perceptive correlates independent of context [41]. It must be noted that the distinctive features are formulated for the distinction of phonemes rather than the representation of speech sound.

Encouraged by the discovery of acoustic cues for stop consonants, Jakobson proposed an extensive use of distinctive features as the basis of speech analysis, such as production and perception [22]. This work has had a big impact on the research of speech perception. In the following decades, distinctive features, together with articulatory features (motivated by the Motor Theory of speech perception), became the basis for the analysis of perceptual data. Typical research questions include: Which feature carries the most information? Which feature systems best represent the perceptual space [64–68]? Extensive effort has been spent in the search for articulatory, acoustic, and perceptual correlates of the universal distinctive features [16], with little success. Most important of all, without the addition of linguistically redundant information the distinctive features are too abstract to be the basis for the quantitative operations needed for speech processing. As Fant, one of the three people who started the idea, puts it (page 18 in [41]): "The limitations of the preliminary study of Jakobson, Fant and Halle are that the formulations are made for the benefit of linguistic theory rather than for engineering or phonetic applications. Statements of the acoustic correlates to distinctive features have been condensed to an extent where they retain merely a generalized abstraction insufficient as a basis for the quantitative operations needed for practical applications. It should also be remembered that most of the features are relational in character and thus imply comparisons rather than absolute identifications."

1.2.3 Cochlear speech processing

The cochlea plays a vital role in speech perception. Once the cochlea is damaged, our ability to process speech in noise is seriously degraded. The main functionalities 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. The auditory neurons then convert the signal into neural spikes and send them to the central auditory system. This is a basic question of 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, while the ear has about 11 to 12 [69]. Neurons are low bandwidth neural channels. The stimulus has a relatively high information rate. The eye and the ear must cope with the bandwidth problem by reducing the stimulus to a large number of low bandwidth signals. It is then the job of the cortex to piece these pixel signals back together, to reconstruct the world as we see and hear it.

Most sensorineural hearing loss can be attributed to the malfunction of cochlear outer hair cells (OHCs) and inner hair cells (IHCs). Damage to OHCs reduces the vibration of the cell's cilia at the stimulus frequency, resulting in an elevated detection threshold. Damage to the IHCs reduces the efficiency of mechanical-to-electrical transduction, also resulting in an elevated detection threshold. The audiometry configuration is not a good indicator of the physiological nature of the hearing loss [6]; specifically, subjects with OHC and IHC loss may show the same amount of shifting in hearing threshold, yet the influence of the two types of hearing loss on speech perception can be very different.

It is well known that damage to "nerve cells" (i.e., OHCs) leads to a reduction of dynamic range, a disorder clinically named *loudness recruitment*. Recruitment, the most common form of neurosensory hearing loss, is best characterized as the reduction in dynamic range. Recruitment results from outer hair cell damage. To successfully design hearing aids that deal with the problem of recruitment, we need models to 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* into the time domain.

The loss of IHCs also has a serious impact on speech perception, as indicated by the results of an elderly subject (AS) with moderate hearing loss, who volunteered in our pilot study of hearing-impaired speech perception. Due to a *cochlear dead region* (an extreme case of IHC loss [6]) from 2 to 3 kHz, where the perceptual cues for /ka/ and

/ga/ are located, AS cannot hear these two sounds with her left ear. In contrast, her right ear can hear /ka/ and /ga/ (with low accuracy), despite the fact that the two ears have an almost identical hearing threshold. A consonant 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/.

1.3 Thesis Outline

The goal of this thesis is to gain more insight into hearing-impaired speech perception and understand why people with hearing loss cannot understand noisy speech, so that more advanced speech enhancing algorithms can be developed to compensate for it. Based on the analysis of a large amount of speech perception data, it is hypothesized that HI listeners may have difficulty with noisy speech because they cannot hear certain sounds, for which the characteristic features are lost due to both noise and the hearing loss. The distorted speech cues may reduce the HI listeners' binaural processing ability and make it difficult to attend to one talker in the case of a cocktail party environment. Thus the corrupted speech can be enhanced by selectively boosting the acoustic features. To explore the hypothesis, the following tasks are addressed sequentially.

Chapter 2 evaluates the validity of the multi-band product rule of frequency integration for consonant recognition, a basic assumption of Articulation Index (AI) theory and its extension, the Speech Intelligibility Index (SII). A speech perception test using high-pass and low-pass filtered nonsense syllables as the stimuli was conducted to investigate the validity of the band-independence assumption for individual consonant sounds. The cutoff frequencies were chosen such that the basilar membrane was evenly divided into 12 segments from 250 Hz to 8000 Hz with the high-pass and low-pass filters sharing the same six cutoff frequencies, in the middle frequency range.

Chapter 3 determines the perceptual cues of consonant sounds by using psychoacoustic methods. To measure the time-frequency importance function of the consonant sounds, speech stimuli (16 nonsense CVs from the LDC-2005S22 database) are high-pass or low-pass filtered and time-truncated before being presented to normal hearing (NH) listeners. Databases of speech perception under various SNR conditions are constructed to investigate the effect of noise on speech recognition. The AI-gram, a visualization

tool that simulates the auditory peripheral processing, is used for the audible part of the speech events under various SNR conditions.

Chapter 4 investigates the impact of sensorineural hearing loss on consonant identification. A perception test of noisy speech is applied to identify the difficult speech sounds for the hearing impaired listeners. Pure tone audiometry (PTA), threshold equalized noise (TEN) and psychoacoustic tuning curve (PTC) tests are used to characterize the sensorineural hearing loss. An extended speech banana that accounts for the effect of steady-state masking noise is developed to find out the correlation between the hearing loss and the intelligibility of individual consonants.

Chapter 5 explores the potential use of prior knowledge about speech cues, as identified in Chapter 3, in speech processing. It was found that natural speech sounds contain conflicting speech cues that are characteristic of confusable sounds. Through the manipulation of these acoustic cues, one phone (a consonant or vowel) can be morphed into another; a weak sound, easily masked by noise, can be converted into a strong one. The fact that the percept of nonsense syllables, words and sentences can be convincingly changed by playing with the perceptual cues indicates that the identified speech cues are indeed the basic units for speech perception. A small speech perception experiment using feature-enhanced speech sounds as the stimuli is conducted on a few normal hearing and hearing impaired listeners.

Chapter 6 summarizes the problems encountered, their solutions, and the contributions of the thesis.

CHAPTER 2

MULTIBAND PRODUCT RULE OF FEATURE INTEGRATION AND CONSONANT IDENTIFICATION

The multiband product rule, also known as band-independence, is a basic assumption of the Articulation Index (AI) and its extension, the Speech Intelligibility Index (SII). Previously Fletcher showed its validity for a balanced mix of CV (20%), VC (20%) and CVC (60%) sounds. This study repeats Miller and Nicely's version of the hi/lo-pass experiment with minor changes to study band-independence for the 16 Miller-Nicely consonants. The cutoff frequencies are chosen such that the basilar membrane is evenly divided into 12 segments from 250 Hz to 8000 Hz with the highpass and lowpass filters sharing the same six cutoff frequencies in the middle. Results show that the multiband product rule is statistically true for consonants on average. It also applies to subgroups of consonants, such as stops and fricatives, which are characterized by a flat distribution of speech cues along the frequency. It fails for individual consonants.

2.1 Introduction

A fundamental problem of human speech perception is how the human auditory system integrates speech cues across frequency. The most relevant study on this topic dates back to the 1920s, when Fletcher and his colleagues at Bell Labs were investigating speech articulation over voice communication systems [70]. Lowpass and highpass filtered nonsense syllables were used for the study of phone recognition. They found that the average phone error of the full-band stimuli e is equal to the product of the error of the lowpass filtered stimuli e_L and the error of the complimentary highpass filtered stimuli e_H, that is,

$$e = e_L \times e_H. \tag{2.1}$$

In other words, the lowpass band and the highpass band are consistent with the assumption that the low band and high band are independent. Equation (2.1) was generalized into a multiple band form [14, 70, 71]

$$e = e_1 e_2 ... e_K. \tag{2.2}$$

The number of independent articulation bands is generally taken to be $K = 20$, which makes each band correspond to about 1 mm along the basilar membrane [71].

Let s denote the average phone articulation (i.e., the probability of the nonsense phones being correctly recognized); then the articulation error $e = 1 - s$, and the articulation band error $e_1 = 1 - s_1$...etc. Given Eq. (2.2),

$$log(1 - s) = \sum_{k=1}^{K} log(1 - s_k). \tag{2.3}$$

Notice that $-log(1 - s_k)$ is similar to the definition of entropy [72], and thus may be interpreted as the information carried by the kth band [14, 71]. Equation (2.3) strongly suggests that the human speech recognition system consists of at least K parallel channels and that the total information is equal to the sum over the information in the K articulation bands. This relation may also be called the *additivity law of frequency integration*. It is the foundation of the two ANSI standards, Articulation Index (AI) [73] and more recently Speech Intelligibility Index (SII) [74].

Based on the assumption of independent articulation bands, French and Steinberg developed a method for calculation of AI based on the intensity of the long-term average speech and noise [15]. Following the verification by Beranek [75] and Kryter [76], French and Steinberg's method became an ANSI standard in 1969. Then in 1970-1980 Steeneken and Houtgast extended the AI to the Speech Transmission Index (STI) by introducing a modulation transfer function (MTF) to account for reverberation and peak clipping [77]. The original AI was developed for the use of normal hearing listeners. Later it was extended to estimate speech intelligibility for hearing-impaired listener [78–81], resulting in a new ANSI standard named the Speech Intelligibility Index (SII). All the three models—AI, STI, and SII—are based on the same Fletcher-Galt assumption that the total articulation is the sum of the contribution from multiple in-

dependent narrow bands.

Despite its importance to the widely used articulation models, the validity of the multiband product rule (Eq. (2.2)) has actually been a key open question [82]. For example, Kryter [83] showed that AI was a valid predictor of the intelligibility of speech under a wide variety of conditions of noise masking and speech distortion except for the cases of three non-contiguous pass bands at 0–600 Hz, 1200–2400 Hz, and 4800–9600 Hz. Grant and Braida [84] found that the predicted AI based on the sum of the AIs from individual bands was greater than the observed AI by approximately 18% for adjacent 1/3-oct bands, while the AI predicted for combinations of non-adjacent bands was less than the observed AI by approximately 41%. Lippmann [85] also found that the stop-band data did not agree with AI calculation. In 2001 Müsch and Buus coined two new terms—*synergistic* and *redundant* interaction between neighboring bands to explain why the AI under- or overestimate the wideband error, compared to the product of the errors associated with the narrow bands [86, 87]. It has been conjectured that a revised model that accounts for the mutual dependency between adjacent bands might give a better prediction [88]. In a recent study, Ronan et al. [89] compared several frequency integration models for the prediction of individual consonant articulation score, for narrow-band cases. Results indicated that Fletcher's product rule (Eq. (2.2)) made satisfactory predictions under various combinations of adjacent and non-adjacent narrow-band speech, except for the case of multiple high-frequency narrow bands, for which none of the evaluated methods are satisfactory. Investigation of SII [90] also found that it greatly over-predicted performance at high sensation levels, and under-predicted performance at low sensation levels for many hearing-impaired listeners. The information contained in each frequency band is not strictly additive.

In 1955, Miller and Nicely (MN55) repeated Fletcher and Galt's highpass and lowpass filtering experiment [64] for the analysis of perceptual confusion. The speech stimuli include 16 consonant sounds, /p, t, k, f, θ, s, ʃ, b, d, g, v, ð, z, ʒ, m, n/, spoken initially before the vowel /a/. Using the data from experiment MN55, we checked the validity of Fletcher's product rule (Eq. (2.1)). Results show that the model applies to the consonants on average, despite that it over-predicts the full-band error by 10%. We then plotted the product of e_L and e_H against the full-band error e for each of the 16

consonant sounds. To our surprise, more than half of the consonant sounds, specifically, /p, k, f, ʃ, b, d, g, ʒ, m, n/, show only small discrepancy.

Designed for the purpose of confusion analysis, the MN55 data is unsuitable for the study of the multiband product rule, for several reasons. First, the frequency samples are limited. Only six lowpass and five highpass conditions are included; in contrast, Fletcher and French and Steinberg suggested $K = 20$ frequency points. Second, the cutoff frequencies are not evenly distributed along the effective range of speech communication. Four out of six lowpass samples are below 1.5 kHz, with only one highpass sample within the same frequency range. Interpolation between data points introduces significant error.

In the present study we investigate the validity of the multiband product rule for consonant sounds. The product rule is evaluated on three levels: (1) 16 consonants on average; (2) subgroups such as stops and fricatives; (3) individual consonants. A computer-based highpass and lowpass experiment, named HL07, is designed for this purpose. The new experiment utilizes the same 16 consonant sounds as experiment MN55. To address the problems listed above, the cut-off frequencies were chosen such that the basilar membrane is evenly divided into 12 bands over the frequency range of [250 Hz, 8000 Hz], with the lowpass and highpass filters sharing the same six cut-off frequencies in the mid-frequency range.

2.2 Methods

2.2.1 Subjects

Nineteen normal hearing subjects were enrolled in the experiment, of which 6 male and 12 female listeners completed. Except for one subject in her 40s, all the subjects were college students in their 20s. The subjects were born in the U.S. with English being their first language. All subjects were paid for their participation. Approval by the University of Illinois Institutional Review Board was obtained for the experiment. In order to make sure that all the data are of high quality, the performances of the listeners were assessed by their average recognition score. Those who had abnormally low scores will be excluded for further analysis. In experiment HL07, no subject has been removed

for that reason.

2.2.2 Speech stimuli

The same 16 nonsense CVs used by Miller Nicely (1955) were chosen. A subset of wide-band syllables sampled at 16,000 Hz were taken from the LDC-2005S22 corpus. Each CV was spoken by 20 talkers, among which only 6 utterances, half male and half female, were finally chosen for the test, to reduce the total duration of the experiment. The 6 utterances were selected such that they were representative of the speech material in terms of confusion patterns and articulation score based on the results of a similar speech perception experiment [91]. The speech sounds were presented to both ears of the subjects at the listener's most comfortable level (MCL), but always less than 80 dB SPL.

2.2.3 Conditions

The subjects were tested under 19 filtering conditions, including one full-band (250–8000 Hz), nine highpass and nine lowpass conditions. The cut-off frequencies were calculated from Greenwood inverse cochlear map function [92] such that the full-band frequency range from 0.25 kHz to 8 kHz was divided into 12 bands, corresponding to equal length along the basilar membrane. Figure 2.1 illustrates the frequency samples and the corresponding distances from the base on the human basilar membrane. The cut-off frequencies of the highpass filtering were 6185, 4775, 3678, 2826, 2164, 1649, 1250, 939, and 697 Hz, with the upper limit at 8000 Hz. The cut-off frequencies of the lowpass filter were 3678, 2826, 2164, 1649, 1250, 939, 697, 509, and 363 Hz, with a lower limit at 250 Hz. The highpass and lowpass filtering shared the same cut-off frequencies over the middle frequency range that contains most of the speech information. The filters were 6^{th} order elliptical filters with 0.02 dB of peak-to-peak ripple and a stop-band attenuation of -60 dB. To make the filtered speech sound more natural and to mask the stop bands, white noise was used to mask the stimuli at the signal-to-noise ratio of 12 dB, based on the average speech spectra of the 96 nonsense syllables.

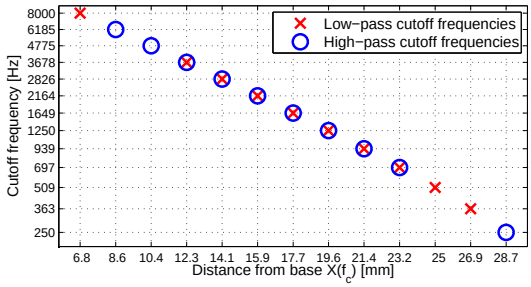


Figure 2.1: Highpass and lowpass cutoff frequencies of experiment HL07.

2.2.4 Procedure

The speech perception experiment was conducted in a sound-proof booth. A Matlab code was developed for the collection of the data. Speech stimuli were presented to the listeners through Sennheisser HD 280-pro headphones. Subjects responded by clicking on the button labeled with the CV that they 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 subjects were instructed to click on a "Noise Only" button. A total of 2208 tokens were randomized and divided into 16 sessions, each of which lasted for about 15 min. A mandatory practice session of 60 tokens was given at the beginning of the experiment. To prevent fatigue, the subjects were instructed to take frequent breaks. The subjects were allowed to play each token up to 3 times. At the end of each session, the subject's test score, together with the average score of all listeners, was shown to the listener to provide feedback on their relative progress, as motivation.

2.2.5 Difference between HL07 and MN55

Although experiment HL07 can be regarded as a repeat of the MN55 study, the two experiments are distinguished in several important aspects. First, the subjects differ in gender and proficiency. In MN55, five extensively trained female subjects served as both talkers and listening crew. This introduced a "coupling" effect between the talkers and the listeners, as well as an awareness of the relative difficulty of the sounds. In HL07 we use recorded speech prepared by 10 male and 8 female talkers from the LDC database. All the 18 subjects (6 male and 12 female) are naive listeners without any experience in speech perception tests. Second, the noise levels are different. Both

experiments use white noise at 12 dB SNR. However, in experiment MN55 the speech level was controlled by a VU meter [93] which measures the speech peaks, while in experiment HL07 the noisy speech was created by setting the RMS level of the speech and noise. Thus 12 dB SNR in MN55 is about the same as 14 dB SNR in HL07 [93]. As a consequence, the fullband error of MN55 is about 12% lower than that of HL07. Third, the filtering conditions are different. In MN55 the fullband speech was created by a wide-band filter of 0.2–6.5 kHz, and then the distorted speech was created by filtering the fullband speech with a lowpass cutoff frequency of 0.3, 0.4, 0.6, 1.2, 2.5, 5 kHz and a highpass cutoff frequency of 0.2, 1.0, 2.0, 2.5, 3.0, 4.5 kHz. In contrast, the fullband speech in HL07 goes to 8 kHz. The loss of information from 6.5 kHz to 8 kHz accounts well for the over-prediction of MN55 in the high frequency. Fourth, the test platforms are different. Data collection in MN55 was paper-based. The listeners were told to choose a response from the 16 nonsense CVs and write it down on the answer sheet within seconds following the presentation. The HL07 experiment is computer-based. No limit is applied for the responding time. Subjects were allowed to play each sound up to three times. In case the subjects could not tell which sound is presented, a "Noise Only" button was added.

2.2.6 Data analysis

The validity of the Fletcher's product rule (Eq. (2.1)) is investigated for average speech and individual consonants. The probability of error of a token (an utterance filtered at a frequency) is defined as the number of mislabeled responses divided by the total number of presentations. The mean error of a consonant is the average over the six tokens pronounced by different talkers. Similarly, the total error of average speech can be calculated by averaging the errors of the 16 consonants. For both average speech and individual consonants, the fitness of the model to the data is evaluated in terms of average bias $B(f_c)$ and $\chi^2(f_c)$ computed from the error of all listeners. The average bias is given by

$$B(f_c) = e - e_L \times e_H \tag{2.4}$$

where $e_L \times e_H$ and e are the model error and observed error at a cutoff frequency f_c. The chi-square statistic is

$$\chi^2(f_c) = N\frac{[(1 - e_L \times e_H) - (1 - e)]^2}{1 - e_L \times e_H} + N\frac{[e - e_L \times e_H]^2}{e_L \times e_H} \tag{2.5}$$

where N is the total number of presentations for the particular condition. The quantities $(1 - e_L \times e_H)$ and $(1 - e)$ are the predicted and observed scores. A significance level (the probability of this result not being due to chance) of 0.05 is chosen as the threshold of the chi-square test. A value of χ^2 greater than the threshold indicates that the measurements do not satisfy Eq. (2.1) at that condition, whereas when χ^2 is less than the threshold of significance, the Fletcher's product rule can be regarded as true.

The above analysis is carried out by treating the 18 listeners as an average normal listener. In order to determine if the same conclusion applies to any individual listener, a one-way ANOVA test is applied to the $e - e_L \times e_H$ of different listeners following each χ^2 test. Due to the small number of responses, the 16 sessions are combined into 4 repeats, 4 sessions each. Let B_i denote the bias of $e_L \times e_H$ against e for subject i, and B_{ij} denote the bias of repeat j from subject i. Assuming that B_i has a Gaussian distribution $N(b_i, \sigma)$, where b_i is the mean of B_i, we can compare the mean of the various listeners by testing the hypothesis that they all have the same bias, against the general alternative that they are not all the same. If no two listeners are significantly different, we may conclude that the conclusion based on the average normal listener is applicable to any individual listener.

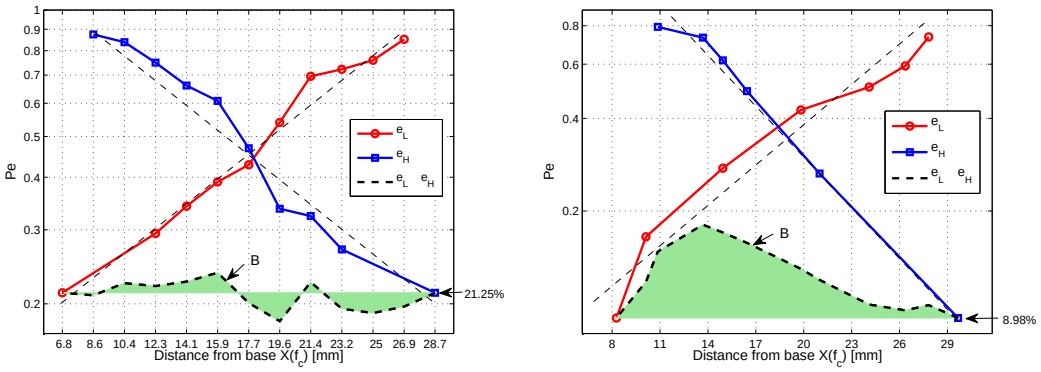
2.3 Results

2.3.1 Multiband product rule for 16 consonants on average

Results indicate that the multiband product rule closely fits the recognition scores averaged over the 16 consonants. Figure 2.2(b) depicts the lowpass error e_L, the highpass error e_H, and their product as a function of cutoff frequency. The fullband error e is equal to the lowpass error e_L at 8000 Hz and the highpass error e_H at 250 Hz. The intermediate points of lowpass and highpass error are linearly interpolated from the

nearest neighboring points. The average bias $B = e - e_L \times e_H$ is depicted by the shaded area. Supposing that the product rule is true, the shaded area would be zero. It is shown in Fig. 2.2(a) that the difference between $e_L \times e_H$ and the fullband error e is typically less than 3%, which is very close to zero.

Figure 2.2(b) depicts the results of experiment MN55 [64]. Fletcher's product rule over-predicts the fullband error over most frequencies for MN55, but still the measurements fit the model with reasonable accuracy. Since the lowpass and the highpass conditions do not use the same set of cutoff frequencies, the lowpass error e_L and highpass error e_H are linearly interpolated along the frequency to create the $e_L \times e_H$ curve, which introduces extra error in the prediction.



(a) HL07 (b) MN55

Figure 2.2: Grand probability of error and the average bias $B = e - e_L \times e_H$ for 16 consonants as a function of cutoff frequency. Figure (a) shows the average lowpass error e_L (circles), the average highpass error e_H (squares), and the product of the two $e_L \times e_H$ (thick dashed) for experiment HL07. The fullband error e is defined as e_L (f_c=8000 Hz) or e_H (f_c=250 Hz). The average bias B is depicted by the shaded area. Figure (b) shows the same data from experiment MN55, in which the fullband error e is defined as e_L (f_c=6500 Hz) or e_H (f_c=200 Hz). Note the log ordinate scale, which makes the figures easily read, actually magnifies the bias visually.

For both experiments, the intersection points of the lowpass and highpass curves that divide the full band into two parts of equal information are about the same (1.5 kHz or 18 mm). The log lowpass error e_L and highpass error e_H have been fitted by two straight lines that are symmetrical at the intersection point. This means the speech

information is evenly distributed across frequency. A significant difference between the results of MN55 and HL07 lies in that the former has a maximum average bias B of 8.02%, which is considerably smaller than that of HL07 (21.25%). This might be due to the aforementioned coupling effect between the talkers and the listeners in experiment MN55, which makes the task relatively easy. Apart from that, the results of the two experiments are generally consistent. Due to the experimental design, experiment HL07 has better precision (smaller bias) than experiment MN55, as we seen in Fig. 2.2. Therefore, in the following sections, we will focus on analyzing the perceptual data of our experiment HL07.

Table 2.1: The average bias of 16 consonants on average in experiment HL07 for various cutoff frequencies.

Frequency (Hz)	363	509	697	939	1250	1649	2164	2826	3678	4775	6185
$B = e - e_L \times e_H$	-1.9	-2.6	-1.8	1.3	-3.1	-1.2	2.5	1.3	0.8	1.7	0.3

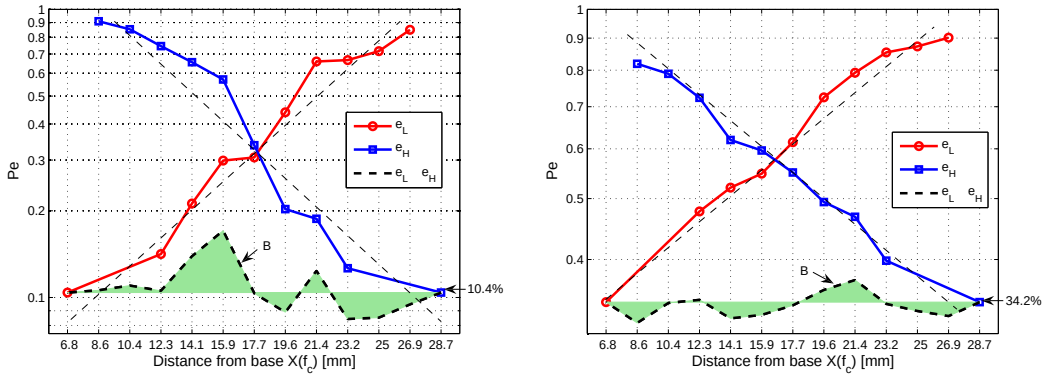
Table 2.1 lists the average bias of the predicted score (the same data is depicted in Fig. 2.2(a) as the shaded area). The results of the χ^2 tests indicate that e_L, e_H and e are consistent with the Fletcher's product rule at all frequencies. An ANOVA test indicates that the differences between the 18 listeners are too small to be statistically significant at the level of 0.05. The discrepancy between the biases of any individual listeners and the overall average bias is generally less than 5%. Therefore the 18 listeners of normal hearing can be regarded as having the same bias $e - e_L \times e_H$ independent of cutoff frequencies. Thus Fletcher's product rule may be applied to any individual normal hearing listener.

2.3.2 Multiband product rule for stops and fricatives

Analysis of the perceptual data indicates that the multiband product rule applies to the stops and fricatives as well. Figure 2.3(a) depicts the average lowpass error e_L, average highpass error e_H and the product of the two $e_L \times e_H$ for the six stop consonants (/pa, ka, ta, ba, ga, da/). The average bias $B = e - e_L \times e_H$, as depicted by the shaded area, is rather small. The highpass error and the lowpass error cross each other at about 1.5 kHz, which is about the same position (18 mm) as the weight of the 16 consonants on average. The logarithms of e_L and e_H are well approximated by straight lines having

complementary but identical slopes.

The results for the eight fricative consonants (/fa, θa, sa, ʃa, va, ða, za, ʒa/) are depicted in Fig. 2.3(b). The average bias B is almost flat with the maximum prediction error being less than 3%. Like the case of average consonants, $e_L(f_c)$ and $e_H(f_c)$ have near constant equal slopes of opposite sign when the two curves are plotted on log scales, suggesting that the fricative information is evenly distributed across the frequency range.



(a) stops (b) fricatives

Figure 2.3: Average probability of error and the average bias $B = e - e_L \times e_H$ for stops (/pa, ka, ta, ba, ga, da/) and fricatives (/fa, θa, sa, ʃa, va, ða, za, ʒa/) as a function of cutoff frequency. Figure (a) shows the average lowpass error e_L (circles), the grand highpass error e_H (squares), and the product of the two $e_L \times e_H$ (thick dashed) for stops. The average bias $B = e - e_L \times e_H$ is the shaded area. Figure (b) shows the same results for the fricatives.

Table 2.2 lists the average bias B for the two sound groups at various cutoff frequencies. All values satisfy the χ^2 test at a significance level of 0.05. An ANOVA test shows no significant difference between the results of the 18 listeners.

Table 2.2: The average bias of stops and fricatives in experiment HL07 for various cutoff frequencies.

	Frequency (Hz)										
subgroup	363	509	697	939	1250	1649	2164	2826	3678	4775	6185
stops	-1.1	-2.0	-2.0	2.0	-1.5	-0.1	6.6	3.5	0.1	0.9	0.5
fricatives	-2.1	-1.5	-0.2	2.9	1.5	-0.4	-1.6	-2.0	0.3	0.8	-1.5

2.3.3 Multiband product rule for individual consonants

Analysis of our HL07 data reveals that Fletcher's product rule applies to the 16 consonants over limited frequencies for about 80% of the cases (CVs×Frequencies). Figure 2.4 depicts the lowpass error e_L, highpass error e_H and the product of the two $e_L \times e_H$ for the 16 consonants. Based on the shape of $e_L \times e_H$, the 16 consonants can be roughly classified into flat and non-flat groups. The flat group includes /pa, ka/ and /ʃa, da, ma, na, za, ga, sa, fa, va/, for which the prediction error $e_L \times e_H - e$ is less than 5% over all frequencies, or less than 5% for most of the cutoff frequencies. The rest of the consonant sounds, /ta, ba, ʒa, θa, ða/, form the biased (non-flat) group.

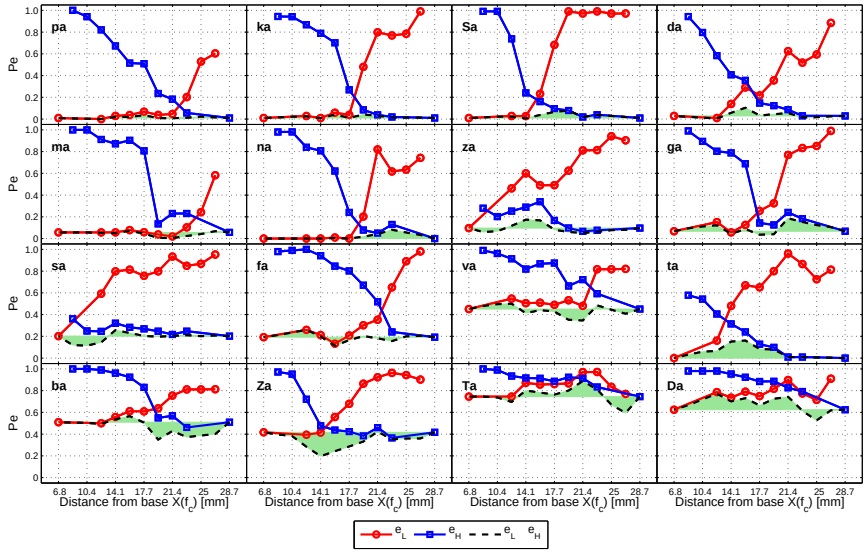


Figure 2.4: Probability of error for 16 consonants as a function of cutoff frequency. The lowpass error $e_L(f_c)$ and the highpass error $e_H(f_c)$ are marked by circles and squares respectively. The dashed curve depicts the product of the two $e_L \times e_H$. The fullband error e is equal to e_L ($f_c = 8000$ Hz) or e_H ($f_c = 250$ Hz). The bias $B(f_c) = e - e_L \times e_H$ is illustrated by the shaded area. The IPA symbols for Ta, Sa, Da, Za are /θa, ʃa, ða, ʒa/, respectively.

Table 2.3 lists the average bias of the predicted score (the same data is depicted in Fig. 2.4 as the shaded area). A χ^2 test of significance level 0.05 was applied to each of the 16 consonants. A total of 136 out of 176 cases (16 CVs×11 Frequencies) statistically satisfy Fletcher's product rule at a significance level of 0.05. Only two consonants /pa,

ka/ passed the χ^2 test over all frequencies. Most of the unsatisfied cases come from the biased group, such as /ta, ba, ða, ʒa/, for which the fail rate is 50%.

Table 2.3: The average biases of 16 consonant sounds in experiment HL07 for various cutoff frequencies. Cases for which the χ^2 test was statistically significant at the 0.05 level are marked with an asterisk.

	Frequency (Hz)										
CV	363	509	697	939	1250	1649	2164	2826	3678	4775	6185
pa	0.3	1.0	0.2	-0.1	-0.1	2.5	0.9	1.0	-1.0	-0.7	-0.4
ka	0.2	0.2	0.5	2.1	3.1	0.1	3.1	-0.2	1.5	1.2	0.7
ʃa	0.7	1.7	3.0	0.9	6.8*	5.6*	2.8	-0.3	1.2	1.4	0.8
da	-0.3	-1.1	-1.3	2.5	1.5	0.3	7.5*	2.8	-2.3	-1.7	-0.9
ma	0.2	-1.8	-3.3*	-5.2	-5.1	-1.0	1.4	-0.7	-0.5	0.0	0.0
na	2.4	4.8*	8.0*	4.0*	1.6	0.0	0.6	0.0	0.0	0.0	0.0
za	-1.4	-1.7	-3.5	-4.3*	-3.5	-1.4	7.0*	7.7*	2.0	-2.2	-2.6
ga	2.8	4.7	8.6*	11.8*	-2.6	-3.0	1.9	-2.2	5.5	4.9	3.4
sa	0.1	-0.4	0.8	0.1	-0.4	0.2	2.8	5.4	-5.7	-7.9*	-7.0*
fa	0.8	0.4	-3.6	-1.0	1.0	-2.6	-7.9*	0.7	6.7	4.8	2.4
va	-5.1	-1.5	3.3	-10.4*	-9.8*	-2.3	-0.9	-3.7	4.9	5.2	3.6
ta	0.2	0.4	0.8	0.9	7.8*	8.3*	16.1*	15.1*	6.5*	6.5*	3.9*
ba	-10.5*	-11.9*	-13.6*	-8.1	-16.0*	-0.4	5.5	2.6	-1.4	-0.7	-0.4
ʒa	-5.3	-5.2	-6.5	0.7	-8.4	-12.9*	-17.3*	-22.0*	-13.2*	-3.6	-2.1
θa	-15.5*	-8.0	6.5	14.2*	5.5	1.8	3.7	5.5	-4.9	-0.7	0.0
ða	-1.7	-10.8*	-1.1	11.8*	9.9*	3.8	10.5*	7.6	14.7*	10.6*	5.5

An ANOVA test was used to investigate the listener's dependence. Since the number of tokens per CV×Frequency for each listener is only six, a number too small for a useful statistical test, the 18 listeners are ranked according to their speech recognition scores and artificially divided into three groups. The top six are attributed to the H group. The middle six are attributed to the M group. The lower six are classified as the L group. For 173 out of 176 combinations (16 CV×11 Frequency) ANOVA tests produce the same result that the H, M, and L groups are not significantly different in terms of the average bias per CV×Frequency. In other words, the three groups of listeners are close to each other in terms of the fitness to the multiband product rule.

The perceptual data provide important information on the perceptual cues for the initial consonants. Usually the primary cue of a consonant is located around the intersection point of e_L and e_H, which divides the full band into two parts having equal information (e.g., score). When the primary speech cue is removed, the error climbs dramatically [94].

2.4 General Discussion

In Section 2.3.1, we demonstrated that Fletcher's product rule (Eq. (2.1)) is true for the average consonants at all cutoff frequencies. This can be regarded as a significant verification of the multiband product rule of frequency integration (Eq. (2.2)). Suppose that Eq. (2.2) is a consequence of the fact that the frequency bands b_k, associated with e_k, are independent in terms of speech perception. A strict proof would require a speech perception test that actually measures the 20 narrow-band recognition scores. This is totally impractical for $K = 20$, as it would require $20! = 2.5 \times 10^{18}$ tests.

If we look at the real perceptual data (Fig. 2.2(a)), it actually provides much more information. The logarithms of both e_L and e_H can be closely fitted by two lines symmetrical across the intersection point of the two curves. This clearly indicates that: (1) The speech information is evenly distributed across the frequency, as independently measured by both lowpass and highpass tests. (2) The articulation bands are additive in log error in speech perception. Similar results are observed for the two groups of stops and fricatives (Fig. 2.3(a) and (b))

Based on the observation, it is conjectured that the multiband product rule is a combined property of the peripheral auditory system that has multiple independent parallel channels, and that the input speech stimuli are characterized by a flat distribution of speech cues along the basilar membrane. It does not apply to individual consonants because the distribution of speech cues is not flat. Due to the a priori dependence between the speech cues, sometimes the highpass and lowpass errors do not fit the model. For example, when the primary cue of a sound covers more than one band, the product of the lowpass and highpass error $e_L \times e_H$ may be lower or higher than fullband error e, due to the fact that the bands neighboring the cutoff frequency are not really independent. To fully understand the interactions between the speech cues and to explain why the multiband product rule fails at certain points require knowledge of the speech features.

2.5 Conclusion

The multiband product rule of frequency integration is an empirical formula justified by the two properties about speech and hearing, specifically, (1) the speech information

is evenly distributed across the frequency, and (2) the auditory critical bands are independent in terms of speech perception. Results of our experiment HL07 show that the multiband product rule is statistically true for consonants on average. It may also apply to subgroups of consonant sounds, such as stops and fricatives, that are characterized by a flat distribution of speech cues along the frequency. It fails for individual consonants, as expected.

CHAPTER 3

PERCEPTUAL CUES OF CONSONANT SOUNDS IN NATURAL SPEECH

Synthetic speech has been widely used in the study of speech cues. A disadvantage of this method is that it requires prior knowledge about the cues to be identified. Incomplete or inaccurate hypotheses about the cues often lead to speech sounds of low quality. In this research, a 3D Deep Search (3DDS) is developed to explore the perceptual cues of stop consonants from naturally produced speech. For a given sound, it measures the contribution of each sub-component to perception by time truncating, highpass/lowpass filtering, or masking the speech with white noise. AI-gram, a visualization tool that simulates the auditory peripheral processing, is used to predict the audible components of the speech sound. Results show that the stop consonants are defined by a short duration burst characterized by its center frequency and the delay to the onset of voicing. Further analysis reveals that the robustness of a consonant sound is determined by the strength of its dominant cue.

3.1 Introduction

Speech sounds are characterized by time-varying spectral patterns called acoustic cues. When a speech wave propagates on the basilar membrane (BM), it creates perceptual cues, named *events*, which define the basic units for speech perception. The relationship between the acoustic cues and perceptual units has been a key research problem for speech perception [14, 95, 96].

Bell Labs (1940): The first search for acoustic cues dates back to the 1940s at Bell Labs, when [40] began their *visible speech* project, with the goal of training the hearing-impaired to read spectrograms. Five normal hearing (NH) and one hearing-impaired (HI) listeners participated in the study. Following a series of lectures on the spectrograph and its use on isolated syllables and continuous speech, the subjects were

successfully trained to "read" speech spectrographs. Even though the acoustic cues identified by visual inspection were not very accurate, this pioneering work laid a solid foundation for subsequent quantitative analysis.

Haskins Laboratories (1950): In the 1950s researchers at the Haskins Laboratories conducted a series of landmark studies on the acoustic cues of consonant sounds. A speech synthesis system called the *Pattern Playback* was created to convert a spectrograph into (low quality) speech sound. Based on the spectrographs of real speech, it was postulated that stop consonants are characterized by an initial burst, followed by a consonant-vowel transition. In [42], the authors investigated the effect of center frequencies of the burst and the second formant (F_2) transition on the percept of unvoiced stop consonants by using a set of "nonsense" synthetic consonant-vowel (CV) speech sounds synthesized from 12 bursts followed by seven F_2 formant frequencies. The subjects were instructed to identify the stimulus as /p/, /t/ or /k/ (a closed-set task). Results show that most people hear /t/ when the burst-frequency is higher than the F_2 frequency; when the two frequencies are close, most listeners report /k/; otherwise they hear /p/. In a following study [43], the authors dropped the burst and examined the effect of F_2 transition only on the percept of stop consonants. It was found that stimuli with rising F_2 transition were identified as /b/; those with F_2 emanating from 1.8 kHz were associated with /d/; and those with a falling transition were reported as /g/.

Follow-up studies (1960-90): The study of Liberman et al. has had a major impact on the research of speech perception. Since their study, speech synthesis has become a standard method for feature analysis. It was used in the search for acoustic correlates for stops [44], fricatives [45, 46], nasals [47–49], as well as distinctive and articulatory features [16, 25, 97]. A similar approach was taken by [56] to generate highly unintelligible "sine-wave" speech; the study concluded that the traditional cues, such as bursts and transitions, are not required for speech perception. More recently, Alwan applied the same method in modeling speech perception in noise [98].

The argument in favor of this method is that the features can be carefully controlled. However, the major disadvantage of synthetic speech is that it requires prior knowledge of the cues being sought. Thus, incomplete and inaccurate knowledge about the acoustic cues has often led to synthetic speech of low quality, and such speech sounds are

commonly unnatural and barely intelligible, which by itself is strong evidence that the critical cues for the perception of target speech sound are poorly represented. For those cases, a fair question is: *How close are the synthetic speech cues to those of natural speech?* Another key issue is the *variability* of natural speech, which depends on the talker [99], accent, masking noise, etc., most of which are well beyond the reach of the state-of-the-art speech synthesis technology. To answer questions such as why /ba/s from some of the talkers are confused with /va/, while others are confused with /ga/, or what makes one speech sound more robust to noise than another, it is critical to study the acoustic cues of naturally produced speech, rather than those of artificially synthesized speech.

This chapter describes a psychoacoustic method for isolating speech cues from natural consonant-vowel (CV) speech. Rather than making assumptions about the cues to be identified, natural speech is modified by (1) adding noise of variable degrees, (2) truncation of the speech from the onset, and (3) high and lowpass filtering the speech with variable cutoff frequencies. For each modification of the speech, the identification of the sound is judged by a large panel of listeners. We then analyze the results to determine where in time and frequency, and at what signal-to-noise ratio (SNR), the speech identity has been masked. In this way we triangulate on the location of the speech cues and the events, along the three dimensions.

3.2 Event Identification

The cochlea is a nonlinear spectrum analyzer. Once a speech sound reaches the cochlea, it is represented by time-varying energy patterns across the *basilar membrane* (BM). A small subset of the patterns contribute to speech recognition. The purpose of event identification is to isolate this small specific feature subset.

3.2.1 Modeling speech reception

The cochlea decomposes each sound through an array of overlapping nonlinear (compressive), narrow-band critical filters, splayed out along the BM, with the base and the apex of BM being tuned to the high frequency (20 kHz) and low frequency (20 Hz),

respectively [69]. Once a speech sound reaches the inner ear, it is represented by a time-varying response pattern along the BM, of which some of the sub-components contribute to speech recognition, while others do not. Many components are masked by the highly nonlinear forward-spread [18, 19] and upward-spread of masking [69]. The purpose of event identification is to isolate the specific parts of the psychoacoustic representation that are required for each consonant's identification [94].

To better understand how speech sounds are represented on the BM, the *AI-gram* is used. This construction is a *what-you-see-is-what-you-hear* (WYSIWYH, IPA: /wɑsi-wɑh/) signal processing auditory model tool, to visualize audible speech components [94, 100]. The *AI-gram* is thus called due to its estimation of the speech audibility via Fletcher's Articulation Index (AI) model of speech perception [71, 95]. The AI-gram tool, first published by [94], crudely simulates audibility using an auditory peripheral processing (a linear Fletcher-like critical band filter-bank).

The AI Model

Fletcher's AI model is an objective appraisal criterion of speech audibility. The basic concept of AI is that any narrow band of speech frequencies carries a contribution to the total index, which is independent of the other bands with which it is associated and that the total contribution of all bands is the sum of the contribution of the separate bands.

Based on the work of speech articulation over communication systems [14,70], French and Steinberg developed a method for the calculation of AI [15].

$$AI(SNR) = \frac{1}{K}\sum_{k=1}^{K} AI_k \tag{3.1}$$

where AI_k is the *specific AI* for the kth articulation band [76, 96], and

$$AI_k = \min(\frac{1}{3} log_{10}(1 + c^2 snr_k^2), 1) \tag{3.2}$$

where snr_k is the speech-to-noise root-mean-squared (RMS) ratio in the k^{th} frequency band and $c \approx 2$ is the critical band *speech-peak* to *noise-rms* ratio [15].

Given $AI(SNR)$ for the noisy speech, the predicted average speech error is [71, 96]

$$\hat{e}(AI) = e_{\min}^{AI} \cdot e_{chance} \tag{3.3}$$

where $e_{\min}$ is the maximum full-band error when $AI = 1$, and e_{chance} is the probability of error due to uniform guessing [96].

The AI-gram

The AI-gram is the integration of the Fletcher's AI model and a simple linear auditory model filter-bank (i.e., Fletcher's SNR model of detection [95]). Figure 3.1 depicts the block diagram of the AI-gram. Once the speech sound reaches the cochlea, it is decomposed into multiple auditory filter bands, followed by an "envelope" detector. Fletcher-audibility of the narrow-band speech is predicted by the formula of specific AI (Eq. (3.2)). A time-frequency pixel of the AI-gram (a two-dimensional image) is denoted $AI(t, f)$, where t and f are the time and frequency respectively. The implementation used here quantizes time to 2.5 ms, and uses 200 frequency channels, uniformly distributed in place according to the Greenwood frequency-place map of the cochlea, with bandwidths according to the critical bandwidth of [70].

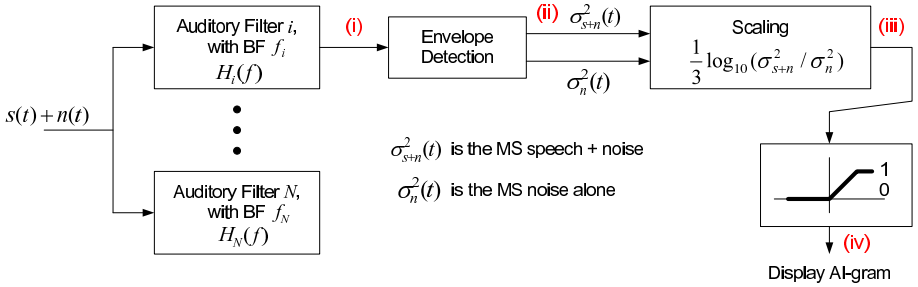


Figure 3.1: Block diagram of AI-gram (modified from [100], with permission).

The average of the AI-gram over time and frequency, and then averaged over a phonetically balanced corpus, yields a quantity numerically close to the AI as described by [96]. An average across frequency at the output of the AI-gram yields the *instantaneous AI*

$$a(t_n) \equiv \sum_k AI(t_n, f_k) \tag{3.4}$$

at time t_n.

Given a speech sound, the AI-gram model provides an approximate "visual detection threshold" of the audible speech components available to the central auditory system. It is silent on which components are relevant to the speech event. To determine the relevant cues, it is necessary to directly relate the results of speech perception experiments (events) with the AI-grams (or perhaps some future nonlinear extensions of the AI-gram).

3.2.2 3D Deep Search (3DDS)

Speech sounds are characterized by three properties: time, frequency and intensity. Event identification involves isolating the speech cues along the three dimensions. In the past studies, confusion test on nonsense syllables has long been used for the exploration of speech features. For example, Fletcher and his colleagues investigated the contribution of different frequency bands to speech intelligibility using highpass and lowpass filtered CV syllables [14, 15], resulting in the *Articulation Index* (AI) model. The study in [101] examined the relationship between dynamic features and the identification of Japanese syllables modified by initial and final truncation. More often, noise masking was used to study consonant [64, 67] and vowel [91] recognition. The study in [94] successfully combined the results of time truncation and noise masking experiments, for the identification of /ta/ event. However, it has remained unclear how many speech cues could be extracted from real speech by these methods. In fact there is high skepticism within the speech research community as to the general utility of such methods.

In the present investigation, we have integrated the three types of tests; thus, we have developed a "3DDS" for exploring the events of consonants from natural speech. To evaluate the acoustic cues along the three dimensions, speech sounds are truncated in time, high/lowpass filtered, or masked with white noise, as illustrated by Fig. 3.2, and then presented to normal hearing (NH) listeners.

Imagine that an acoustic cue, critical for speech perception, has been removed or masked. Would this degrade the speech sound and reduce the recognition score significantly? For the sound /t/, [94] has answered this question: The /t/ event is entirely due to a short $\approx$ 20 ms burst of energy, between 4 and 8 kHz. To estimate the im-

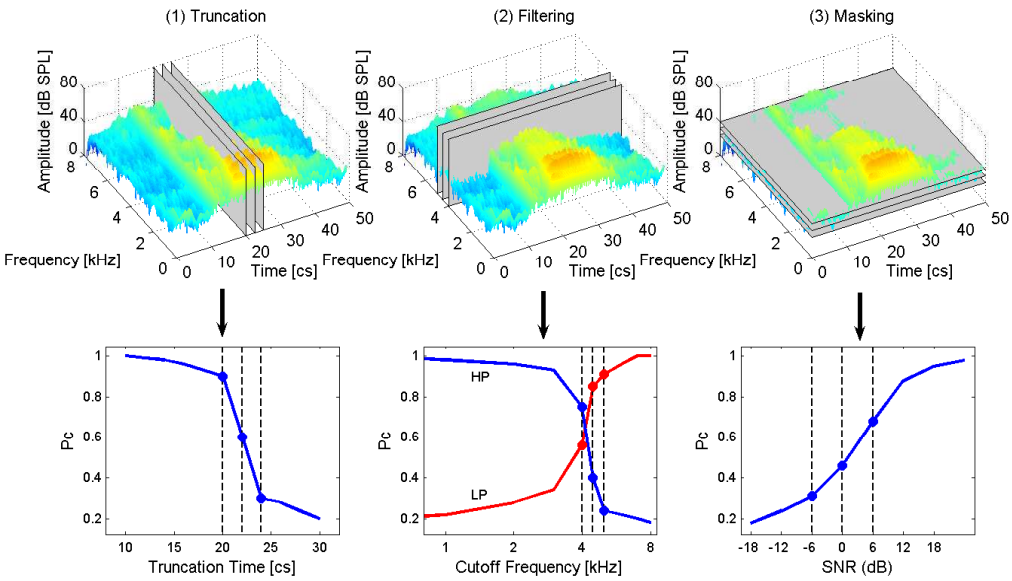


Figure 3.2: The 3DDS for the identification of acoustic cues: (1) To isolate the cue along the time axis, speech sounds are truncated in time from the onset with a step-size of 5, 10, or 20 ms, depending on the duration and type of consonant. (2) To locate the cue along the frequency axis, speech sounds are highpass and lowpass filtered before being presented to normal hearing listeners. (3) To measure the strength of the cue, speech sounds are masked by white noise of various signal-to-noise ratio. The three plots on the top row illustrate how the speech sound is processed. Typical correspondent recognition scores are depicted in the plots on the bottom row.

portance of individual speech perception events for sounds other than /t/, the 3DDS requires three independent experiments for each CV utterance. 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 sound and its duration. The *second* experiment divides the fullband into multiple bands of equal length along the BM, and measures the score in different frequency bands by using highpass/lowpass filtered speech as the stimuli. Once the time-frequency coordinate of the event have been identified, a *third* experiment assesses the strength of the speech event by masking the speech at various signal-to-noise ratios. To reduce the length of the experiments, the three dimensions, i.e., time, frequency and intensity, are assumed to be independent. This is not always true, though; therefore, the identified events are verified by a special software designed for the manipulation of acoustic cues, based on the short-time Fourier transform [102, 103].

In order to understand continuous speech, it is necessary to first identify the acoustic

correlates of the individual phonemes, for which the movement of the articulators are more easily interpretable [41]. For this reason, we will first look at the normal events of individual consonants as they occur in isolated syllables in which their acoustic properties are well formed. The interaction between the events in continuous speech must be addressed in future studies. Finally, the 3D method has been successfully applied to the remainder of the 16 Miller-Nicely consonants [104], but for both space and pedagogical reasons, the discuss will be limited to the six stop consonants.

3.3 Methods

The details of the time-truncation (TR07), high/lowpass filtering (HL07) and noise masking "Miller-Nicely (2005)" (MN05) experiments are described below. Each abbreviation gives the experiment type followed by the year the experiment was executed. An analysis of the MN05 experiment has since been published [9].

Subjects: Sixty-two listeners were enrolled in the study, of which 19 subjects participated in HL07, and 19 subjects participated in TR07. One subject participated in both experiments. The remaining 25 subjects were assigned to experiment MN05 [9]. The large majority of the listeners were undergraduate students, while the rest were mothers of teenagers. No subject was older than 40 years, and all self-reported no history of speech or hearing disorder. All listeners spoke fluent English, with only slight regional accents. Except for two listeners, all the subjects were born in the U.S. with their first language (L1) being English. The subjects were paid for their participation. University of Illinois's Institutional Review Board approval was attained.

Speech Stimuli: A significant characteristic of natural speech is the variability of the acoustic cues. Thus we designed the experiment by manually selecting six different utterances per CV consonant, based on the criterion that the samples be representative of the corpus.

The 16 [64] (MN55) CVs /pa, ta, ka, fa, θa, sa, ʃa, ba, da, ga, va, ðа, za, ʒa, ma, na/ were chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus," which were used as the common test

material for the three experiments. The speech sounds were sampled at 16 kHz using a 16 bit analog-to-digital converter. Each CV was spoken by 18 talkers of both genders. Experiment MN05 uses all 18 talkers $\times$ 16 consonants. For the other two experiments (TR07 and HL07), 6 talkers, half male and half female, each saying each of the 16 MN55 consonants, were manually chosen for the test. These 96 (6 talkers $\times$ 16 consonants) utterances were selected such that they were representative of the speech material in terms of confusion patterns and articulation score based on the results of earlier speech perception experiment [9,91]. The speech sounds were presented diotically (same sounds to both ears) through a Sennheisser HD 280 Pro headphone, at each listener's most comfortable level (MCL) (i.e., between 75 to 80 dB SPL, based on a continuous 1 kHz tone in a homemade 3 cc coupler, as measured with a Radio Shack sound level meter. All experiments were conducted in a single-walled IAC sound-proof booth.

Conditions: Three experiments were performed, denoted TR07, HL07 and MN05. All three experiments included a common condition of fullband speech at 12 dB SNR, as a control.

Experiment TR07 evaluates the temporal property of the events. Truncation starts from the beginning of the utterance and stops at the end of the consonant. These times were all determined by hand. The truncation times were also manually chosen, such that the duration of the consonant was divided into non-overlapping consecutive intervals of 5, 10, or 20 ms. An adaptive scheme was applied for the calculation of the sample points. The basic idea was to assign more points where the speech changed rapidly, and fewer points where the speech was in a steady condition, in a manner consistent with the findings of [101]. Starting from the end of the consonant, near the consonant-vowel transition, eight frames of 5 ms were allocated, followed by twelve frames of 10 ms, and as many 20 ms frames, as needed, until the entire interval of the consonant was covered. To make the truncated speech sounds more natural, and to remove an possible onset truncation artifacts, white noise was used to mask the speech stimuli, at an SNR of 12 dB.

Experiment HL07 investigated the frequency properties of the events. Nineteen filtering conditions, including one full-band (250-8000 Hz), nine highpass and nine lowpass conditions, were included. The cutoff frequencies were calculated using the Greenwood

function, so that the full-band frequency range was divided into 12 bands, each having an equal length along the basilar membrane. The highpass cutoff frequencies were 6185, 4775, 3678, 2826, 2164, 1649, 1250, 939, and 697 Hz, with an upper-limit of 8000 Hz. The lowpass cutoff frequencies were 3678, 2826, 2164, 1649, 1250, 939, 697, 509, and 363 Hz, with the lower-limit being fixed at 250 Hz. Note that the highpass and lowpass filtering share the same cutoff frequencies over the middle range. The filters were implemented in Matlab (The Mathworks Inc.) via a 6^{th} order elliptical filter, with a stop band of 60 dB. White noise having a 12 dB SNR was again added, to make the modified speech sounds more natural sounding.

Experiment MN05 assesses the strength of the event in terms of noise robust speech cues, under adverse conditions of high noise. Besides the quiet condition, speech sounds were masked at eight different SNRs: -21, -18, -15, -12, -6, 0, 6, 12 dB, using white noise. The details of MN05 may be found in [91].

Procedures: The three experiments employed similar procedures. A mandatory practice session was given to each subject at the beginning of the experiment. In each experiment, the general methods were to randomize across all variables when presenting the stimuli to the subjects. There was one important exception to this rule, being MN05, where effort was taken to match the experimental conditions of [64] as closely as possible, as discussed in [9]. Following each presentation, subjects responded to the stimuli by clicking on the button labeled with the CV that they heard. In case the speech was completely masked by the noise, the subject was instructed to click a "Noise Only" button. If the presented token did not sound like any of the 16 consonants, the subject had the option to either guess one of the 16 sounds, or click the "Noise Only" button. To prevent fatigue, listeners were asked to take frequent breaks, or break whenever they feel tired. Subjects were allowed to play each token for up to 3 times before making their decision, after which the sample was placed in the list, at the end. A Matlab program was created for the control of the three procedures. The audio was played using a SoundBlaster 24 bit sound card in a standard PC Intel computer, running Ubuntu Linux.

A preliminary analysis of the raw data

The experimental results of TR07, HL07 and MN05 take the form of *confusion patterns* (CP), which display the probabilities of all possible responses (the target and competing sounds), as a function of the experimental conditions, i.e., truncation time, cutoff frequency and signal-to-noise ratio.

Notation: Let $c_{x|y}$ denote the probability of hearing consonant /x/ given consonant /y/. When the speech is truncated to time t_n the score is denoted $c^T_{x|y}(t_n)$. The score of the lowpass and highpass experiment at cutoff frequency f_k is indicated as $c^{L/H}_{x|y}(f_k)$. Finally the score of the masking experiment as a function of signal-to-noise ratio is denoted $c^M_{x|y}(SNR_k)$.

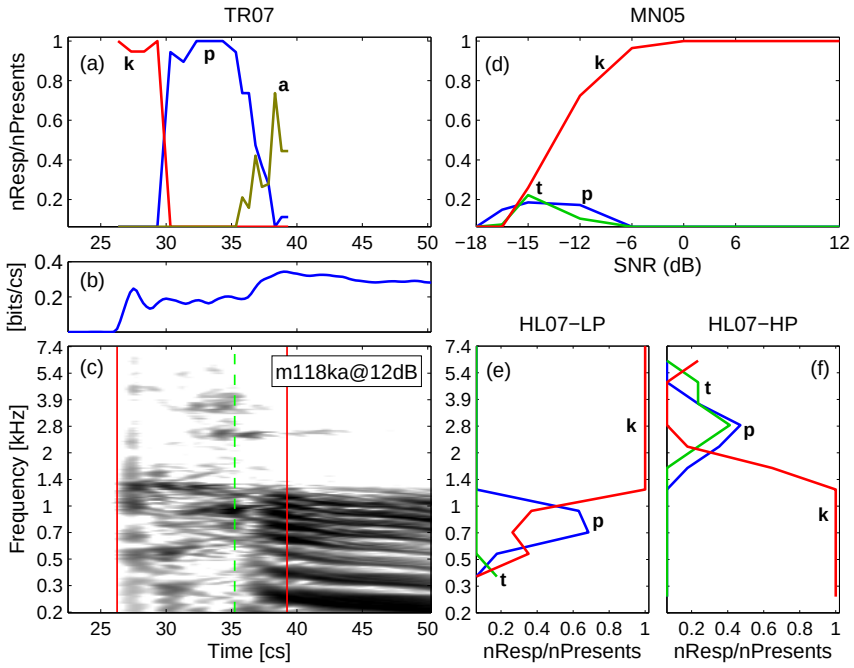


Figure 3.3: Various CPs of /ka/ spoken by talker m118, under various experimental conditions: (a) The temporal truncation CP as a function of truncation time t_n [cs], from experiment TR07. (b) Instantaneous AI $a_n \equiv a(t_n)$ at truncation time t_n. (c) AI-gram at 12 [dB] SNR. The left and right vertical lines denote the start and end time for truncation. The middle (green) line denotes the time of voice (sonorant) onset. (d) CP as a function of SNR for experiment MN05. Finally, the (e) high and (f) low CPs as a function of cutoff frequency for HL07. The text provides further details.

A specific example is helpful to explain the 3D method, and to show how speech perception is affected by the events. Figure 3.3 depicts the CPs of /ka/ produced

by talker m118 (utterance `m118_ka`). The results of experiment TR07 are given in Fig. 3.3(a), HL07-lowpass in Fig. 3.3(e), HL07-highpass in Fig. 3.3(f) and MN05 in Fig. 3.3(d). The instantaneous AI (Eq. 3.4) is shown in panel (b), and the AI-gram in (c). To facilitate the integration of the three experiments, the AI-gram and the three scores are aligned in time (t_n in centiseconds [cs]) and frequency (along the cochlear place axis, but labeled in frequency), and thus depicted in a compact manner.

The CP of TR07 [Fig. 3.3(a)] shows that the probability of hearing /ka/ is 100% for $t_n \leq 26$ cs, while no speech component is removed. However, at around 29 cs when the /ka/ burst has been completely truncated, the score for /ka/ drops sharply to 0% within a span of 1 cs. At this time (32–35 cs) only the transition region is heard, and 100% of the listeners report hearing a /pa/. Once even the transition region is truncated, listeners report hearing only the vowel /a/.

A related conversion occurs in the lowpass and highpass experiment HL07 for /ka/ [Fig. 3.3(e,f)], in which both the lowpass score $c^L_{k|k}$ and highpass score $c^H_{k|k}$ plunge from 100% to less than 10% at a cutoff frequency of $f_k = 1.4$ kHz, thereby defining the frequency location of the /ka/ cue. For the lowpass case, listeners reported a morphing from /ka/ to /pa/ with score $c^L_{p|k}$ reaching 70% at 0.7 kHz, and for the highpass case, /ka/ morphed to /ta/, but only at the $c^H_{t|k} = 0.4$ (40%) level. To reduce the clutter, the remaining confusions are not shown.

The MN05 masking data [Fig. 3.3(d)] shows a somewhat related CP. When the noise level increases from quiet to 0 dB SNR, the recognition score of /ka/ is close to 1 (i.e., 100%), which usually signifies the presence of a robust event.

It is satisfying (and significant) that the [42] finding, that the acoustic cue for /ka/ is a mid-frequency burst at the beginning of the sound, is directly confirmed by the above experimental results.

3.4 Results

In this section we demonstrate how the events of stop consonants are identified by applying the 3DDS. Again the results from the three experiments are arranged in a compact form for convenience. Take Fig. 3.4 for example; panel (a) shows the AI-gram of the speech sound at 18 dB SNR, upon which each event hypothesis is highlighted by a

rectangular box. The middle vertical dashed line denotes the voice-onset time, while the two blue vertical solid lines on either side of the green dashed line denote the starting and ending points for the time truncation experiment (TR07). Above the AI-gram (a), panel (b) shows the scores from TR07, while to the right, panel (d) shows the scores from HL07. Panel (c) depicts the scores from experiment MN05. The CP functions are plotted as solid (lowpass) or dashed (highpass) curves, with competing sound scores with a single letter identifier next to each curve. The $*$ in panel (c) indicates the SNR at which the listeners just begin to confuse the sound in MN05, while the $\star$ in panel (d) indicates the intersection point of the highpass and lowpass scores, measured in HL07. The six small figures (e) along the bottom show partial AI-grams of the consonant region, delimited in panel (a) by the solid lines, at -12, -6, 0, 6, 12, 18 dB SNR. A box in any of the seven AI-grams of panels (a) or (e) indicates a hypothetical event region, and for (e), indicates its visual threshold, according to the AI-gram model. The methods presented in these figures are significant extensions of the work of [94].

3.4.1 Stops

In the following sections we shall study the stop consonants /p/, /t/, /k/, /b/, /d/ and /g/ followed by vowel /a/ as in "father." For each consonant, six utterances were analyzed, discussed by the research group, and the most representative example was subjectively chosen to be presented.

/pa/

Figure 3.4 for /pa/ spoken by female talker f103 (LDC file `s_f103_pa.wav`) reveals that there may be two different events: (1) a formant transition at 1–1.4 kHz, which appears to be the dominant cue, maskable by white noise at 0 dB SNR; and (2) a wide band click running from 0.3–7.4 kHz, maskable by white noise at 12 dB SNR.

Stop consonant /pa/ is traditionally characterized as having a wide-band click which is seen in this /pa/ example, but not in the five others we have studied. For most /pa/s, the wide-band click diminishes into a low frequency burst. The click does appear to contribute to the overall quality of /pa/ when it is present.

Time: Figure 3.4(b) shows the truncated /p/ score $c^T_{p|p}(t_n)$. It starts at 100% from

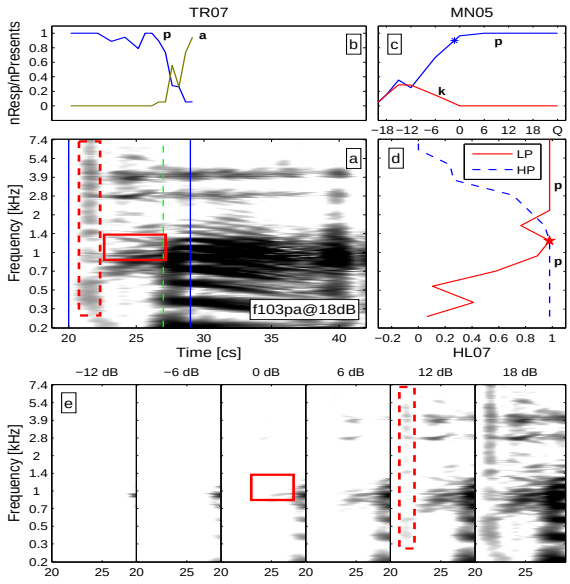


Figure 3.4: Hypothetical events for /pa/ from talker f103: (a) AI-gram. A dashed vertical line labels the onset of voicing (sonorance), indicating the start of the vowel. The solid boxes indicate hypothetical sources of events. (b) CPs as a function of truncation time t_n. (c) CPs as a function of SNR_k. (d) CPs as a function of cutoff frequency f_k. (e) AI-grams of the consonant region [defined by the solid vertical lines on panel (a)], at -12, -6, 0, 6, 12, 18 dB SNR. While the wide-band click becomes barely intelligible when SNR<12 dB, the F_2 transition remains audible at 0 dB SNR.

the beginning. When the wide band click, which includes the low frequency burst, is truncated at around 23 cs, the score is seen to drop, but not significantly. It is only when the transition is removed at 27 cs that the score suddenly drops to the chance level (1/16). At this time subjects begin to report hearing the vowel /a/ alone. Thus, even though the wide-band click contributes slightly to the perception of /pa/, the F_2 transition appears to play the main role.

Frequency: The lowpass and highpass scores, as depicted in Fig. 3.4(d), start at 100% at each end of the spectrum, and only begin to drop near the intersection point, close to 1.3 kHz. This intersection (indicated by a $\star$) appears to be a clear indicator of the center frequency of the dominant perceptual cue, which is the F_2 region running from 22 cs to 26 cs, as labeled by the truncation data in panel (b).

Amplitude: The recognition score $c^M_{p|p}$ as a function of SNR [Fig. 3.4(c)] drops to 90% at 0 dB SNR (SNR_{90} denoted by $*$), at the same time the /pa/→/ka/ confusion $c^M_{k|p}$ starts a slow but steady increase. In the 6 aigrams of panel (e) we can see that the

audible threshold for the F_2 transition is at 0 dB SNR, the same as the SNR_{90} point in panel (c) where the listeners begin to lose the sound, giving credence to the energy of F_2 sticking out in front of the sonorant portion of the vowel, as the main cue for /pa/ event.

Summary and other /pa/ data: The 3D displays of the other five /pa/s (not shown) are in basic agreement with that of Fig. 3.4, with the main difference being the existence of the wideband burst at 22 cs for f103, and slightly different highpass and lowpass intersection frequency, ranging from 0.7 to 1.4 kHz, for the other five sounds. The required duration of the F_2 energy before the onset of voicing was seen around $3-5$ cs before the onset of voicing and this timing, too, is very critical to the perception of /pa/. The existence of excitation of F_3 is evident in the AI-grams, but it does not appear to interfere with the identification of /pa/, unless F_2 has been removed by filtering (a minor effect for f103). Also /ta/ was identified in a few examples, as high as 40% when F_2 was masked.

/ta/

From Fig. 3.5, the /ta/ event for talker f105 is a short high frequency burst above 4 kHz, 1.5 cs in duration and 5-7 cs prior to the vowel.

Time: In panel (b), the score for the truncated /t/ drops dramatically at 28 cs, and remains at chance level for later truncations, suggesting that the high frequency burst is critical for /ta/ perception. It is interesting to see that at around 29 cs, when the burst has been completely truncated and the listeners can only listen to the transition region, listeners start reporting a /pa/. By 32 cs, the /pa/ score climbs to 85%. This is in total agreement with the results of /pa/ events in the previous section. Once the transition region is also truncated (as indicated in Fig. 3.5(a) by the dashed line at 36 cs) subjects report hearing only the vowel, with the transition from 50% /pa/ $\rightarrow$ /a/ occurring at 37 cs.

Frequency: In panel (d), the intersection of the highpass and the lowpass perceptual scores (indicated by the $\star$) is at around 5 kHz, showing the dominant cue to be the high frequency burst. From the lowpass CPs (solid curve) we see that once the high frequency burst has been removed, the /ta/ score $c^L_{t|t}$ drops dramatically. From the

off-diagonal lowpass CP data $c^L_{p|t}$ (solid curve labeled "p" at 1 kHz), confusion with /pa/ is very high once all the high frequency information is removed. This can be easily explained by referring to our results of Fig. 3.4, which shows the significance of the F_2 transition around 1 kHz for /pa/ identification. Given only low frequency bands, while /ta/ cannot be perceived, it can be guessed (chance must play an important role when the set-size is small). The best alternative in such cases seems to be a low frequency /pa/, as found from our previous results of Fig. 3.4. The highpass results are in agreement with the view that /ta/ results from the high frequency burst.

Amplitude: The /ta/ burst has an audible threshold of -1 dB SNR in white noise, defined as the SNR where the score drops to 90%, namely SNR_{90} [labeled by a $*$ in panel (c)]. Once the /ta/ burst is masked at -6 dB SNR, subjects report /ka/ and /ta/ equally, with a reduced score around 30%. From the AI-grams in (e) we see that the high frequency burst is lost between 0 dB and -6 dB, consistent with the results of Fig. 3.5(c) that SNR_{90}=-1 dB SNR.

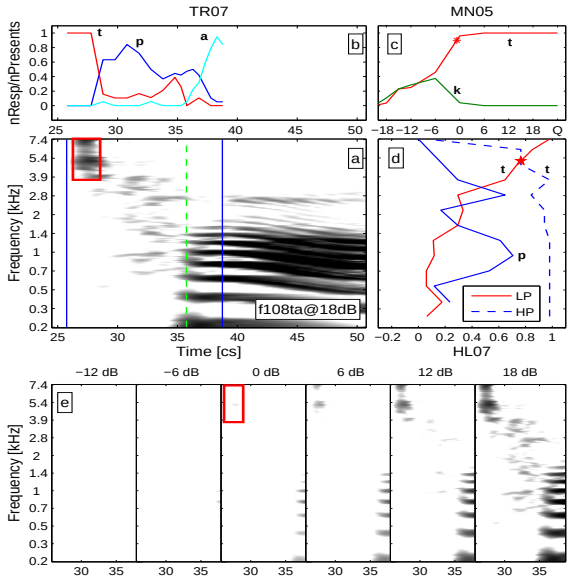


Figure 3.5: Hypothetical event for /ta/ from talker f105. (a) AI-gram with identified event highlighted by a rectangular box. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR. The event becomes masked at 0 dB SNR.

Summary and other /ta/ data: In summary, the event of /ta/ is verified to be a high frequency burst above 4 kHz. The perception of /ta/ is highly dependent on the

identified event, which explains the sharp drop in scores when the high frequency burst is masked. These results are therefore in complete agreement with the earlier analysis of /t/ by [94], as well as many of the conclusions from the 1950s Haskins Laboratories research.

Of the six /ta/ sounds, five morphed to /pa/ once the /ta/ burst was truncated (e.g., Fig. 3.5(b)), while one morphed to /ka/ (m112ta), with an impressive 90% score. This same sound also became /ka/ rather than /pa/ following lowpass filtering below 2.8 kHz, with a 100% score. For this particular sound, it is seen that the /ta/ burst precedes the vowel only by around 2 cs as opposed to 5–7 cs which is the case for a normally articulated /ta/. This timing cue is especially important for the perception of /pa/ since the transition region and relative timing of this transition region are critical to /pa/ perception.

/ka/

Analysis of Fig. 3.6 reveals that the event of /ka/ is a mid-frequency burst around 1.6 kHz, articulated $5-7$cs before the vowel, as highlighted by the rectangular boxes in panels (a) and (e). The following /ka/ event analysis is relatively straightforward.

Time: Figure 3.6(b) shows that once the mid-frequency burst is truncated at 16.5 cs, the recognition score $c^T_{k|k}$ jumps from 100% to chance level within 1–2 cs. At the same time, most listeners begin to hear /pa/ with the score ($c^T_{p|k}$) rising to 100% at 22 cs, which is in excellent agreement with previous conclusion about the /pa/ feature. Frequently [as seen in panel (a)] there are high frequency (e.g., 3-8 kHz) bursts of energy, but usually not of sufficient amplitude to trigger /t/ responses. Since these /ta/-like bursts occur around the same time as the mid-frequency /ka/ feature, time truncation of the /ka/ burst results in the simultaneous truncation of these potential /t/ cues. Thus truncation beyond 16.5 cs results in confusions with /p/, not /t/. Beyond 24 cs, subjects report only the vowel.

Frequency: According to Fig. 3.6(d) the highpass score $c^H_{k|k}$ and the lowpass score $c^L_{k|k}$ cross each other at 1.4 kHz. Both curves have a sharp dive around the intersection point, suggesting that the perception of /ka/ is dominated by the mid-frequency burst as highlighted in panel (a). The highpass $c^H_{t|k}$ [dashed curve of panel (d)] shows minor

confusions with /ta/ (e.g., 40%) for $f_c > 2$ kHz. This is in agreement with the conclusion about the /ta/ feature being a high frequency burst. Similarly, the lowpass CP around 1 kHz shows strong confusions with /pa/ ($c^L_{p|k} = 90\%$), when the /ka/ burst is absent.

Amplitude: From the AI-grams [panel (e)], the burst is just above its detection threshold at 0 dB SNR; accordingly, the recognition score of /ka/ $c^M_{k|k}$ [panel (c)] drops rapidly at 0 dB SNR. At -6 dB SNR the burst has been fully masked, and most listeners report /pa/ instead of /ka/.

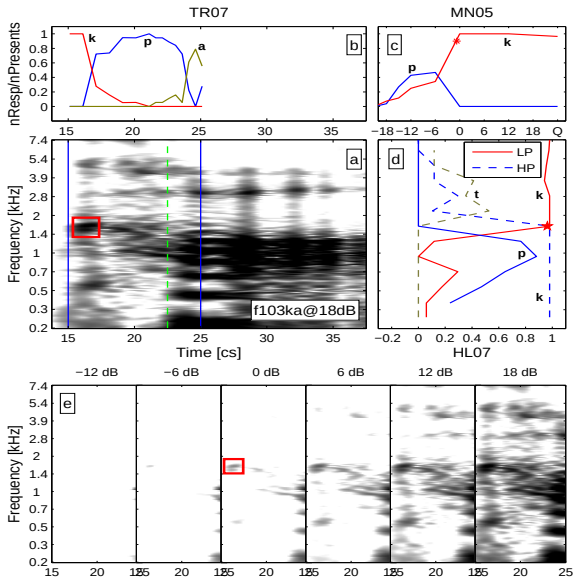


Figure 3.6: Hypothetical event for /ka/ from talker f103. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR. The event remains audible at 0 dB SNR.

Summary and other /ka/ data: Not all of the six sounds strongly morphed to /pa/ once the /ka/ burst was truncated, as is seen in Fig. 3.3(a) and 3.6(b). Two out of six had no morphs, and just remained a very weak /ka/ once the onset-burst was removed (m114ka, f119ka). Again, these scores are consistent with guessing.

In casual experiments (not reported on here) we have tried shifting the burst along the frequency axis, reliably morphing /ta/ into /ka/ or /pa (or *vice versa*). When the burst of /ka/ or /ta/ is masked or removed, the auditory system can pick up residual transitions in the low frequency, which would cause the sound to morph to /pa/.

In all the speech perception tests, /pa, ta, ka/ commonly form a confusion group.

This can be explained by the fact that the three sounds share the same type of event patterns, i.e., burst and F_2 transition. The relative timing for these three unvoiced sounds is nearly the same. The major difference lies in the center frequencies of the bursts, with /pa/ cue in the low frequency, /ka/ cue in the mid-frequency, and /ta/ cue in the high frequency.

/ba/

The perceptual events for /ba/ are perhaps the most difficult of the six stops. For the 3D method to work well, high scores in quiet are essential. Among the six /ba/ sounds, only the one shown (f111) had 100% scores at 12 dB SNR and above.

Based on the analysis of Fig. 3.7, the hypothetical features for /ba/ include: (1) a wide-band click in the range of 0.3 to 4.5 kHz; (2) a low frequency burst around 0.4 kHz; and (3) a F_2 transition around 1.2 kHz.

Time: When the wide-band click is completely truncated at t_n =28 cs, the /ba/ score $c^T_{b|b}$ [Fig. 3.7(b)] drops dramatically from 80% to chance level, at the same time the /ba/→/va/ confusion $c^T_{v|b}$ for and /ba/→/fa/ confusion $c^T_{f|b}$ increase quickly, indicating that the wide-band click is important for distinguishing /ba/ from the two fricatives /va/ and /fa/. However, since the three events overlap on the time axis, it is hard to tell which event plays the major role.

Frequency: Figure 3.7(d) shows that the highpass score $c^H_{b|b}$ and lowpass score $c^L_{b|b}$ cross each other at 1.3 kHz, both change fast within 1–2 kHz, indicating that the F_2 transition, centered around 1.3 kHz, is very important. Without the F_2 transition, as we see in the lowpass data while f_c <1 kHz, most listeners guess /da/ instead of /ba/. Besides, the small jump in the lowpass score $c^L_{b|b}$ around 0.4 kHz suggests that the low frequency burst may also play a role in /ba/ perception.

Amplitude: From the AI-grams in Fig. 3.7(e), the F_2 transition and wide-band click become masked by the noise somewhere below 0 dB SNR. Accordingly the listeners begin to lose /ba/ in the masking experiment around the same SNR, as represented by SNR_{90} ($*$) in panel (c). Once the wideband click has been masked, the confusions with /va/ increase, and become equal to /ba/ at -12 dB SNR with a score of 40%.

Summary and other /ba/ data: There are only two LDC /ba/ sounds out of 18

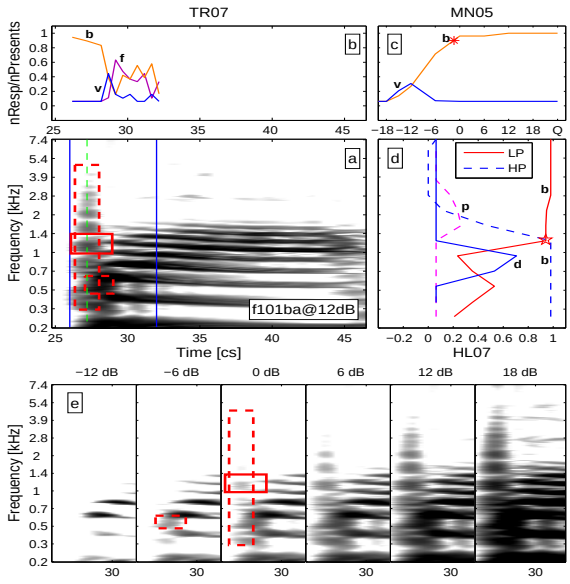


Figure 3.7: Hypothetical events for /ba/ from talker f101. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR. The F_2 transition and wide-band click become masked around 0 dB SNR, while the low frequency burst remains audible at -6 dB SNR.

with 100% scores at and above 12 dB SNR, i.e., /ba/ from f101/ shown here and /ba/ from f109, which has a 20% /va/ error rate for $SNR \leq -10$ dB SNR. The remaining 16 /ba/ utterances have /va/ confusions between 5 and 20%, in quiet. We do not know if the recordings in the LDC database are responsible for these low scores, or if /ba/ is inherently difficult. Low quality consonants with error rates greater than 20% were also observed in the LDC study by [91]. These very low starting (quiet) scores are part of our difficulty in identifying the /ba/ event with certainty, since the 3D method requires high scores in quiet for its proper operation.

From unpublished research that is not fully described here, we have found that in order to achieve a high quality /ba/ (defined as 100% identification in quiet), the wide-band burst must exist over a wide frequency range. For example, a well defined 3 cs burst from 0.3 to 8 kHz will give a strong percept of /ba/, which, if missing or removed, may likely be heard as /va/ or /fa/.

/da/

Consonant /da/ (Fig. 3.8) is the voiced counterpart of /ta/. It is characterized by a high frequency burst above 4 kHz and an F_2 transition near 1.5 kHz, as shown in panels (a) and (e).

Time: Truncation of the high frequency burst [panel (b)] leads to an immediate drop in the score of $c^T_{d|d}$ from 100% at 27 cs to about 70% at 27.5 cs. The recognition score keeps decreasing until the F_2 transition is removed completely at 30 cs, when subjects report only hearing vowel /a/. The truncation data indicate that both the high frequency burst and F_2 transition are important for /da/ identification.

Frequency: The lowpass score $c^L_{d|d}$ and highpass score $c^H_{d|d}$ cross at 1.7 kHz. Notice that subjects need to hear both the F_2 transition and the high frequency burst to get a full score of 100%, meaning that both events are critical for a high quality /da/. Lack of the burst usually leads to the /da/→/ga/ confusion, as shown by the lowpass confusion of $c^L_{g|d}$ =30% at f_c=2 kHz [solid curve labeled "g" in panel (d)].

Amplitude: From the AI-grams [panel (e)] the F_2 transition becomes masked by noise at 0 dB SNR; accordingly the /da/ score $c^M_{d|d}$ in panel (c) drops quickly at the same SNR. When the remnant of the high frequency burst is finally gone at -6 dB SNR, the /da/ score $c^M_{d|d}$ decreases even faster, until $c^M_{d|d} = c^M_{m|d}$ at -10 dB SNR, namely, until the /d/ and /m/ scores are equal.

Summary and other /da/ data: Two other /da/ sounds (`f103, f119`) showed a dip where the lowpass score decreases abnormally as the cutoff frequency increases, similar to that seen for /da/ of m118 (i.e., 1.2-2.8 kHz). Two showed larger gaps between the lowpass score $c^L_{d|d}$ and highpass score $c^H_{d|d}$. The 6^{th} /da/ had a very wide-band burst going down to 1.4 kHz. In this case the lowpass filter did not reduce the score until it reached this frequency. For this example the cutoff frequencies for the high- and lowpass filtering were such that there was a clear crossover frequency having both scores at 100%, at 1.4 kHz. Some of the /da/s are much more robust to noise than others. For example, the SNR_{90}, defined as the SNR where the listeners begin to lose the sound (P_c=0.90), is -6 dB for /da/-m104, and +12 dB for /da/-m111. The variability over the six utterances is impressive, yet the story seems totally consistent with the requirement that both the burst and the F_2 transition need to be heard.

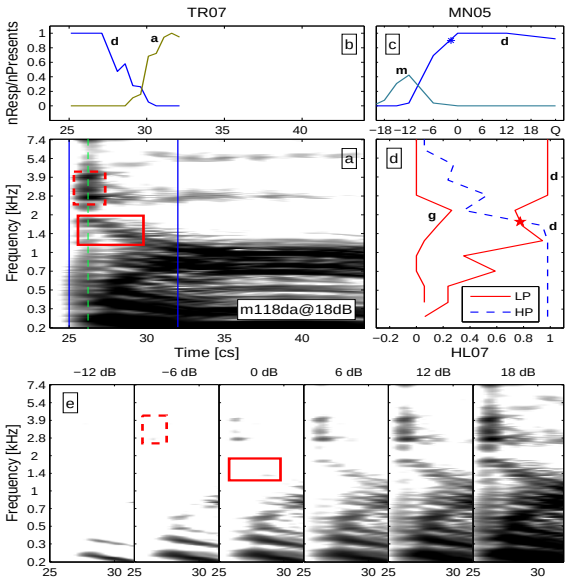


Figure 3.8: Hypothetical events for /da/ from talker m118. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR. The F_2 transition and the high frequency burst remain audible at 0 and -6 dBSNR respectively.

/ga/

The events of /ga/ include a mid-frequency burst from 1.4 to 2 kHz, followed by an F_2 transition between 1 and 2 kHz, as highlighted with boxes in Fig. 3.9(a).

Time: According to Fig. 3.9(b), the recognition score of /ga/ $c^T_{g|g}$ starts to drop when the mid-frequency burst is truncated beyond 22 cs. At the same time the /ga/→/da/ confusion appears, with $c^T_{d|g}$=40% at 23 cs. From 23 to 25 cs the probabilities of hearing /ba/ and /da/ are equal. And the reason for this low-grade confusion is that the two sounds have similar patterns of F_2 transitions. Beyond 26 cs, where both events have been removed, subjects only hear the vowel /a/.

Frequency: From Fig. 3.9(d) the highpass (dashed) score and lowpass (solid) score fully overlap at the frequency of 1.6 kHz, where both show a sharp decrease of more than 60%, which is consistent with the statements about /ga/ events. There is minor /ba/ confusion $c^L_{b|g}$ =20% at 0.8 kHz and /da/ confusion $c^H_{d|g}$ =25% at 2 kHz. This can be explained by the fact that /ba/, /da/ and /ga/ all have the same types of events, i.e., bursts and transitions, allowing for guessing within the confusion group, given a burst onset coincident with voicing.

Amplitude: Based on the AI-grams in panel (e), the F_2 transition is masked by 0 dB SNR, corresponding to the turning point of $c^M_{g|g}$ labeled by a $*$ in panel (c). As the mid-frequency burst gets masked at -6 dB SNR, /ga/ becomes confused with /da/.

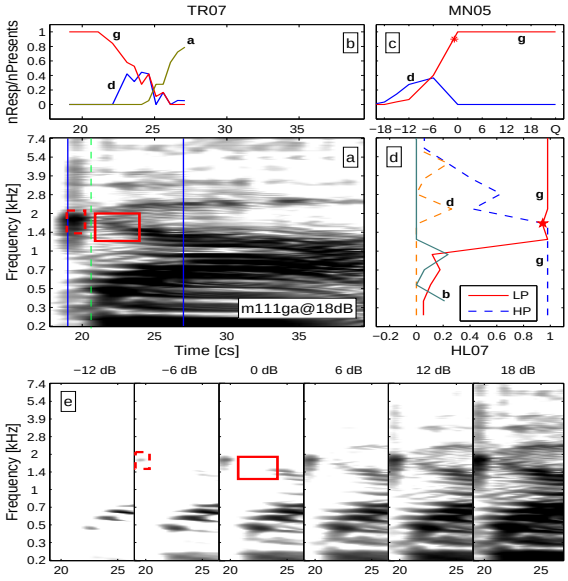


Figure 3.9: Hypothetical events for /ga/ from talker m111. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR. The F_2 transition is barely intelligible at 0 dB SNR, while the mid-frequency burst remains audible at -6 dBSNR.

Summary and other /ga/ data: All six /ga/ sounds have well defined bursts between 1.4 and 2 kHz with well correlated event detection threshold [predicted by AI-grams in panel (e)] versus SNR_{90} [denoted by $*$ in panel (c)], the turning point of recognition score where the listeners begin to lose the sound. Most of the /ga/s (`m111, f119, m104, m112`) have a perfect score of $c^M_{g|g}$ =100% at 0 dB SNR. The other two /ga/s (`f109, f108`) are relatively weaker; their SNR_{90} are close to 6 dB and 12 dB respectively.

3.4.2 Fricatives

Fricatives are sounds produced by an incoherent noise excitation of the vocal tract. This noise is generated by turbulent air flow at some point of constriction. In order to produce a fricative, a talker must position the tongue or lips to create a constriction

width of 2–3 mm and allow air pressure to build behind the constriction so that the turbulence needed is created. Fricatives may be voiced like the consonants /v, ð, z, ʒ/ or unvoiced like the consonants /f, θ, s, ʃ/.

/fa/

The dominant perceptual cue is between 0.8 and 2.8kHz and lasts for about 80 cs in duration, as shown in Fig. 3.10.

Time: According to Fig. 3.10(b) the recognition score of the target sound decreases gradually from 1 at $t = 26$ cs to chance level at $t = 34$ cs, where the probability of reporting /ba/ equals that of /fa/. After that the recognition score of /fa/ goes to zero and the confusion of /ba/ increases dramatically, suggesting that the /fa/ cue of is from 25 to 34 cs.

Frequency: Refering to Fig. 3.10(d), the high frequency score (dashed) and low frequency score (solid) cross at 1.6 kHz. Both curves change fast within the mid-frequency range. The recognition accuracy saturates once the high-pass and low-pass cutoff frequencies reach around 700 Hz and 2.8 kHz respectively, thus we can conclude that the dominant cue is in the range of 0.8–2.8 kHz.

Amplitude: Figure 3.10(c) depicts the confusion pattern of noise masking data. The recognition score of /fa/ drops dramatically at 0 dB SNR, where the mid-frequency cue (Fig. 3.10(e)) becomes inaudible. Below that, the probability of correctness goes to chance level.

Summary and other /fa/ data: For talkers m111 and f101, the high-pass curve and low-pass curve cross around 1.4 kHz. For the other four talkers, the two curves cross at 1 kHz or slightly below. The event strength of other /fa/s shows a large amount of variance. Except for m111 and f101, the fricative cues are relatively weak for the other four talkers m112, f103, m117, and f105. As a result, their recognition scores begin to drop at 12 dB SNR.

/θa/

The perceptual data indicates that /θa/ does not have a dominant cue. Figure 3.11 depicts the AI-grams and perceptual scores of speech sound /θa/. The truncation score

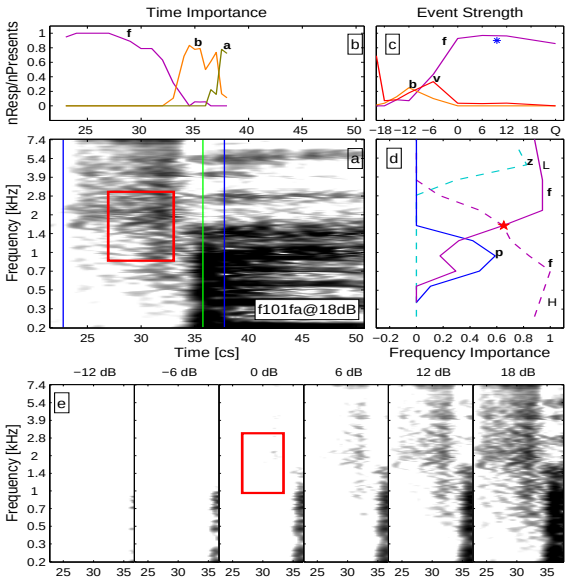


Figure 3.10: Hypothetical events for /fa/ from talker f101. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

(Fig. 3.11(b)) starts at a low level (less than 0.8), indicating that the perceptual cues are weak. The same is true for the high/low-pass data (Fig. 3.11(d)) and noise-masking data (Fig. 3.11(c)). Looking at the confusion plots embedded in the upper left panel, it can be seen that /θ/ does not have a fixed confusion group. It is confused with a large number of speech sounds and there is no fixed pattern for these confusions. Based on all of this information, it is safe to say that /θa/ does not have a compact dominant cue; therefore, it is confused with many sounds.

/sa/

The dominant perceptual cue of /sa/ is seen to be between 4 and 8 kHz and spans for around 100 ms just before the vowel is articulated (refer to Fig. 3.12). This cue is seen to be robust to white noise of around 0 dB SNR.

Time: As depicted in Fig. 3.12(b), the truncation score starts from 1 at t=26 cs. It begins to decrease quickly at t=32 cs, and then reaches the same level as /ta/ at t=34 cs.

Frequency: According to Fig. 3.12(d), the lowpass experiment data shows that it is

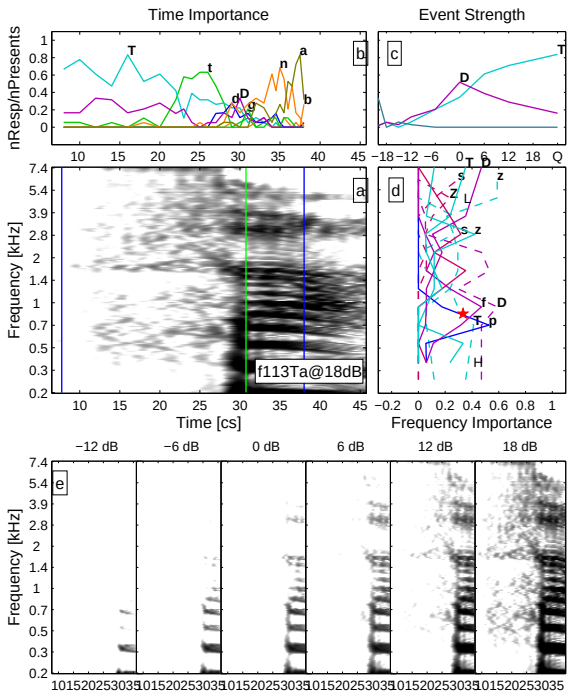


Figure 3.11: Hypothetical events for /θa/ from talker f113. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

only after the cutoff frequency goes above around 3 kHz that the score steadily rises till a score of 0.9 is reached at 7.4 kHz. For the highpass filtering, there is a steady rise in score as the cutoff frequency goes below 7.4 kHz and it goes almost to 0.9 at around 4 kHz. In both cases, the change in score is pretty abrupt, signifying that the feature is well defined in frequency.

Amplitude: Based on Fig. 3.12(c), the recognition score remains 100% until the signal-to-noise ratio reaches 0 dB SNR, suggesting that the dominant cue is masked so that the normal hearing listeners begin to lose the sound. The AI-grams in Fig. 3.12(e) confirm the conjecture. At -6 dB SNR, the high frequency cue is totally inaudible.

Summary and other /sa/ data: The identified /sa/ cues are consistent across different talkers. Except for a /sa/ from talker m117, which has a low recognition accuracy of less than 0.6 even in quiet, the other five /sa/s all have a salient high frequency cue on the AI-gram. For m112, f108, f109, the /sa/ cues are still audible at 0 dB SNR. It is relatively weaker for the other two utterances.

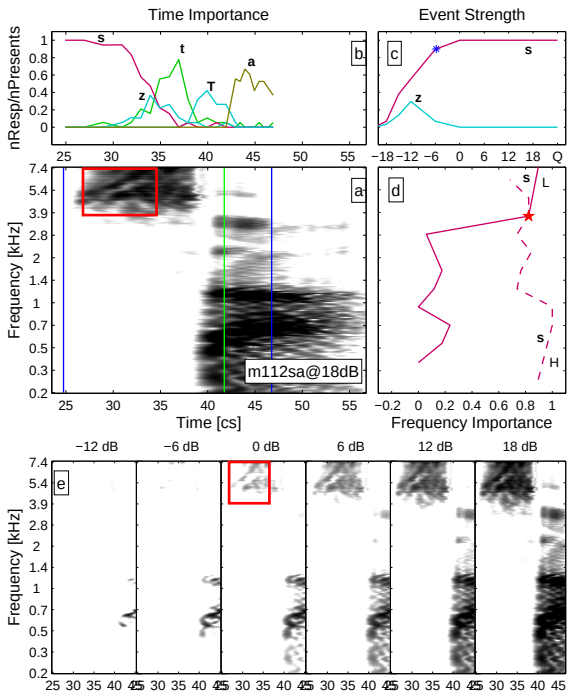


Figure 3.12: Hypothetical events for /sa/ from talker m112. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

/ʃa/

For consonant /ʃ/, the dominant perceptual cue is from 2 to 4 kHz and lasts more than 160 ms before the vowel starts, as shown in Fig. 3.13.

Time: Refering to Fig. 3.13(b), the probability of hearing /ʃa/ decreases from $t = 22$ cs to $t = 37$ cs continuously, suggesting that the duration is a key parameter for the perception of /ʃa/. Due to the long duration, the sound is confused with no other sounds until t =36 cs, where the target sound morphs into /ʒa/.

Frequency: Based on Fig. 3.13(d), the lowpass score shows a sharp increase when lowpass cutoff frequency equals 2 kHz. The highpass score remains at chance levels, but once the cutoff frequency is above 4 kHz. The score increases significantly and reaches the peak value when the cutoff frequency goes below 2 kHz. These scores clearly suggest that the /ʃ/ perceptual feature lies in the range of 2–4 kHz.

Amplitude: The noise-masking data (Fig. 3.13(c)) shows a sharp decrease at -6 dB SNR, meaning that the perceptual cue of /ʃ/ is strong enough to resist white

noise at that level. The perceptual data is consistent with the prediction of AI-grams (Fig. 3.13(e)) which shows that the dominant cue of /ʃa/ is barely intelligible at -6 dB SNR.

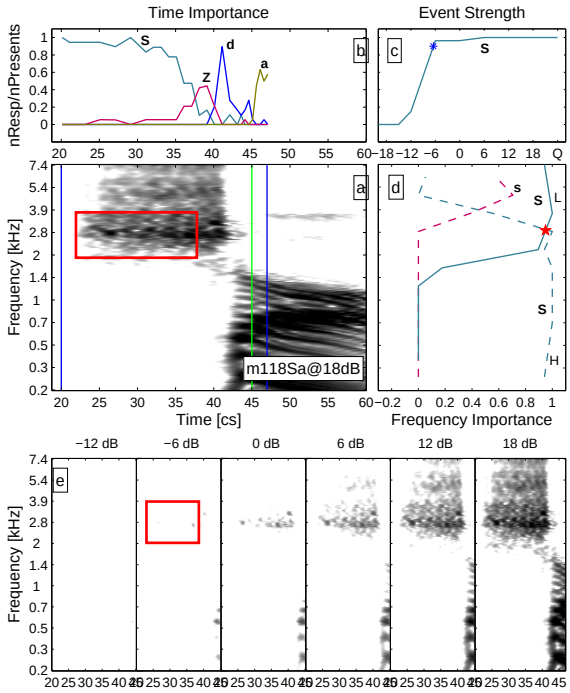


Figure 3.13: Hypothetical events for /ʃa/ from talker m118. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR. The speech cue is strong enough to resist white noise at -6 dB SNR.

Summary and other /ʃa/ data: The perceptual cue of /ʃa/ is consistent across six male and female talkers in that the feature has about the same long duration and covers the same frequency range. One of the cues (m118) is still audible at -6 dB SNR, others are at 0 or 6 dB SNR.

The unvoiced fricatives all have the feature regions around and above 2 kHz and span a considerable duration before the vowel. For the case of /sa/ and /ʃa/, the events of both sounds come on at the same time, with the only difference being that the burst for /ʃa/ is slightly lower in frequency than /sa/. Eliminating the burst at that frequency in the case of /ʃ/ should give rise to the sound /sa/. Even though the /θ/ feature is not unknown, when the masking is applied these four sounds are confused with each other, as shown by the Miller-Nicely experiments and verified by this study. Masking by white

noise in particular can cause these confusions increasingly as the white noise would act as a lowpass filter on these sounds that have relatively high frequency cues, and this would considerably alter the cues of the masked sounds, thus leading to confusions between /f/, /θ/, /s/ and /ʃ/.

/ða/

Much like /θ/, /ð/ (see Fig. 3.14) has a large number of confusions with several different sounds, indicating that it does not have a strong compact perceptual cue. For the highpass, lowpass and truncation experiments especially, the highest scores are around 0.4-0.5 on average. It is very difficult to make any sort of conjecture with /θ/ and /ð/ as far as feature regions are concerned.

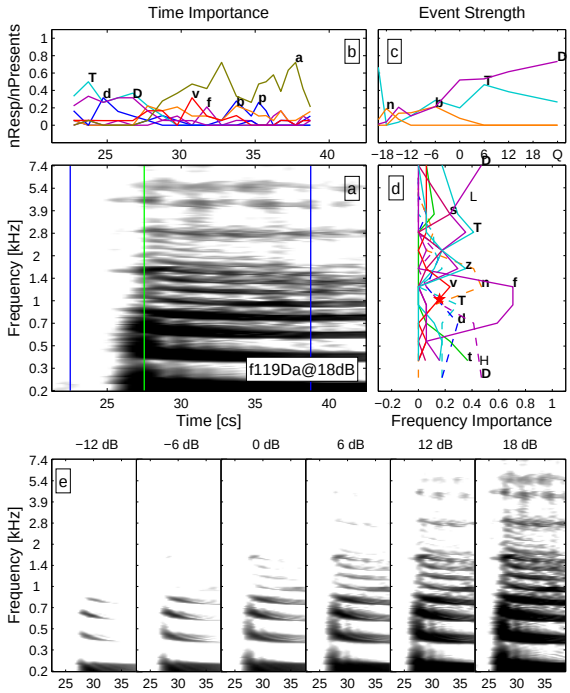


Figure 3.14: Hypothetical events for /ða/ from talker f119. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

/va/

The /va/ cue is from 0.5 to 1.4 kHz, highlighted in the mid-left panel of Fig. 3.15. Due to the similarity between /va/ and /ba/, the fricative sound is often confused with the bilabial stop.

Time: According to Fig. 3.15(b) the truncation score drops quickly from close to 1 at t=25 cs to chance level at t=29 cs. Beyond that, most listeners report /ba/ or vowel /a/ only, suggesting that the highlighted area is critical for /va/ identification.

Frequency: Based on Fig. 3.15(d) the highpass and lowpass scores cross at $f = 0.7$ kHz. Both curves change fast from 0.5 to 1.4 kHz, which isolates the feature area.

Amplitude: Figure 3.15(c) depicts the confusion patterns of /va/ when the sound is masked by white noise. It is confused with /ba/ even in quiet. The recognition score decreases gradually as the noise level increases and turns sharply at 0 dB SNR, when the perceptual cue is completely wiped out (refer to the AI-grams at Fig. 3.15(e)).

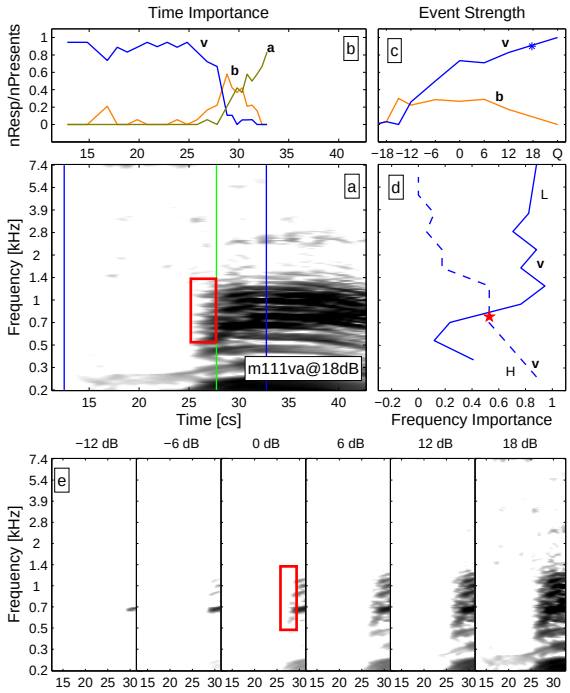


Figure 3.15: Hypothetical events for /va/ from talker m111. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

Summary and other /va/ data: The perceptual cue of /va/ is consistent across

different talkers. However, the fricative shows a big variance in terms of the intensity of the cue. Two of the talkers, f103 and f105, have low scores, suggesting that the perceptual cues are weak, or barely audible. The /va/ from talker m104 has a much longer duration.

/za/

The /za/ feature is seen to lie between 3 and 7.5 kHz and spans around 50-70 ms before the vowel is articulated as highlighted in the mid-left panel of Fig. 3.16. This feature is seen to be robust to white noise of -6 dB SNR.

Time: According to Fig. 3.16(b), the truncation score drops dramatically at t=36 cs, then the fricative morphs into /da/. Combining the perceptual data of /sa/, it is easy to tell that the perceptual cue of /za/ is within the highlighted box.

Frequency: Refering to Fig. 3.16(d), the lowpass score climbs only when the cutoff frequencies reach around 2.8kHz. The highpass score remains constant above 4 kHz. There is a brief dip in the score which is an indication of an interfering cue of /ʒa/. The perceptual cue is in the same frequency range as /sa/.

Amplitude: The perceptual cue of /za/ is strong enough to be audible at -6 dB SNR in white noise, as suggested by the sharp turning point in Fig. 3.16(c) and the AI-grams depicted in Fig. 3.16(e).

Summary and other /za/ data: Except for /za/ from talker f109, which has a low recognition score of less than 0.8 in quiet, the /za/s produced by the other five talkers all have a salient cue in the high-frequency. Most of them can resist white noise at 0 or even -6 dB SNR.

/ʒa/

The /ʒa/ perceptual cue (see Fig. 3.17) is present at around 2–4 kHz spanning around 50–70 ms before the vowel is articulated. This cue is robust to white noise of 0 dB SNR.

Time: Fig. 3.17(b) depicts the confusion patterns of /ʒa/ when the sound is truncated in time. The recognition score changes gradually from t=14 to 25 cs, then drops quickly to chance level, at the same time the fricative morphs into a /da/. The /ʒa/ sound is similar to /ʃa/ in that both have a long duration.

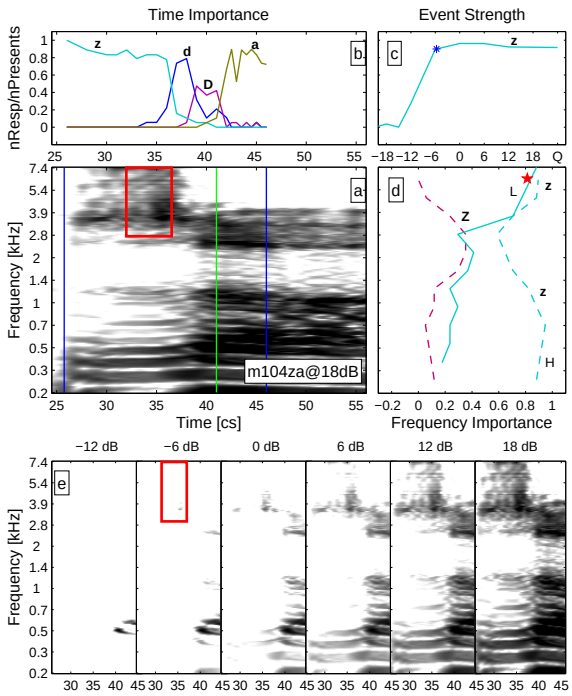


Figure 3.16: Hypothetical events for /za/ from talker m104. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

Frequency: The lowpass score and the highpass score (Refer to Fig. 3.17(d)) cross at f=2.8 kHz. The lowpass score (dashed) saturates when the cutoff frequency reaches 4 kHz. The highpass score drops quickly when the cutoff frequency goes below 2 kHz. Combining the high/low-pass data, it is easy to tell that the perceptual cue is from 2 to 4 kHz.

Amplitude: Based on Fig. 3.17(c) the masked recognition score remains constant until the signal-to-noise level drops to 0 dB SNR, suggesting that the /ʒa/ cue is barely audible at that noise level. The AI-grams (Fig. 3.17(e)) show that the /ʒa/ cue is missing at -6 dB SNR. Due to the imperfection of the AI-gram, which over predicts the cue audibility, the perceptual data and the AI prediction mismatch by a few dB.

Summary and other /ʒa/ data: The identified /ʒa/ cues are consistent across all six talkers. The intersection points fall within the same frequency range, i.e., from 2 to 4 kHz. Without exception, the feature shows a long duration for all cases. Most of the /ʒa/ cues are still audible at 0 dB SNR in white noise.

In the case of the voiced fricatives, it is noticed that /f/ and /θ/ are not prominent

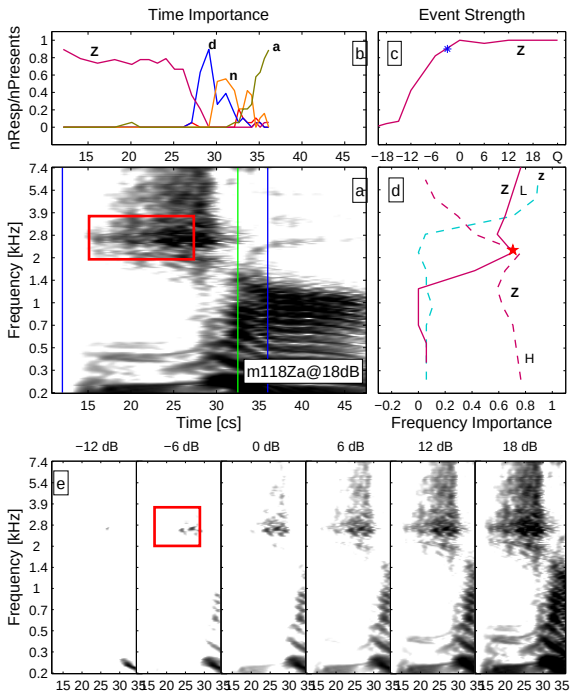


Figure 3.17: Hypothetical events for /ʒa/ from talker m118. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

in the confusion group of /f/, /θ/, /s/ and /ʃ/ primarily as /f/ has stronger confusions with the voiced consonant /b/ and unvoiced fricative /v/ and /θ/ has no consistent patterns as far as confusions with other consonants are concerned. Similarly for the unvoiced fricatives, /v/ and /ð/ are not prominent in the confusion group as /v/, too, is often confused with /b/ and /f/, and /ð/ shows no consistent confusions.

3.4.3 Nasals

As the name suggests, nasal sounds are those for which the nasal tract provides the main sound transmission channel. A complete closure is made toward the front of the vocal tract, either by the lips, by the tongue at the gum ridge or by tongue at the hard or soft palate, and the velum is opened wide. As may be expected, most of the sound radiation takes place at the nostrils. The nasal consonants used in this study include /m/ and /n/.

/ma/

The perceptual cues of /ma/ include the nasal murmur around 100 ms before the vowel is articulated and the F2 region between 0.5 and 1.2 kHz as highlighted in Fig. 3.18(a).
Time: The recognition score of truncated /ma/ (Fig. 3.18(b)) remains constant until t hits 24 cs, indicating that the onset of the F2 formant is critical for /ma/ distinction. Notice that the /ma/ sound is confused with no other sound because of the nasal murmur. The truncation score does not change with the nasal murmur, suggesting that it is a feature for both /ma/ and /na/.
Frequency: The highpass and lowpass scores cross at 1 kHz (refer to Fig. 3.18(d)), suggesting that the F2 region is important. The lowpass score changes dramatically as the cutoff frequency increases from 50% at 0.3 kHz to 100% at 0.6 kHz when the cutoff frequency hit the feature area. With the highpass experiment, a sudden decrease in score is seen when the cutoff frequency changes from 1.4 to 2 kHz. A further decrease in the cutoff frequency leads to increasing scores again, which reach 1 at around 1 kHz. The above information clearly isolates the /ma/ feature in the frequency domain.
Amplitude: Based on the confusion patterns of /ma/ in white noise, as depicted in (Fig. 3.18(c)), the /ma/ sound is very robust. It has no other confusions until the signal-to-noise ratio drops to -12 dB SNR.
Summary and other /ma/ data: The /ma/ cues are consistent across different talkers in that all the identified cues are within the beginning area of F_2. Most /ma/s are still highly intelligible at -12 dB SNR.

/na/

The perceptual cue of /na/ (see Fig. 3.19) includes a low frequency nasal murmur about 80–100 ms before the vowel and an F_2 transition around 1.5 kHz, as indicated by the clear peak on the frequency importance function.
Time: Like /m/, the time importance function for /n/ is also seen to have a peak till around the transition region. The truncation score (Fig. 3.19(b)) decreases as the cutoff frequency cuts into the F2 transition, the perceptual cue that distinguishes /na/ from /ma/.
Frequency: Based on the high/low-pass data (Fig. 3.19(d)), the two curves cross each

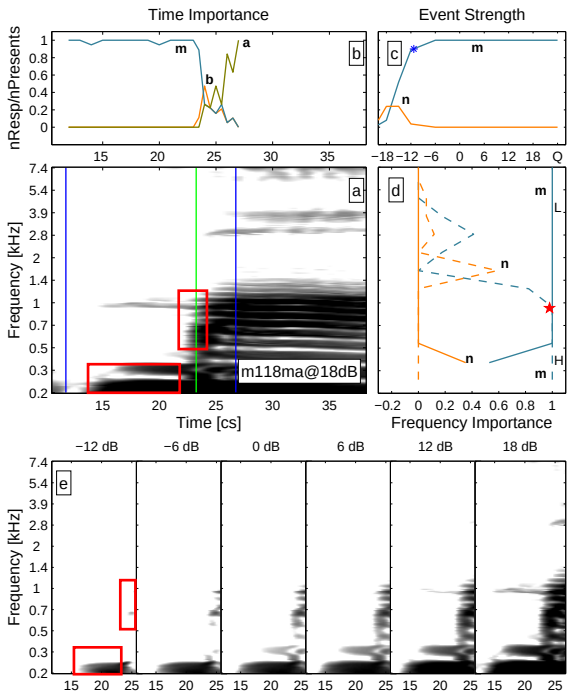


Figure 3.18: Hypothetical events for /ma/ from talker m118. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

other at 1.3 kHz, which pinpoints the weight of the speech sound. The low-pass score is seen to be at chance until the cutoff frequency goes above 0.4 kHz, then it steadily increases. An intermittent peak is observed in the low-pass curve at around 0.5–1 kHz. For the case of the high pass data, the score is high when the cutoff frequency goes below 1.4 kHz.

Amplitude: According to the noise masking data (Fig. 3.19(c)), the perceptual score of /na/ remains 100% until the signal-to-noise ratio drops to -6 dB SNR, suggesting that the /na/ is strong enough to resist white noise at -6 dB SNR, which is consistent with the AI-grams in (Fig. 3.19(e)).

Summary and other /na/ data: The perceptual cue of /na/ is consistent across multiple talkers. The high- and lowpass scores crosse at the F2 area without exception for the /na/s from all six talkers. Most of the /na/ cues are still audible at -6 dB SNR.

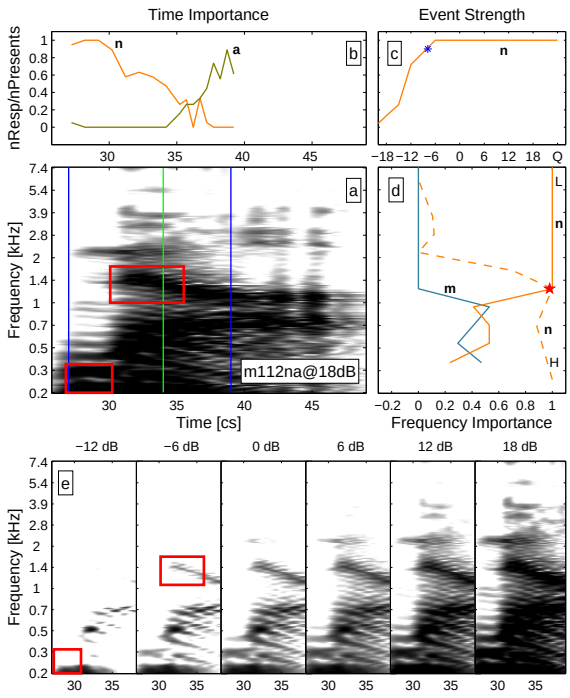


Figure 3.19: Hypothetical events for /na/ from talker m112. (a) AI-gram with identified events highlighted by rectangular boxes. (b, c, d) CPs of TR07, HL07 and MN05. (e) AI-grams of the consonant part at -12, -6, 0, 6, 12, 18 dB SNR.

3.4.4 Robustness

The robustness of consonant sounds is determined mainly by the strength of the dominant cue. In our experiment it is common to see that the recognition score of a speech sound remains unchanged as the masking noise increases from a low intensity; it then drops when the noise reaches a certain level, at which point the dominant cue becomes barely intelligible. The study in [94] found that the threshold of speech perception with the probability of correctness being equal to 90% (SNR_{90}) is proportional to the threshold of the /t/ burst, using a Fletcher critical band measure (the AI-gram). In the present study a related rule is identified for the remaining five stop consonants. Figure 3.20 depicts the scatter-plot of SNR_{90} versus the threshold of audibility for the dominant cue. For a particular sound (each point on the plot), the SNR_{90} is interpolated from the PI function, while the threshold of audibility for the dominant cue is estimated from the 36 AI-gram plots [panel (e)] of Figs. 3.5–3.9. The two thresholds are nicely correlated in this chart, indicating that the recognition of each stop consonant is mainly dependent

on the audibility of the dominant cues. Speech sounds with stronger cues are easier to hear in noise than weaker cues because it takes more noise to mask them. When the dominant cue (typically the burst) becomes masked by noise, the target sounds are easily confused with other consonants. The masking of an individual cue is typically over about a 6 dB range, and never more. It is an all-or-nothing detection task. It is the spread of the event threshold that is large, not the masking of a single cue.

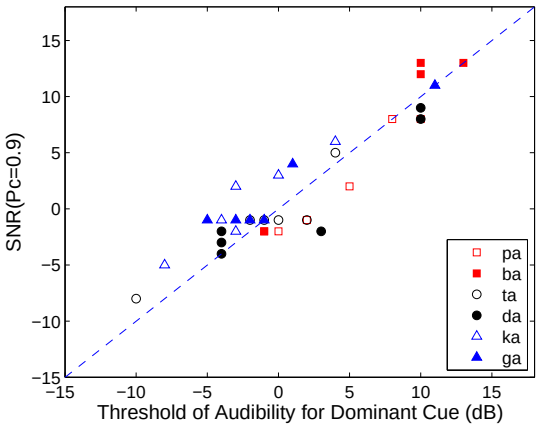


Figure 3.20: Correlation between the threshold of consonant identification and the audible threshold of dominant cues.

3.4.5 Event distributions

A significant characteristic of natural speech is the large variability of the acoustic cues across the speakers. Typically this variability is characterized by using the spectrogram. It was for exactly this reason that we designed the experiment as we did, by manually selecting six different utterances per consonant based on our criterion that the samples have the natural variability representative of the corpus. Since we did not, at the time, know the exact acoustic features, this was a guess at best.

We now know that the key parameters are the timing of the stop burst, relative to the sonorant onset of the vowel (i.e., the center frequency of the burst peak and the time difference between the burst and voicing onset). These variables are depicted in Fig. 3.21 for the 36 utterances. The figure shows that the burst times and frequencies for stop consonants are well separated across the different talkers. The utterance groups seem to nicely separate in this space, but placing a statistical value, given only 6 in each

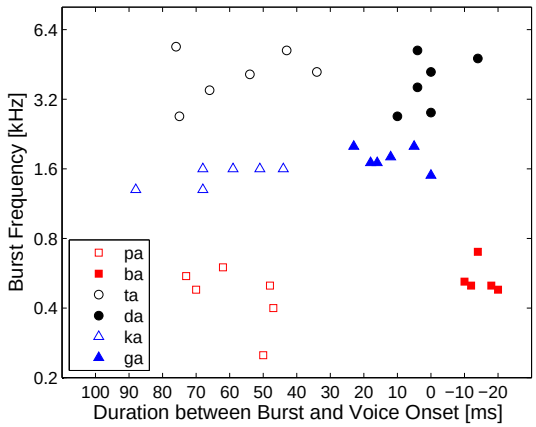


Figure 3.21: Variability of the bursts for stop consonants preceding vowel /a/.

class, while not impossible, seems a stretch not worth the trouble.

3.4.6 Speculations on the source of events

The following is a summary of the acoustic features that define stop consonant events, for all of the 6 examples for the 6 stop consonants that we studied in this report.

Unvoiced stops: **/pa/:** As the lips abruptly release, they are used to excite primarily the F_2 formant relative to the others (e.g., F_3). This resonance is allowed to ring for approximately 5–20 cs before the onset of voicing (sonorance) with a typical value of 10 cs. For the vowel /a/, this resonance is between 0.7–1.4 kHz. A poor excitation of F_2 leads to a weak perception of /pa/. Truncation of the resonance does not totally destroy the /p/ event until it is very short in duration (e.g., $\leq$ 2 cs).

A wideband burst is sometimes associated with the excitation of F_2, but is not necessarily audible (or visual in the AI-grams). Of the six example /pa/ sounds, only f103 showed this wideband burst. When the wideband burst was truncated, the score dropped from 100% to just above 90%.

/ta/: The release of the tongue from its starting place behind the teeth mainly excites a short duration (1-2 cs) burst of energy at high frequencies ($\geq$ 4 kHz). This burst typically is followed by the sonorance of the vowel about 5 cs later. The case of /ta/ has been well studied by [94], and the results of the present study are in good agreement.

All but one of the /ta/ examples morphed to /pa/, with that one morphing to /ka/, following lowpass filtering below 2 kHz, with a maximum /pa/ morph of close to 100%, when the filter cutoff was near 1 kHz.

/ka/: The release for /k/ comes from the soft-pallet, but like that of /t/, is represented with a very short duration, high energy burst near F_2, typically 10 cs before the onset of sonorance (vowel). In our six examples there is almost no variability in this duration. In many examples the F_2 resonance could be seen following the burst, but at reduced energy relative to the actual burst. In some of these cases, the frequency of F_2 could be seen to change following the initial burst. This seems to be a random and unimportant variation, since several /ka/ examples showed no trace of F_2 excitation. A proper test of this question remains an open question.

Five of the six /ka/ sounds morphed into /pa/ when lowpass filtered to 1 kHz. The sixth morphed into /fa/, with a score around 80%.

Voiced stops: /ba/: Since the sounds we chose to analyze have weak events, it is difficult to generalize the source. Only two of the six /ba/ sounds had scores above 90% in quiet (f101 and f111). Based on our 3D analysis of these two /ba/ sounds, it appears that the main source of the event is the wide-band burst release itself rather than the F_2 formant excitation as in the case of /pa/. This burst can excite all the formants, but since the sonorance starts within a few cs, it seems difficult to separate the excitation due to the lip excitation from that due to the glottis.

The four sounds with low scores had no visible onset burst, and all have scores below 90% in quiet. Consonant /ba-f111/ has 20% confusion with /va/ in quiet, and had only a weak burst, with a 90% score above 12 dB SNR. Consonant /ba-f101/ has a 100% score in quiet and is the only /b/ with a well developed burst, as shown in Fig. 3.7.

/da/: This consonant shares many properties in common with /ta/ other than its onset timing since it comes on with the sonorance of the vowel. The range of the burst frequencies tends to be lower than with /ta/, and in one example (`m104`), the lower frequency went down to 1.4 kHz. The low burst frequency was used by the subjects in identifying /da/ in this one example, in the lowpass filtering experiment. However, in all cases the energy of the burst always included 4 kHz. The large range seems significant, going from 1.4 to 8 kHz. Thus, while release of air off the roof of the mouth may be

used to excite the F_2 or F_3 formants, to produce the burst, several examples showed a wide-band burst seemingly unaffected by the formant frequencies.

/ga/: In the six examples studied here, the /ga/ consonant was defined by a burst that is compact in both frequency and time, and very well controlled in frequency, always being between 1.4 and 2 kHz. In 5 out of 6 cases, the burst is associated with both F_2 and F_3, which can clearly be seen to ring following the burst. Such resonance was not seen with /da/, which seems notable.

3.5 General Discussion

The speech events are the information bearing aspects of the speech code. From what we have found, the acoustic cues that support the events have a low density in time-frequency space.

It was shown by Shannon [105] that the performance of a communication system is dependent on the code of the symbols to be transmitted. The larger the "distance" between two symbols, the less likely the two will be confused. This principle also applies to the case of human speech perception. For example, the /pa, ta, ka/ have common perceptual cues, i.e., a burst followed by a transition. Once the burst is removed or masked by noise, the three sounds are highly confusable.

It is interesting to see that many speech sounds contain acoustic cues that conflict with each other. Take f103ka (Fig. 3.6) for example. In addition to the mid-frequency /ka/ burst, it also contains two burst in the high and low frequency ranges that greatly increase the probability of perceiving the sound as /ta/ and /pa/ respectively [Fig. 3.6(d)]. We call this type of misleading onset a *conflicting cue*.

An especially interesting case is the confusions between /ba/ and /va/. Traditionally these two consonants were attributed to two different cothnfusion groups based on their articulatory and distinctive features. However, in our experiments, we find that consonants with similar events tend to form a confusion group. Thus /ba/ and /va/ are highly confusable with each other because they share the common F_2 transition. This is strong evidence that events are the basic units for speech perception.

Summary: The stop consonants are defined by a short duration burst (e.g., 2 cs), characterized by its center frequency (high, medium and wide-band), and the delay to

the onset of voicing. This delay, between the burst and the onset of sonorance, is a second parameter called "voiced/unvoiced."

There is an important question about the relevance of the wide-band click at the onset of the bilabial consonants /p/ and /b/. For /pa/ this click *appears* to be an option that adds salience to the sound. For /ba/ our data is clearly insufficient, but it *appears* that the click is the key to the /ba/ event.

In contrast, /ta/ and /ka/ are dominated by the burst frequency and delay to the sonorant onset. The voiced and unvoiced stops differ in the duration between the burst and the voicing onset. Confusion is much more common between /g/ and /d/ than between /t/ and /k/. The unvoiced bilabial /b/ is most often confused with the fricatives /v/ and /f/, seen in many CPs.

The fricatives (/v/ being an exception) are characterized by an onset of wide-band noise created by the turbulent airflow through lips and teeth. Duration and frequency range are the two critical parameters of the events. A voiced fricative usually has a considerably shorter duration than its unvoiced counterpart. /θ/ and /ð/ are not included in the schematic drawing because no stable events have been found for these two sounds.

The two nasals /m/ and /n/ share a common feature of nasal murmur in the low frequency. As a bilabial consonant, /m/ has a formant transition similar to /b/, while /n/ has a formant transition close to /g/ and /d/. Recall that for each consonant, we selected six utterances based on the criterion that the samples are representative of the corpus. The events of the consonants are very consistent across the different talkers, despite the fact that the parameters, such as timing, frequency and strength, may change to a certain degree within the given range.

3.5.1 Limitations of the method

It is important to point out that the AI-gram is highly imperfect, in that it is based on a linear model which does not account for cochlear compression, forward masking, upward masking and other neural nonlinear responses. Such important nonlinearities are discussed at length in many places, e.g., [18, 19, 69, 106]. A major extension of the AI-gram is in order, but not easily obtained. Given the extent of such a project, we

have continued to use the linear version of the AI-gram until a fully tested time-domain nonlinear cochlear model becomes available. The model of [107] may presently be the only candidate for such testing.

Nevertheless, based on our many listening tests, we believe that the linear AI-gram generates a useful approximation under many circumstances [94,100]. It is easy (trivial) to find cases where time-frequency regions in the speech signals are predicted to be audible by the AI-gram, but when removed, result in a signal with inaudible differences. In this sense, the AI-gram contains a great deal of "irrelevant" information. Thus it is a gross "over-predictor" of audibility. There are rare cases where the AI-gram "under-predicts" audibility, namely where it fails to show an audible response; yet when that region is removed, the modified signal is audibly different. Such cases, to our knowledge, are rare, but when discovered, are examples of a serious failure of the AI-gram. Finally, and perhaps most important, the relative strengths of cues will be misrepresented. For example, it is well known that onsets are strongly represented in neural responses due to adaptation [19]. Such cues are not properly present in the AI-gram, and this weakness might be relatively easily fixed, using existing hair-cell and neural models.

CHAPTER 4

IMPACT OF SENSORINEURAL HEARING LOSS ON CONSONANT IDENTIFICATION

This paper investigates the impact of sensorineural hearing loss on the perception of consonant sounds. In addition to pure tone audiometry (PTA), threshold equalized noise (TEN) test [108] and psychoacoustic tuning curve (PTC) tests [109] are utilized to diagnose possible cochlear dead regions. A speech perception test is conducted to measure the hearing impaired listeners' performance on 16 consonants /p, t, k, f, θ, s, ʃ, b, d, g, v, ð, z, ʒ, m, n/ in speech-weighted noise. To determine the correlation between the shift in hearing threshold and speech intelligibility, an extended speech banana that accounts for the steady-state masking noise is developed to predict the audibility of the dominant cue for individual consonants, given the pure tone audiogram and the signal-to-noise ratio of speech stimuli. Five subjects with bilateral sensorineural hearing loss volunteered for the study. Results show that audibility successfully accounts for the disability of speech perception for subjects with mild-flat hearing loss, but fails for the cases of cochlear dead region and extremely unbalanced (e.g., severe high-frequency) loss.

4.1 Introduction

People with hearing loss often complain about the difficulty of hearing speech in cocktail party environments [7]. Depending on the configuration of hearing loss, a HI listener may easily hear certain sounds and have serious problems with some others. To explain why, the following two questions need to be addressed: (1) What are the perceptual cues making up speech sounds? (2) What is the impact of sensorineural hearing loss (SNHL) on speech perception? In [110], a systematic psychoacoustic method has been developed to identify the perceptual cues of consonant sounds. With that information, this chapter investigates the effect of SNHL on the perception of individual consonants.

A key research question is: *Does audibility, as characterized by the pure tone audiogram, fully account for the loss of intelligibility for individual consonants?*

Most sensorineural hearing loss can be attributed to the malfunctioning of cochlear 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, and also results in an elevated detection threshold. It is generally assumed that a mild-to-moderate elevation in threshold primarily reflects OHC loss, while a moderate-to-severe hearing loss indicates an additional IHC loss. In the past decade Moore and his colleagues coined the concept of cochlear dead regions (CDR), an extreme case of IHC loss, and developed the threshold equalized noise (TEN) test [6,108] and a psychoacoustic tuning curve (PTC) test [109] for the detection of CDR. An important implication of those studies is that a pure tone audiogram is not a good indicator of the physiological nature of the hearing loss [6]; specifically, subjects with OHC loss and IHC loss may show the same amount of shift in hearing threshold.

Due to the lack of means for the quantization of OHC and IHC hearing loss, most studies on hearing impaired speech perception have been focused on the correlation between the disability in speech perception and the shift in hearing threshold. Bilger and Wang investigated the effect of hearing loss on articulatory features using INDSCAL and claimed that there is generally a relationship between the audiometric configuration and consonant confusions [111, 112]. In [113–115] the cochlear hearing loss was simulated by frequency-specific attenuation (filtering) or masking normal hearing listeners with spectrally shaped broadband noise; no consistent difference was observed between hearing-impaired listeners and masked normal hearing listeners. On the other hand, the Speech Intelligibility Index (SII), which uses audiometric configuration to compensate for the hearing loss, has been found inaccurate in predicting the performance of speech perception for hearing impaired listeners [90]. To fill the gap between the detection of pure tone and complex speech signals, Plomp and his colleagues proposed a test of speech intelligibility named speech-reception threshold (SRT), defined as the level of speech for a fixed 50% score under fluctuating noise and steady-state noise [2,116–118].

The impact of hearing loss on the perception of individual sounds is basically unknown.

The speech banana is a pure tone audiogram labeled with common sounds. It has long been used for the qualitative assessment of general speech perception ability for HI listeners. The speech cues are based on the formant data of vowels and consonants measured by Fant [119] during the 1940s. When all the speech sounds are plotted on the audiogram, they can be contained by the shape of banana. Two factors limit the wide use of the speech banana in audiological clinics. First, it does not account for noise, which by itself is a much more important factor than intensity (loudness) in terms of speech perception. Second, most consonants are not defined by the formants. During the last several years, the relevant speech cues have been established for normal hearing listeners [110,120,121]. Based on the accurate information of speech cues and Fletcher's method of calculating effective hearing loss in masking noise [70], the speech banana can be extended and used for the analysis of hearing impaired speech perception in speech-weighted noise.

In this chapter we investigate the impact of SNHL on consonant identification by integrating the accurate information about speech cues and configuration of hearing loss. In addition to PTA, the TEN test and PTC test are applied to diagnose possible cochlear dead regions. Based on the analysis of a large amount of data, it is hypothesized that HI listeners have difficulty understanding noisy speech because they cannot hear the weak sounds for which the characteristic acoustic cues are missing due to their hearing loss and the masking effect introduced by the noise.

4.2 Extended Speech Banana

The extended speech banana is aimed for the prediction of intelligibility of individual sounds in steady-state noise for the hearing impaired listeners. It requires two components: accurate information about speech cues and effective hearing loss in the presence of masking noise.

Since the stop consonants are characterized by a compact burst that falls within a single auditory filter, while the fricatives all have a wide-band noise-like cue that covers multiple critical bands, it is easier to study the stop consonants than the fricatives. In the rest of this chapter we will focus on stop consonants and leave the fricatives and

nasals for future study.

4.2.1 Acoustic cues of stop consonants

Letting D_C denote the peak intensity density of a speech cue, the RMS level can be approximated by $D_C - 6$ dB SPL. Assuming that the speech cue only covers the bandwidth of a single auditory filter, the intensity of speech cue I_C can be estimated by integrating the RMS level along the critical bandwidth,

$$I_C = D_C - 6 + 10 \log_{10} ERB \tag{4.1}$$

where *ERB* is the *Equivalent Rectangular Bandwidth* of the auditory filter described by the following equation [122]:

$$ERB = 24.7(4.37F + 1) \tag{4.2}$$

where *ERB* is in Hz and F is the center frequency in kHz.

Table 4.1 lists the peak intensity density and center frequency (CF) of the dominant cue, the burst, for 36 stop consonants. The center frequency is measured by using the 3D deep search method as described in Chapter 3. The peak intensity density is estimated by comparing the amplitude of the short-time Fourier transform coefficients to that of a pure tone with given intensity.

Notice that the speech sounds have been normalized to 80 dB SPL before the measurement. For speech sounds at other levels the actual D_C needs to be corrected by the difference in intensity.

4.2.2 Effective hearing loss in masking noise

Hearing loss and noise masking are equivalent in the sense that both cause a shift in the hearing threshold. Past studies demonstrated that hearing loss can be closely simulated by introducing a steady-state masking noise with a certain spectral shape [113–115].

Table 4.1: Peak intensity density and center frequency of dominant cue for stop consonants (speech intensity normalized to 80 dB SPL).

Utter	D_C (dB SPL/Hz)	CF (Hz)	Utter	D_C (dB SPL/Hz)	CF (Hz)
m115-ta	69.5	4141	f103-da	47.3	5145
f105-ta	70.4	5043	f119-da	71.0	4848
m112-ta	62.7	4238	m118-da	67.0	2727
f108-ta	68.2	5219	m104-da	53.0	2543
m104-ta	42.3	3855	m111-da	46.0	4027
f106-ta	48.1	3602	m115-da	50.2	3320
m111-ka	76.0	1695	m111-ga	52.0	1769
f103-ka	70.8	1641	f119-ga	66.5	1965
m118-ka	70.5	1273	m104-ga	51.6	1449
f105-ka	53.0	1617	f108-ga	51.6	1918
m114-ka	57.9	1301	m112-ga	69.3	1652
f119-ka	47.6	1637	f109-ga	56.9	1699
m118-pa	64.5	859	m118-ba	37.9	527
f103-pa	63.3	844	f119-ba	28.3	441
m114-pa	52.6	820	m107-ba	57.4	426
f106-pa	66.2	1117	f105-ba	41.4	457
m104-pa	59.5	758	m111-ba	57.7	387
f109-pa	60.1	727	f101-ba	45.1	1102

Given the signal-to-noise ratio of the noisy speech stimuli and the audiogram of the hearing-impaired listener, what is the *effective hearing loss*? Fletcher provided a method of adding up the impact of hearing loss and the effect of masking noise [70]. The idea is to treat the hearing loss as the consequence of a masking noise from the inner ear; then the effective hearing loss H_M is the sum of the masking effect from the internal noise and the external noise:

$$H_M = 10\log_{10}[10^{\frac{M}{10}} + 10^{\frac{(H+H_0)}{10}}] - H_0 \tag{4.3}$$

where the first item M in the log function is the masking level introduced by the external noise; H is the hearing loss in dB HL; and H_0 is the standard hearing threshold for normal hearing people. $H + H_0$ can be regarded as the equivalent internal noise.

Next we show how the masking level of external noise M can be calculated from the signal-to-noise ratio. Assuming that the speech stimuli are presented at the most comfortable level (MCL) for the HI listener, the total energy of the noisy speech is equal

to the sum of the clean speech and the masking noise

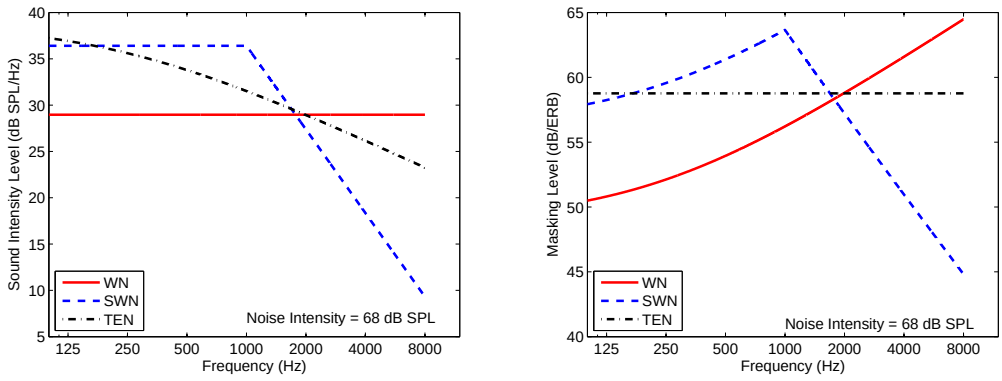
$$10\log_{10}(10^{S/10}+10^{N/10}) = MCL \tag{4.4}$$

where S and N are the intensities of speech and noise, respectively. Given the signal-to-noise ratio $SNR = S - N$ of the noisy speech stimuli, the intensity of masking noise is

$$N = 10\log_{10}(10^{MCL/10}/(1+10^{\frac{SNR}{10}})). \tag{4.5}$$

Notice that the intensity of noise N is the integration of the energy across the frequency. For white noise (WN), the intensity density is relatively simple $D_N = N - 10\log FBW$, where FBW is the full bandwidth of the white noise.

Depending on the spectral shape, the masking noise may have varied intensity density over different frequency ranges. To give a brief comparison between white noise (WN), speech-weighted noise (SWN) and threshold equalized noise (TEN), Fig. 4.1(a) depicts the intensity density D_N of the three types of masking noise when the intensity of noise N is equal to 68 dB SPL. The full bandwidth FBW is 8000 Hz. The masking levels M of the three types of noises are plotted in Fig. 4.1(b). As we see, TEN provides an equal amount of masking over all frequencies, while WN and SWN tend to produce more masking in the high-frequency and low-frequency ranges respectively.



(a) Noise intensity density

(b) Noise masking level

Figure 4.1: Comparison of WN, SWN and TEN at 68 dB SPL.

Letting Δ denote the difference between the intensity density per frequency D_N and

the overall intensity N, the intensity density of the masking noise can be calculated by

$$D_N(f) = N + \Delta(f) \tag{4.6}$$

The values of $\Delta(f)$ for WN, SWN and TEN are listed in Table 4.2.

Table 4.2: Intensity density increase of WN, SWN and TEN at various frequencies.

$\Delta(f)$ (Hz)	125	250	500	1000	2000	3000	4000	6000	8000
WN	-39.0	-39.0	-39.0	-39.0	-39.0	-39.0	-39.0	-39.0	-39.0
SWN	-31.5	-31.5	-31.5	-31.5	-40.6	-45.8	-49.6	-54.9	-58.6
TEN	-31.0	-32.3	-34.1	-36.4	-39.0	-40.6	-41.8	-43.4	-44.7

Given the noise intensity density D_N, the masking level within an auditory filter can be calculated by integrating the noise energy over the critical band

$$M = D_N + 10\log_{10} ERB \tag{4.7}$$

It is well known that the masking effect of noise is much greater for hearing impaired people than for normal hearing people; therefore, we introduced a correction factor $K(f)$ to account for the widening of critical bandwidth. Replacing the masking level M with KM, Eq. (4.3) can now be rewritten as

$$H_M = 10\log_{10}[10^{\frac{KM}{10}} + 10^{\frac{(H+H_0)}{10}}] - H_0 \tag{4.8}$$

Thus the correction factor $K(f)$ can be estimated by

$$\hat{K} = \frac{10}{M}\log_{10}[(10^{\frac{H_M}{10}} - 10^{\frac{H}{10}})10^{\frac{H_0}{10}}] \tag{4.9}$$

where the masking level M, hearing loss H and effective hearing loss H_M are available from the TEN test.

Given the presentation level of noisy speech MCL and the signal-to-noise ratio SNR, we derived the formulas for the calculation of speech cue intensity I_C and the effective hearing loss H_M. A consonant sound is intelligible if $I_C - H_0 > H_M$.

4.2.3 Prediction of recognition score

Assuming that a listener can hear a speech sound if and only if the dominant cue is audible, the information about speech cue audibility can be used to predict the recognition score.

Letting P_c denote the probability of correctness in recognizing a sound, we have

$$P_c = \begin{cases} 1 & \text{dominant cue audible} \\ P_{chance} & \text{dominant cue inaudible} \end{cases} \tag{4.10}$$

where P_{chance} is the probability of reporting the right sound by random guessing.

For the case of multiple talkers, the probability of correctness of a particular nonsense syllable is the average over all talkers. Thus we derived a simple way of estimating the recognition score of stop consonants based on the information of speech cue audibility.

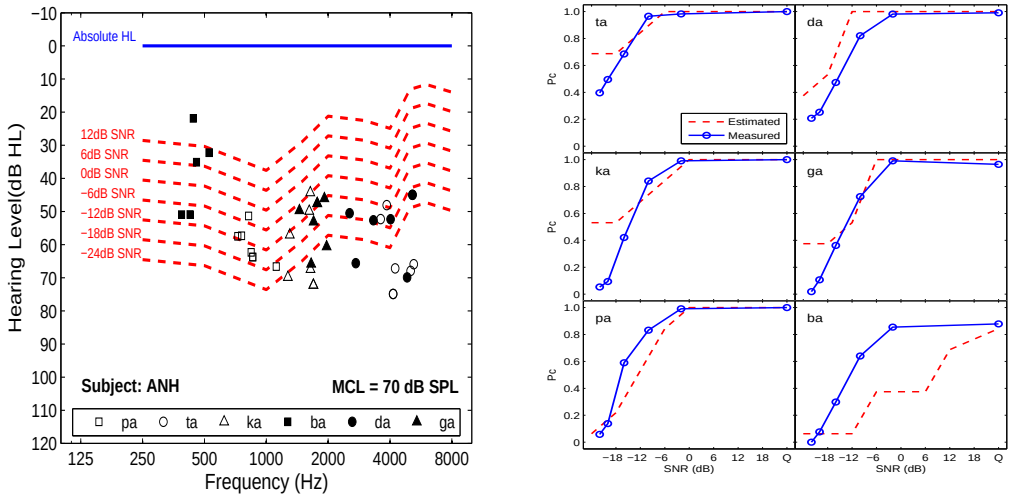
To give an example, Fig. 4.2(a) shows the extended speech banana in SWN for average normal hearing (ANH) people. Using the above method, we convert the information of speech cue audibility into P_c functions and compare them to the real perceptual data measured in [91]. The results are depicted in Fig. 4.2(b). The predicted scores fit the real data closely for five stop consonants except for /ba/, suggesting that the information of speech cue audibility, as predicted by the extended speech banana, is pretty accurate for ANH listeners. Meanwhile it is still uncertain why the extended speech banana under-predicts the audibility of the /ba/ cue. A possible explanation is that the /ba/ cue covers more than one auditory band.

4.3 Methods

A speech perception test (SL07), using 16 nonsense CVs as the stimuli, is employed to collect the confusion patterns. The detail of experiment SL07 is given below.

4.3.1 Subjects

Five subjects (10 ears) with bilateral SNHL participated the study. All subjects spoke fluent English. Tympanometry, DPOAE and MEPA (a wide-band acoustic reflectance



(a) Extended Speech Banana

(b) P_c (Predicted vs. Real)

Figure 4.2: Extended speech banana (left) and the probability of correctness (P_c) (right) for stop consonants. (a) The absolute hearing loss and effective hearing loss in SWN are depicted by the solid and dashed curves. For a given SNR, speech cues above the curve are inaudible. (b) Real perceptual data (solid) versus predicted scores based on the extended speech banana.

system for the diagnosis of middle ear problems) were used to decide the HL listeners' type of hearing loss. The shift of hearing threshold was measured by PTA. TEN test [6, 108, 123] and PTC test [6, 109] were combined for the diagnosis of cochlear dead regions. The first two subjects, AS and DC, were tested with a Matlab version of TEN test created by the author. The other subjects were tested using the TEN CD provided by the inventor from Cambridge University. The speech study and the hearing tests were conducted ear by ear. All subjects were paid for their participation. Approval by the University of Illinois Institutional Review Board was obtained before the experiment.

4.3.2 Speech stimuli

Sixteen nonsense CV: /p, t, k, f, T, s, S, b, d, g, v, D, z, Z, m, n/ + /a/, chosen from the LDC-2005S22 corpus, were used as the test material for the HI listeners. Each CV has only 6 talkers, half male and half female. The speech sounds were sampled at 16,000 Hz. The speech sounds were presented at the listener's most comfortable level (MCL) through an ER-2 earphone. All experiments were conducted in a sound-proof booth.

4.3.3 Conditions

Speech sound were masked at six different signal-to-noise ratios [-12, -6, 0, 6, 12] and quiet conditions using speech-weighted noise.

4.3.4 Procedure

To save the subjects' time and reduce the total length of the speech perception experiment, the speech stimuli were presented to the HI listeners by an adaptive procedure which starts from high SNR and stops at a certain condition when the recognition accuracy falls below a given threshold. 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 felt tired. Subjects were allowed to play each token up to 3 times. A PC-based Matlab program was created for the control of the procedure.

4.4 Results

Two male and three female subjects with bilateral sensorineural hearing loss participated in this study. Table 4.3 provides an overview of the demographic information of the participants. Except for the case of subject MC who has genetic hearing loss, all the subjects report having acquired hearing loss such as presbycusis, noise-induced or other. The hearing loss configuration ranges from mildly flat to severely sloping. In the rest of this section we will summarize the results of the hearing-impaired study case by case.

Table 4.3: Demographic information for the hearing-impaired subjects. AS-L and AS-R represent the left ear and right ear of subject AS, respectively.

Ear ID	Sex	Age	PTA (dB HL)	Hearing Loss Configuration
AS-L	F	81	45	moderate SNHL
AS-R	F	81	46.7	moderate SNHL
DC-L	M	78	21.7	mild ski-slope SNHL
DC-R	M	78	25	mild-to-moderate ski-slope SNHL
BD-L	M	49	38.3	mild flat SNHL
BD-R	M	49	35	mild flat SNHL
MJ-L	F	43	41.7	mild SNHL
MJ-R	F	43	41.7	mild SNHL
MC-L	F	21	58.3	moderate-to-severe SNHL
MC-R	F	21	90	severe SNHL

4.4.1 Subject: AS

Hearing Configuration

The result of pure tone audiometry (Fig. 4.3) shows that subject AS has a bilateral moderate sloping hearing loss with the pure tone average (PTA) values being equal to 40 dB HL and 42 dB HL for the left ear and right ear, respectively. Based on the results of MEPA (middle ear power reflectance) and DPOAE tests (both provided by Mimosa Inc.), subject AS does not have any problem in the middle ear.

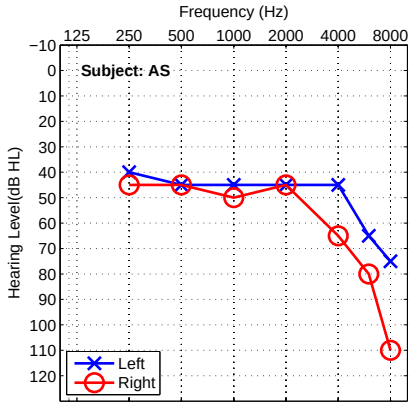
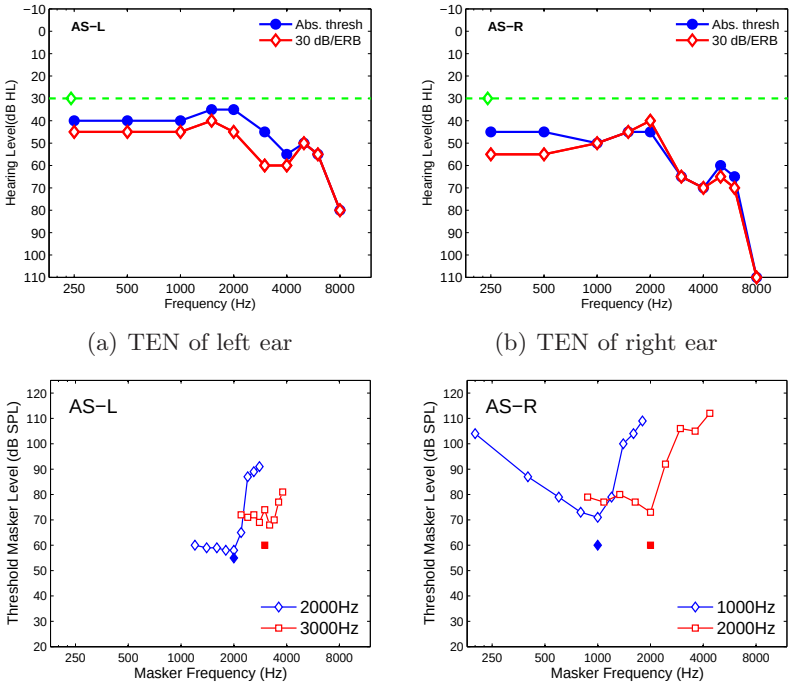


Figure 4.3: Pure tone audiogram of subject AS.

Subject AS may have a big cochlear dead region around 2–3 kHz in the left ear and another cochlear dead region above 8 kHz in the right ear. The TEN test for the left ear (Fig. 4.4(a)) shows a gap of more than 10 dB between the absolute hearing loss (marked with circles) and the effective hearing loss in TEN noise (marked with diamonds) from



(a) TEN of left ear

(b) TEN of right ear

(c) PTC of left ear

(d) PTC of right ear

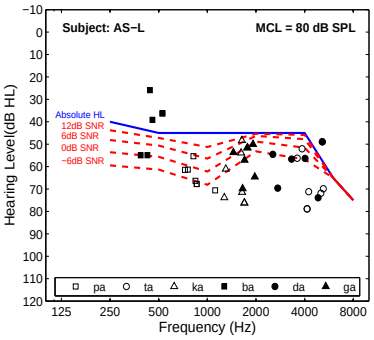
Figure 4.4: Results of TEN and PTC tests. (a) TEN of AS-L: a gap of more than 10 dB between the absolute HL (filled circles) and the TEN-masked HL (open diamonds) suggests a big CDR around 2–3 kHz. (b) TEN of AS-R: no CDR identified. (c) PTC of AS-L: shallow PTC curves at 2 and 3 kHz indicate poor ability of frequency selectivity, CDR possible. (d) PTC of AS-R: the tuning curves are shallow but no tip shifts along the frequency.

2–3 kHz, suggesting a possible cochlear dead region around that frequency range. We then measured the PTC curves at the two frequencies of interest, 2 and 3 kHz, as depicted in Fig. 4.4(c). Both curves have very shallow (close to flat) tips, meaning that the frequency selectivity ability is extremely poor at those frequencies. The situation for the right ear is much better in that the mid-frequency range, which is important for speech perception, has no dead region. The subject may have a cochlear dead region around 8 kHz in the right ear because of the severe hearing loss of 110 dB.

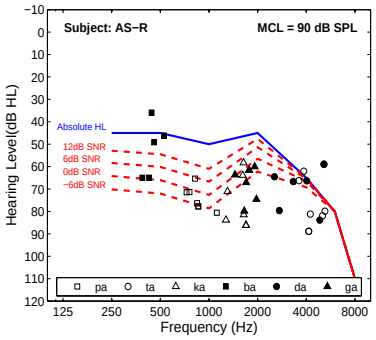
Speech Perception

Results of the speech perception experiment indicate that cochlear dead regions may have a significant impact on the perception of speech sounds. Figure 4.5(c) depicts the recognition scores of subject AS for six stop consonants /pa, ta, ka, ba, da, ga/. Due to the cochlear dead region in the mid-frequency range, where the perceptual cues for /ka/ and /ga/ are located, the subject cannot hear these two sounds completely with her left ear even in quiet conditions. In contrast, her right ear can hear these two sounds with low accuracy, despite the fact that the two ears have close configurations of hearing loss in terms of pure tone audiometry.

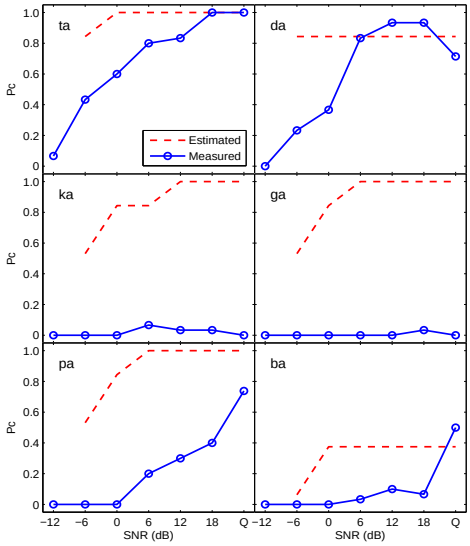
The predictions of the extended speech banana are poorly correlated with the real perceptual data (refer to Fig. 4.5(c) and 4.5(c)), suggesting that pure tone audiometry alone is not a good predictor for speech perception. According to the extended speech banana, subject AS should not have much difficulty hearing the mid-frequency sounds /ka, ga/ in either ear. The real data show the opposite results. The existence of the cochlear dead region and loss of frequency selectivity, which cannot be characterized by the pure-tone audiogram, may account for the failure of the extended speech banana.



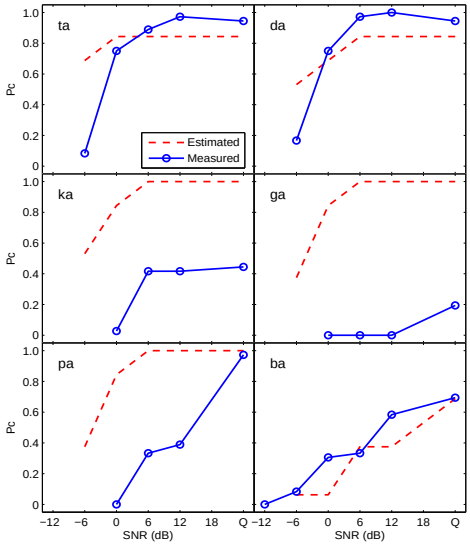
(a) Left ear (AS-L)



(b) Right ear (AS-R)



(c) Left ear (AS-L)



(d) Right ear (AS-R)

Figure 4.5: Extended speech banana (upper panels) and the probability of correctness (P_c) (lower panels) for stop consonants. (a, b) The absolute hearing loss and effective hearing loss in SWN are depicted by solid curve and dashed curves respectively. Speech cues above the curve of (effective) hearing loss are inaudible. (c, d) Real perceptual data (solid) versus estimated scores (dashed) based on the extended speech banana.

4.4.2 Subject: DC

Hearing Configuration

Subject DC has a severe sloping high-frequency hearing loss in both ears (refer to Fig. 4.6). The pure tone average (PTA) for the left ear and right ear are 21.7 and 25 dB HL, respectively. Based on the results of MEPA (middle ear power reflectance) and DPOAE tests (both provided by Mimosa Inc.), subject DC does not have any problems in the middle ear.

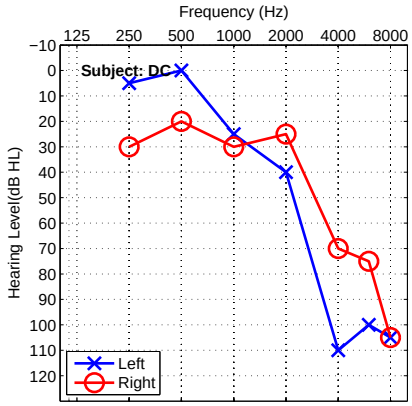
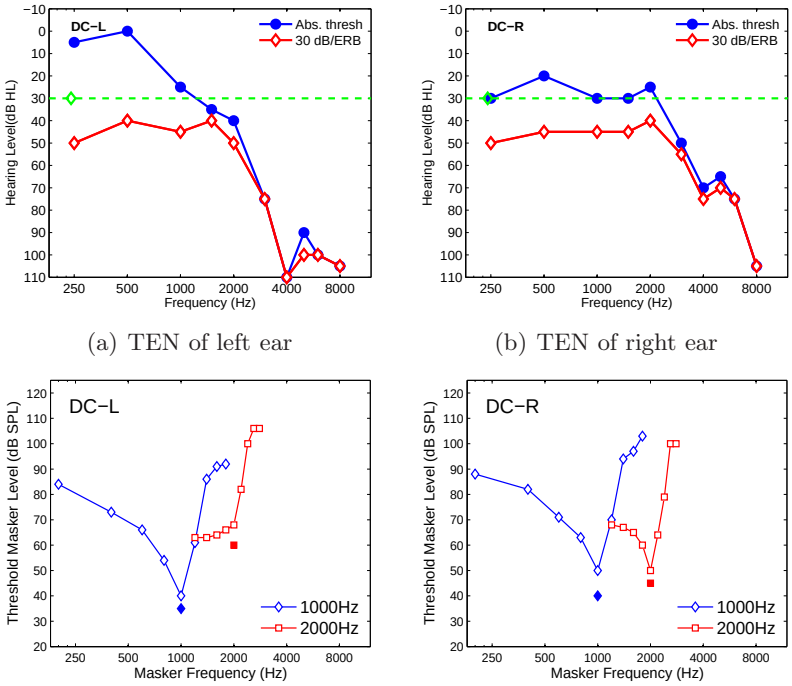


Figure 4.6: Pure tone audiogram of subject DC.

The results of the TEN and PTC tests suggest that the left ear may have a cochlear dead region above 4 kHz, associated with the severe hearing loss (Fig. 4.7(a)), and a small cochlear dead region around 2kHz, as revealed by the 10 dB gap between the absolute hearing loss and the TEN masked hearing loss (Fig. 4.7(a)) and the tip shift of the 2 kHz tuning curve (Fig. 4.7(c)). Despite the large increase of masked hearing loss below 1 kHz, the possibility of a cochlear dead region is negligible because the absolute hearing loss is close to normal hearing. For the right ear, the TEN test (Fig. 4.7(b)) and the PTC test (Fig. 4.7(d)) generate conflicting results, which makes them difficult to interpret. However, considering the fact that the two PTC curves have normal shape and the absolute hearing loss is lower than 30 dB below 2 kHz, it seems that subject DC has no cochlear dead region in the right ear below 8 kHz.



(a) TEN of left ear

(b) TEN of right ear

(c) PTC of left ear

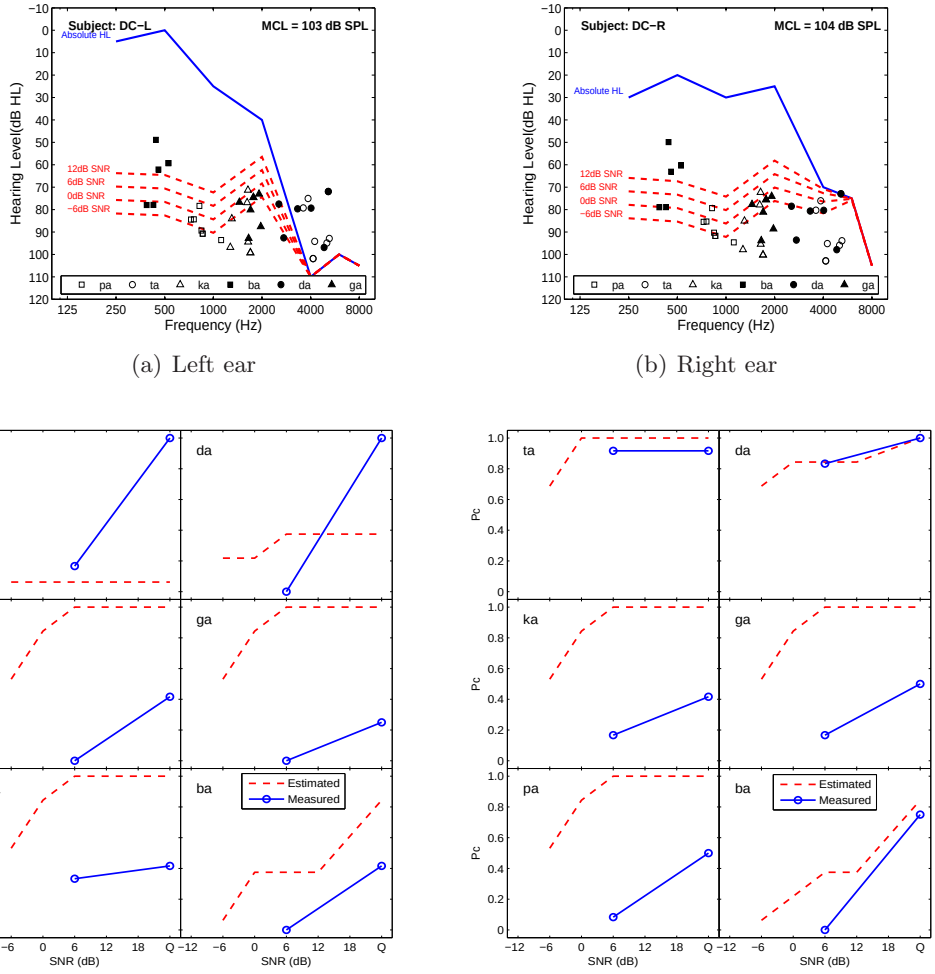
(d) PTC of right ear

Figure 4.7: Results of TEN and PTC tests. (a) TEN of DC-L: a gap of 10 dB between the absolute HL (filled circles) and the TEN-masked HL (open diamonds) suggests a tiny CDR around 2 kHz. (b) TEN of DC-R: no CDR below 2 kHz due to slight hearing loss. (c) PTC of DC-L: tip shift at 2 kHz signifies a possible CDR. (d) PTC of DC-R: tuning curves are normal.

Speech Perception

Due to the time limit imposed by the participant, we only tested subject DC at 6 dB SNR and quiet conditions. Despite the subject having fairly good low-frequency residual hearing and severe high-frequency loss above 4 kHz, both the left and right ears show high performance, in quiet, on the two high-frequency sounds /ta, da/ and poor performance on the two low-frequency sounds /pa, ba/ (Figure 4.8(c)), suggesting that the subject may learn to use a new set of perceptual cues for the two sounds as a consequence of long-term high-frequency hearing loss. The small cochlear dead region around 2 kHz in the left ear, if it exists, has little effect on the perception of mid-frequency sounds. Subject DC can still hear /ka, ga/ under quiet conditions.

Figure 4.8(c) and 4.8(c) compare the prediction of the extended speech banana to the actual perceptual data. Again, the estimated recognition scores are poorly correlated to the actual perceptual data. For example, it fails to explain why subject DC can hear the high-frequency sounds /ta, da/ better than the mid-frequency sounds /ka, ga/. It is conjectured that the severely unbalanced hearing loss along the frequency has changed the mapping of perceptual cues in the auditory cortex; in other words, the HI listener may have tuned to a new set of perceptual cues different from those used by normal hearing listeners.



(a) Left ear

(b) Right ear

(c) Left ear (DC-L)

(d) Right ear (DC-R)

Figure 4.8: Extended speech banana (upper panels) and the probability of correctness (P_c) (lower panels) for stop consonants. (a, b) The absolute hearing loss and effective hearing loss in SWN are depicted by the solid curve and dashed curves respectively. Speech cues above the curve of (effective) hearing loss are inaudible. (c, d) Real perceptual data (solid) versus estimated scores based on extended speech banana.

4.4.3 Subject: BD

Hearing Configuration

Subject BD has mild flat sensorineural hearing loss in both ears (Fig. 4.9) with the right ear (PTA = 35 dB HL) being slightly better than the left ear (PTA = 38.3 dB HL). Both ears have normal middle ear functions based on the results of tympanometry (type A) and DPOAE tests.

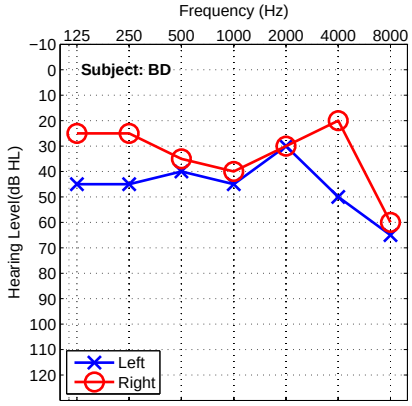
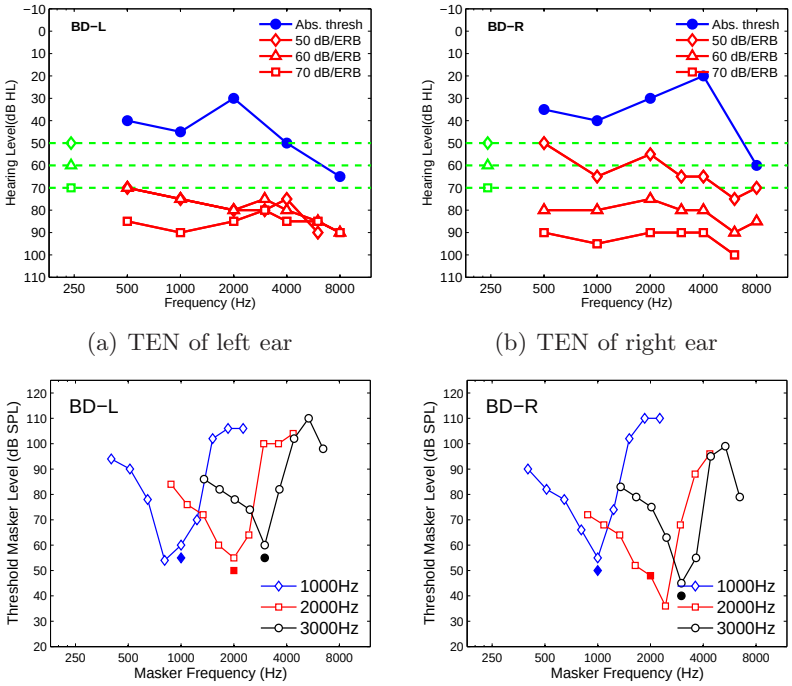


Figure 4.9: Pure tone audiogram of Subject BD.

Based on the PTA, TEN and PTC results, it is difficult to tell whether subject BD has any cochlear dead regions. The TEN test (Fig. 4.10(b)) and PTC test (Fig. 4.10(d)) suggest that BD-R may have a cochlear dead region in the right ear around 2 kHz. There is a big gap between the absolute hearing loss and TEN-masked hearing loss over all frequencies, but only the one at 2 kHz is verified by the PTC curve (Fig. 4.10(b)), which has a tip shifted toward upper frequency. Similar results are observed for BD-L, for which the TEN test (Fig. 4.10(a)) and the PTC test (Fig. 4.10(c)) are consistent only at 1 kHz, other than that the PTC curves have normal shape. On the other hand, the PTA test shows a bilateral mild flat hearing loss lower than 40 dB, and it is hard to believe that a subject who demonstrates near perfect functionality of frequency selectivity, as implied by the close to normal shape of the PTC curves, would have any cochlear dead regions at all.



(a) TEN of left ear

(b) TEN of right ear

(c) PTC of left ear

(d) PTC of right ear

Figure 4.10: Results of TEN and PTC tests. (a) TEN of BD-L: a gap of more than 10 dB between the absolute HL (filled circles) and the TEN-masked HL (open diamonds) suggests a possible CDR below 1 kHz. (b) TEN of BD-R: TEN test fails. (c) PTC of BD-L: tip shift at 1 kHz suggests a possible CDR. (d) PTC of BD-R: tip shift suggests a possible CDR at 2 kHz.

Speech Perception

Figure 4.11(c) depicts the perceptual results of BD-L on the stop consonants. The subject can hear most of the stop consonants with an accuracy of over 90% at 6 dB SNR. Compared to the left ear, his right ear has similar and slightly better recognition scores for all the stop consonants, which is consistent with the symmetrical hearing loss configuration. The results of speech perception do not confirm the existence of a cochlear dead region. It is uncertain why the recognition score of /ga/ increases when the *SNR* decreases from $+\infty$ to 12 for the left and right ears.

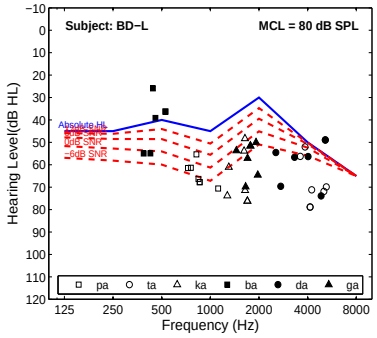
The prediction of the extended speech banana closely matches the actual perceptual score for the stop consonants. According to the extended speech banana (Figure 4.11(a)), subject BD should have little difficulty with stop consonants in quiet or slightly noisy conditions because most of the speech cues are still audible at 6 dB SNR. This is confirmed by the comparison between the real recognition score and the predicted scores based on the extended speech banana, as depicted by Fig. 4.11(c). Except for /ba/, for which the burst is not necessarily the dominant cue, most other stop consonants show small discrepancy between the two curves. Similar results are observed in the right ear, suggesting that the principle of audibility does apply to hearing-impaired people when the subject has mild hearing loss.

4.4.4 Subject: MC

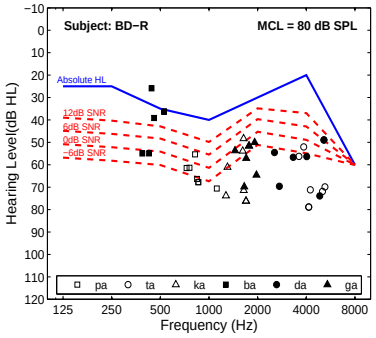
Hearing Configuration

Pure tone audiometry (Fig. 4.12) shows that subject MC has moderate hearing loss in the left ear (PTA average = 58 dB HL) and a severe hearing loss in the right ear (PTA average = 90 dB HL). Based on the results of tympanometry and DPOAE, the subject has no middle ear problems in either side.

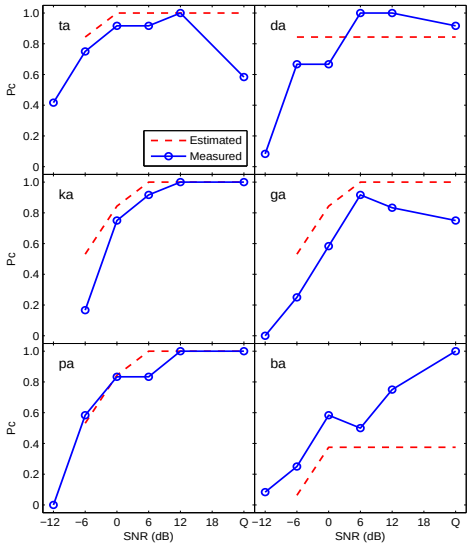
Results of TEN tests (Fig. 4.13(a) 4.13(a)) indicate no sign of a cochlear dead region in either ear; therefore, we skipped the time-consuming PTC tests for the subject.



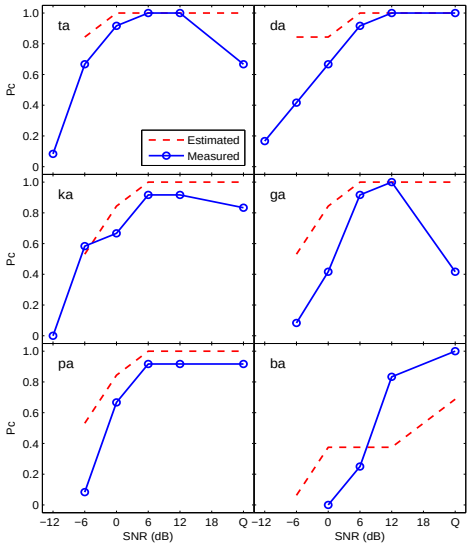
(a) Left ear



(b) Right ear



(c) Left ear (BD-L)



(d) Right ear (BD-R)

Figure 4.11: Extended speech banana (upper panels) and the probability of correctness (P_c) (lower panels) for stop consonants. (a, b) The absolute hearing loss and effective hearing loss in SWN are depicted by the solid curve and dashed curves respectively. Speech cues above the curve of (effective) hearing loss are inaudible. (c, d) Real perceptual data (solid) versus estimated scores based on extended speech banana.

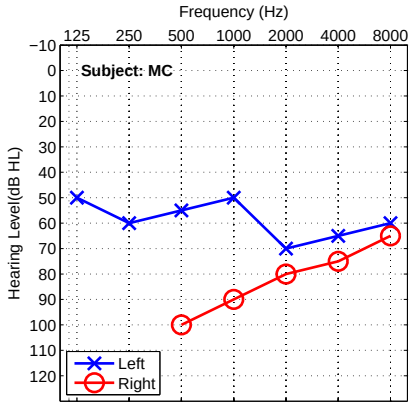
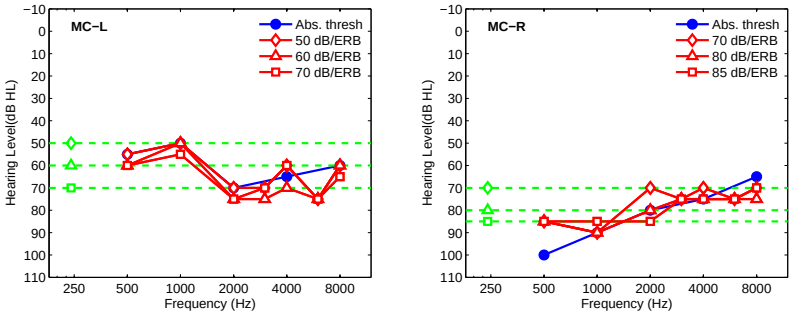


Figure 4.12: Pure tone audiogram of subject MC.



(a) TEN of left ear

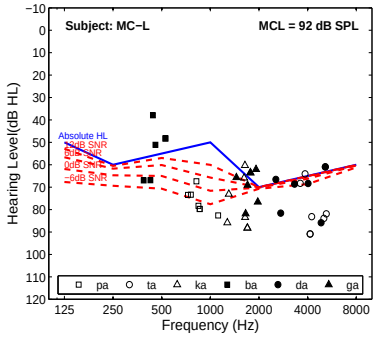
(b) TEN of right ear

Figure 4.13: Diagnosis of cochlear dead regions. (a) TEN of MC-L: no large gap between the absolute HL (filled circles) and the TEN-masked HL (open symbols) suggests no CDR. (b) TEN of MC-R: no CDR detected.

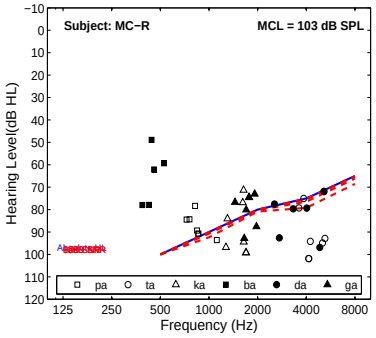
Speech Perception

MC-R has a rare case of severe-to-moderate up-sloping hearing loss; the associated results of speech perception are depicted in Fig. 4.14(d). Because of the severe hearing loss in the low frequency, subject MC cannot hear /ba/ completely ($Pc = 0$) in the right ear. The recognition accuracy increases as the center frequency of the speech cue increases; i.e., the two low-frequency sounds /pa, ba/ have lower scores than the two mid-frequency sounds /ka, ga/, which again are lower than the two high-frequency sounds /ta, da/. All the recognition scores saturate early at a P_c much lower than 1. The results of speech perception match the configuration of hearing loss. The predicted scores of the extended speech banana (Fig. 4.14(b)) fit the real perceptual data fairly well when the signal-to-noise ratio is greater than 6 dB SNR.

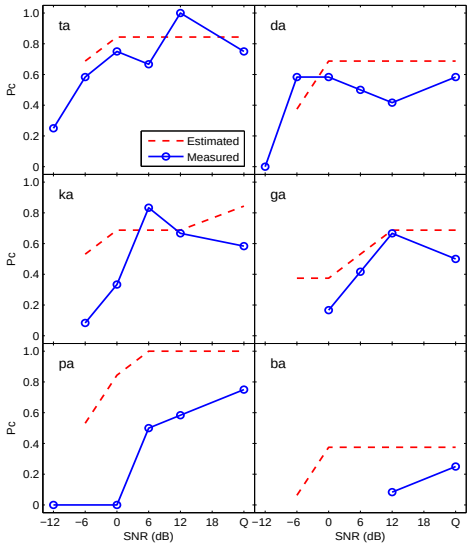
MC-L has moderate flat hearing loss. The extended speech banana (Fig. 4.14(a)) predicts that the subject cannot hear /ta, da/ produced by some of the talkers even in quiet conditions, and the existence of masking noise, no matter whether *SNR* equals 6, 0, or -6 dB, makes little difference for the perception of the two high-frequency sounds. The actual perceptual data (Fig. 4.14(d)) confirms the prediction of the extended speech banana, which also applies to the two mid-frequency sounds /ka, ga/, but it fails for the two low-frequency sounds /pa, ba/. It is uncertain whether the low-frequency hearing loss in the right ear has any impact on the auditory cortex and therefore interferes with the speech perception in the other ear.



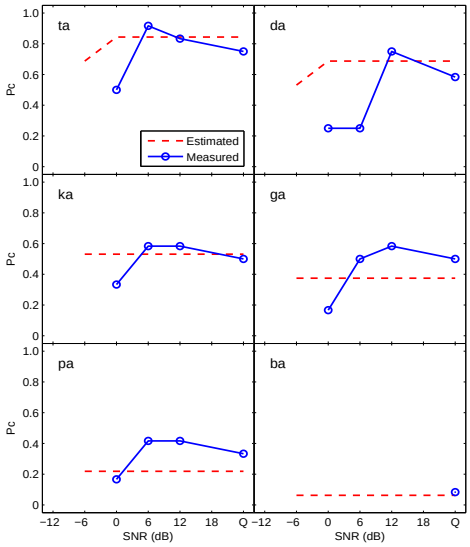
(a) Left ear



(b) Right ear



(c) Left ear (MC-L)



(d) Right ear (MC-R)

Figure 4.14: Extended speech banana (upper panels) and the probability of correctness (P_c) (lower panels) for stop consonants. (a, b) The absolute hearing loss and effective hearing loss in SWN are depicted by the solid curve and dashed curves respectively. Speech cues above the curve of (effective) hearing loss are inaudible. (c, d) Real perceptual data (solid) versus estimated scores based on extended speech banana.

4.4.5 Subject: MJ

Hearing Configuration

Pure tone audiometry (Fig. 4.15) shows that subject MJ has a symmetric bilateral mild hearing loss that is identical for the left and right ears. The PTA average is 41.7 dB HL. Results of MEPA (middle ear power reflectance) and DPOAE tests (both provided by Mimosa Inc.) indicate that subject MJ has normal middle ear functioning. The subject does not participate in the TEN and PTC tests. Considering the fact that the hearing loss is a mild, we assume that subject MJ does not have any cochlear dead regions.

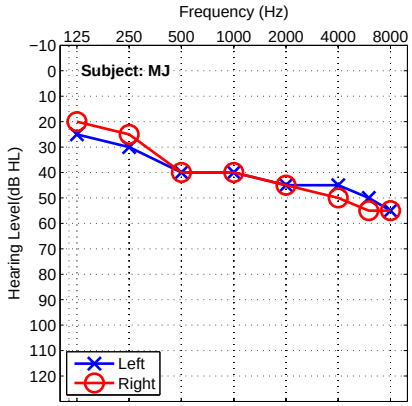
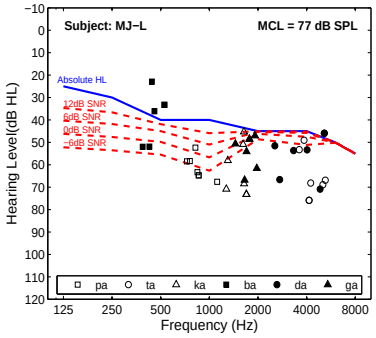


Figure 4.15: Pure tone audiogram of subject MJ.

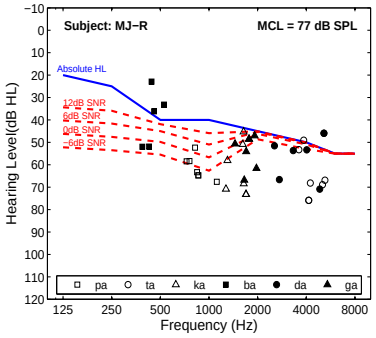
Speech Perception

Results of the speech perception test indicate that subject MJ has no significant problem with any stop consonants. Figure 4.16(c) depicts the recognition scores for her left ear. The subject can still achieve a score of nearly perfect accuracy (close to 100%) for most sounds at 6 dB SNR. Her right ear mirrors the performance of the left ear, consistent with the identical configuration of hearing loss between the two ears.

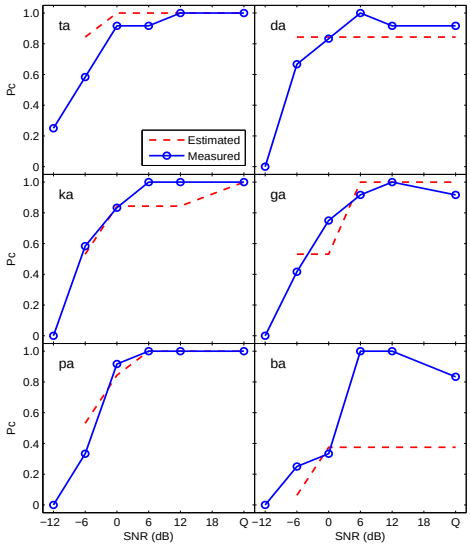
Similar to the other case of mild hearing loss (subject BD), the prediction of the extended speech banana fits accurately to the actual perceptual data. Figure 4.16(c) compares the estimated P_c based on the extended speech banana and the real P_c from experiment MN64, which also uses SWN as the masking noise. The two curves are close to each other. Similar results (Fig. 4.16(d)) are observed for the other ear, meaning that audibility of speech cue fully accounts for the intelligibility of consonant sounds.



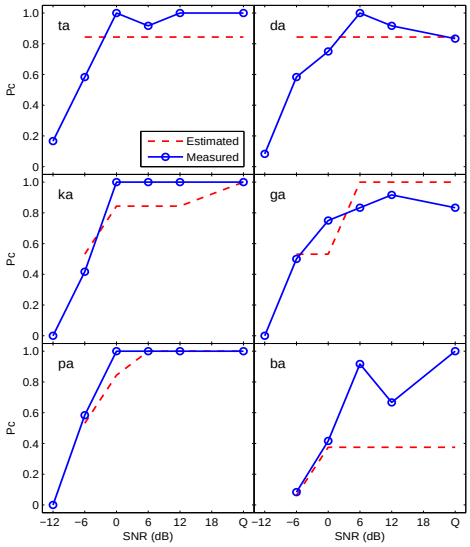
(a) Left ear



(b) Right ear



(c) Left ear (MJ-L)



(d) Right ear (MJ-R)

Figure 4.16: Extended speech banana (upper panels) and the probability of correctness (P_c) (lower panels) for stop consonants. (a, b) The absolute hearing loss and effective hearing loss in SWN are depicted by the solid curve and dashed curves respectively. Speech cues above the curve of (effective) hearing loss are inaudible. (c, d) Real perceptual data (solid) versus estimated scores based on extended speech banana.

4.5 Discussion and Conclusion

The goal of this study is to investigate the impact of sensorineural hearing loss on the perception of consonant sounds. Previous studies [113–115] concluded that audibility, as characterized by the pure-tone audiogram, is the only factor that accounts for the disability in speech perception. The conclusion is reached based on the comparison of average perceptual score between the hearing-impaired listeners and normal hearing listeners with equal amount of noise-simulated hearing loss. The detailed information about individual sounds is ignored.

Based on the identified perceptual cues and the theory of simultaneous masking, the traditional idea of the speech banana is developed into a quantitative tool, named the extended speech banana, for the evaluation of speech intelligibility. Given the intensity of the speech sound and the pure-tone audiogram of a particular HI listener, it can predict the audibility of speech cues for the subject under various noise conditions. If the prediction matches the actual perceptual data, it must be true that audibility, as characterized by the shift in pure-tone hearing threshold, is the only factor critical for speech perception. If the two mismatch, there might be some other factors that override the factor of audibility.

Five hearing-impaired subjects with bilateral sensorineural hearing loss participated the study. The prediction of the extended speech banana closely fits the perceptual data of two subjects (BD and MJ) with mild flat hearing loss, and matches fairly well for a subject (MC) with moderate hearing loss. It fails on two other subjects, one (AS) with a big cochlear dead region in one ear, the other (DC) with a steep sloping high-frequency loss. For both cases the PTC tests show abnormal shallow tuning curves, suggesting that frequency selectivity is an important factor for speech perception. Besides, plasticity may also play a role in speech perception for the long-term hearing-impaired listeners. For example, the left ear of subject DC can hear /ta/ and /da/ with a probability of 100% under quiet conditions despite the fact that most of the speech cues are inaudible. The subject may learn to use a set of minor perceptual cues that are ignored by normal hearing listeners because of the existence of the dominant cue. The results generally support our hypothesis that the HI listeners have problems understanding speech because they cannot hear certain sounds whose

events are missing due to their hearing impairment or the masking effect introduced by the noise. How the speech cues are affected by the cochlear dead regions or flattening auditory filters requires further study.

CHAPTER 5

MANIPULATION OF CONSONANT SOUNDS IN NATURAL SPEECH

This chapter explores the potential use of knowledge about perceptual cues of consonant sounds in speech processing. Analysis of nonsense consonant-vowel syllables from the LDC database reveals that natural speech, especially stop consonants, often contains *conflicting cues* that are characteristic of confusable sounds. Through the manipulation of these acoustic cues, one phone (a consonant or vowel sound) can be morphed into another; a weak sound, easily masked by noise, can be converted into a strong one. Results of speech perception experiments on feature-enhanced /ka/ and /ga/ show that amplification of speech cues significantly improves the performances in noise for both normal and hearing-impaired listeners.

5.1 Introduction

After more than 50 years of study, many speech processing techniques, such as communication, synthesis, noise reduction, as well as automatic speech recognition (ASR), have reached a plateau in performance. A widely held view is that bio-inspired speech processing schemes that take advantage of prior knowledge about human speech perception (HSP) could potentially lead to better solutions for those speech applications [124–126]. Take speech recognition, for example: the performance of the state-of-the-art ASR systems is still far below that of humans, despite more than 40 years of research effort [125]. The phone classification accuracy in ASR systems varies from 82% in quiet [127] to chance performance at 0 dB SNR. Human performance is quite different. The average phone classification accuracy in quiet is near 98-98.5% (1.5-2% error) [95, 96]. The SNR required for chance performance is below -20 dB SNR [91]. For many sounds, the performance in humans is unchanged from quiet to 0 dB SNR [9].

A fundamental problem of human speech perception is: *How is the speech coded*

in the auditory system? In order to find out the basic spectrum patterns for different speech sounds, Liberman and his colleagues built a machine called *pattern playback* that generates artificial speech from a spectrogram and conducted a series of psychoacoustic studies on the perception of synthetic stop consonants [42, 43]. Later, this method was applied in the search for acoustic correlates for stops [44, 128], fricatives [45, 46], nasals [47–49], and distinctive or articulatory features ([16,25,97]). To understand how speech sounds are represented in the auditory system, a small number of researchers have studied the recordings of single auditory neurons in response to speech and speech-like stimuli [19–21, 129]. Since it is unethical to record in the human auditory nerve, and it is difficult to do extensive speech psychophysics in non-human animals, it was impossible to correlate those neurophysiological studies with psychophysical data. We have skirted this problem by creating a computational model of speech reception, called the AI-gram, which crudely predicts the audibility of speech components to the central auditory system [94,100]. In [110], this method is extended to include a psychophysical test, named 3D Deep Search (3DDS), to measure the contribution of different time-frequency components to speech perception.

This chapter explores the potential uses of our new knowledge of perceptual cues in speech processing. It is frequently said that speech contains redundant cues. To the contrary, it was discovered that natural speech often contains conflicting cues that are representative of competing speech sounds. Through the manipulation of these cues, usually a tiny spot on the AI-gram, we can convert one phone into another phone, or turn a weak speech sound into a strong one, so that hearing impaired listeners can detect them with higher probability in noisy situations.

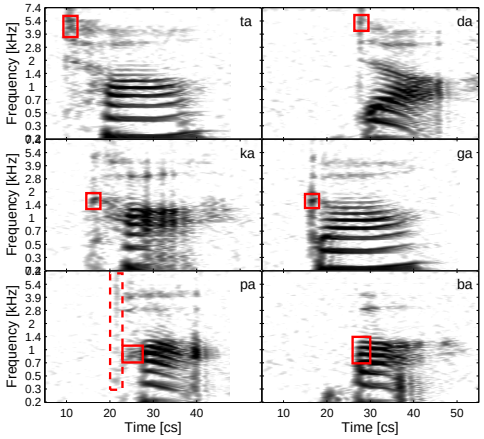
5.2 Perceptual Cues of Consonant Sounds

In natural speech, due to the physical constraints on the articulators (mouth, tongue, lips, etc.), it is widely accepted that their ideal position is often compromised due to neighboring sounds (e.g., a V on a C). Namely, speech cues of successive {C,V} units frequently interact, which is an effect called *coarticulation* [34]. Since coarticulation does not extend beyond the syllable, it is common to separate continuous speech into syllable segments, such as CV or CVC [64].

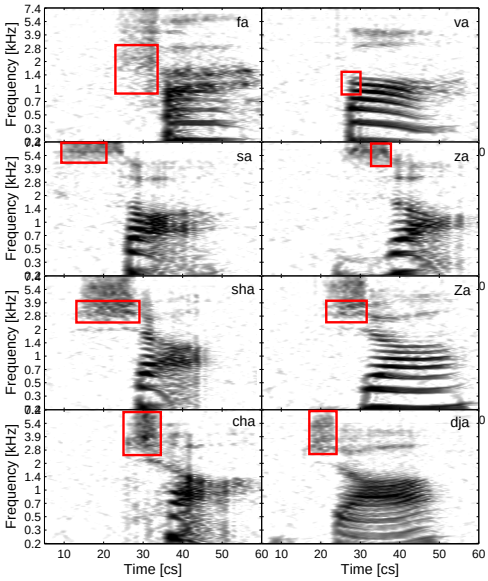
Using the aforementioned 3DDS method, we have identified the perceptual cues of initial consonants preceding vowel /a/ [110, 120, 121].

5.2.1 Overview of consonant cues

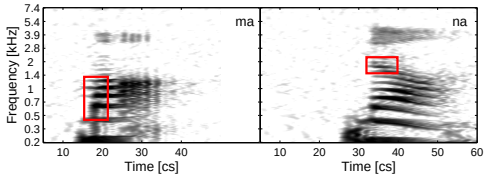
Figure 5.1 depicts the AI-grams of 16 consonants preceding vowel /a/ with the dominant perceptual cues highlighted by the rectangular frames. The stop consonants /p, t, k, b, d, g/ are characterized by a compact burst of short duration (less than 15 ms) caused by the sudden release of pressure in the oral cavity. Within the same group, the stop consonants distinguish themselves by the center frequency of the burst; specifically, /ta/ and /da/ are labeled by a high-frequency burst above 4 kHz, /ka/ and /ga/ are defined by a mid-frequency burst from 1.4 to 2 kHz, and /pa/ and /ba/ are represented by a soft wide-band click, which often degenerates into a low-frequency burst from 0.7 to 1 kHz due to the masking effect of surrounding noise. The voiced and unvoiced stops differ mainly in the duration of the gap between the burst and the start of sonorance. The fricatives /f, s, ʃ, ʧ, v, z, ʒ, ʤ/ are characterized by a salient noise-like cue caused by the turbulent air flow through lips and teeth. Duration and bandwidth are the two key parameters for the discrimination of these sounds. Specifically, the /fa/ cue is within 1-2.8 kHz and lasts for about 80 ms; the /sa/ cue falls within 4-8 kHz and lasts for about 160 ms; /ʃa/ is also labeled by a cue of long duration, but it has a lower frequency (2-4 kHz); and the /ʧa/ cue ranges from 2 to 8 kHz and lasts for more than 100 ms. These results are summarized from [121]. The voiced fricatives have similar patterns of perceptual cues, except that the durations are considerably shorter than their unvoiced counterparts. The two nasals /m/ and /n/ share a common feature of nasal murmur at low frequency and differ from each other in the mid/low-frequency (below 2.4 kHz). These consonant events have been found to be consistent across different talkers. Similar data for the two other vowels /i/ and /u/ is currently being analyzed. In running speech, the acoustic cues are expected to change, depending on the preceding and following vowels, or other factors [42]. This variation is governed by the physical principles of speech production [41, 130].



(a) Stops: /ta, ka, pa, da, ga, ba/



(b) Fricatives: /fa, sa, ʃa, ʧa, va, za, ʒa,ʤa/



(c) Nasals: /ma, na/

Figure 5.1: AI-grams for the 16 Miller-Nicely consonants at 12 dB SNR in white noise: (a) stops, (b) fricatives and (c) nasals. All sounds are pronounced by a female talker f103 except for /fa/, which is produced by talker f101. A rectangular frame highlights the perceptual cue that distinguishes each sound from its competing sounds, as determined by the 3DDS procedure [105, 120]. A dashed frame means that the perceptual cue is often masked by noise. The conflicting cues are labeled by ellipses. These plots form a baseline starting point for speech modifications of the boxed regions.

5.2.2 Conflicting cues

Due to the physical limitations of the human speech articulation system, it seems to be difficult to produce ideal speech sounds, as generated by a speech synthesizer, and it is interesting to see that many natural sounds contain conflicting cues that are characteristic of competing sounds. Analysis of the Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus" from University of Pennsylvania indicates that most stop consonants contain conflicting cues that may lead to confusions in speech perception under adverse circumstances. As an example, the /ka/ from talker f103 is shown in Fig. 5.1(a). The talker (f103) intended to make a /ka/ sound, and the listeners all reported hearing /ka/ 100% of the time at 0 dB in both white noise (WN) and speech-weighted noise (SWN), and even 98% of the time at -10 dB in SWN. Yet, the produced speech contains a high-frequency burst around 5 kHz and a low-frequency burst from 0.4 to 0.7 kHz (highlighted by ellipses), indicative of /ta/ and /pa/ productions, respectively. When such conflicting cues are removed, the speech is perceptually indistinguishable. The listeners report a robust /ka/ because the mid-frequency /ka/ burst (highlighted by the box) perceptually dominates the interfering cues. Another example is /ga/. In addition to the typical /ga/ burst in the mid-frequency (highlighted by the box), it also contains a high-frequency burst above 4 kHz and a low-frequency burst below 1 kHz (labelled by ellipses) that could lead to the perception of /da/ and /ba/, respectively. The same situation applies to /ta/, /da/ and /pa/.

Conflicting cues also exist for fricative consonants. For example, the fricative part of /ʃa/ also contains a /sa/ cue at high frequency above 4 kHz (labelled by an ellipse). Within the fricative part of /sa/ there is also the perceptual cue of /za/. Apart from that, /sa, ʃa, ʧa, za, ʒa/ all have a high frequency burst above the head of the F3 transition (labelled by ellipses) that could lead to the perception of /ða/.

All the example speech sounds, displayed in Fig. 5.1, are produced by a female talker f103, other than /fa/, which is produced by talker f101. According to our speech perceptual data, talker f103 is ranked as one of the best talkers in the LDC database. The problem of producing conflicting cues in a single speech sound is a common observation across all talkers. Because of the conflicting cues, the percept of the sound will predicably change, if the dominant cue is removed or masked. Here we discuss the manipulation

of speech cues in natural speech to: (1) show how the percept of consonant sounds can be controlled, and (2) explore the potential use of perceptual cues in speech processing applications.

5.3 Manipulation of Speech Cues

Speech perception is a complex process where the integration of events is governed by high-level language, such as lexical, morphological, syntactic, and semantic context. In order to understand this process of event integration, it is necessary to start from nonsense syllables, for which the high-level constraints on speech perception are maximally controlled [82]. For this reason, we first look at the manipulation of initial consonants as they occur in isolated nonsense CV syllables. Following that, we show that the speech cues may be modified in isolated meaningful syllables (words) and sentences. The examples discussed in this report can be found at http://hear.ai.uiuc.edu/wiki/Files/VideoDemos. For example, the sample "ka→ka→ta→pa" in Fig. 5.2 is listed as "ka2ka2ta2pa" on the web site.

Our speech modification procedure begins by analyzing the speech sounds using the short-time Fourier transform (STFT). The boxed regions of Fig. 5.1 are then modified, as described below. Finally, the modified speech is returned to the time domain via an overlap-add synthesis [102].

5.3.1 Speech analysis and synthesis

Letting $x[n]$ denote the sampled speech signal at sample times n. For analysis, the original signal $x[n]$ is divided into N point overlapping frames $x[m,n] \equiv w(n)x[mR-n]$ of 20 ms duration with a step size $R \equiv N/4$ samples of 5 ms total duration. A Kaiser window $w[n]$ having -91 dB attenuation (i.e., first side lobe is 91 dB smaller than the main lobe) is used. Note that the speech is time-reversed and shifted across the fixed window prior to being Fourier transformed:

$$X[m,k] = \sum_{n=0}^{N-1} x[m,n]e^{-j2\pi kn/N}. \tag{5.1}$$

The resulting STFT $X[m,k]$ is a two-dimensional complex signal matrix, indexed in time m and frequency k.

The region of a speech cue is modified by multiplying $X[m,k]$ with a two-dimensional mask $M[m,k]$ that specifies the gain g within the feature area. Specifically, $g = 0$ is feature removal, a gain $0 < g < 1$ corresponds to a feature attenuation, and a gain $g > 1$ is feature enhancement, resulting in the modified speech spectrum

$$Y[m,k] = X[m,k] \cdot M[m,k]. \tag{5.2}$$

The gain is expressed in dB as $G = 20 * \log_{10}(g)$ dB. Following modifications, the single frame signal can be recovered by applying an inverse Fourier transform

$$y[m,n] = \frac{1}{N} \sum_{k=0}^{N-1} Y[m,k] e^{j2\pi kn/N} \tag{5.3}$$

followed by the overlap add (OLA) synthesis, resulting in the modified speech signal $\hat{x}[n]$

$$\hat{x}[n] = \sum_{m=-M_0}^{0} y[mR,n] \tag{5.4}$$

over all past samples [102]. In practice this series truncates on the lower side (the modifications have finite memory), as determined by $M_0 = N/R$, which is typically taken to be 4. Due to the modifications, zero-padding of the windowed speech is necessary.

5.3.2 Nonsense syllable

Plosives

To demonstrate that the unvoiced stop consonants /pa/, /ka/ and /ta/ are sensitive to the conflicting cues, we selected /ka/ from talker f103 as an example. Using the signal processing method described in the previous section, we have modified the speech by varying the relative levels of three speech cues (highlighted by the three blocks in Fig. 5.2). When the mid-frequency /ka/ burst in block 1 is removed (lower-left panel of Fig. 5.2), the percept of /ka/ is dramatically changed and listeners report either /pa/

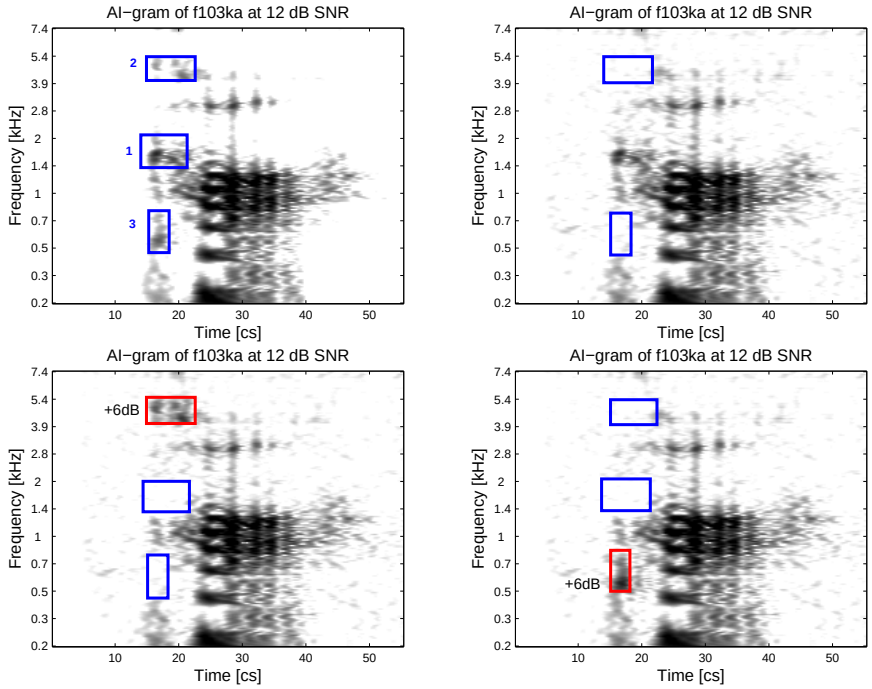


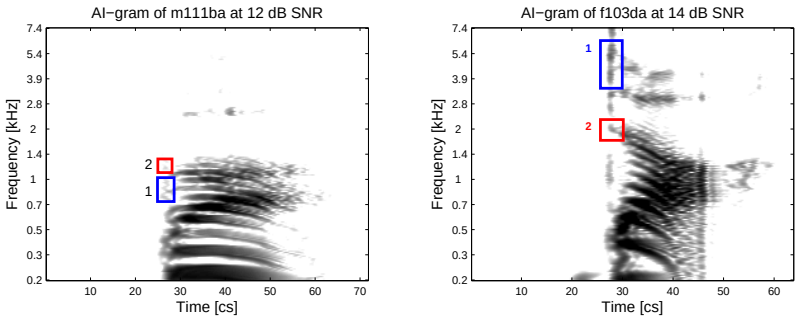
Figure 5.2: Three-way manipulation of unvoiced stop consonant /ka/: In the upper-left, the AI-gram shows the original /ka/ from talker f103 at 12 dB SNR. When the two conflicting cues (blocks 2 and 3) are removed (upper-right panel), the sound is heard as unmodified. When block 1, containing the /k/ cue, is removed (lower-left) and the /t/ cue (block 2) is enhanced by 6 dB, a /t/ is robustly reported. Finally, when both the /k/ and /t/ cues are removed (blocks 1 and 2) (lower-right), /pa/ is robustly reported. (Example: "ka→ka→ta→pa".)

or /ta/. This ambiguous situation is called *priming*, and defined as the auditory illusion where prior expectation of the perceived sound affects the sound reported. When both short bursts for /ta/ and /ka/ (blocks 1, 2) are removed, the sound is robustly perceived as /pa/. Boosting the low-frequency burst within 0.5 and 0.7 kHz (block 3) strengthens the initial aspiration and makes the sound a clearly articulated /pa/ (lower-right panel of Fig. 5.2).

We conjecture that the presence of the 1.4 kHz burst both triggers the /ka/ report and renders the /ta/ and /pa/ bursts either inaudible, via the upward spread of masking (USM, defined as happening when a low frequency sound reduces the magnitude of a higher frequency sound), or irrelevant, via some neural signal processing mechanism. The existence of the USM effect makes high frequency sounds less significant when present with certain low frequency sounds. The auditory system, having learned this, could be programmed to ignore these higher frequency sounds under these uncertain conditions.

An important implication of this example (Fig. 5.2) is that the F2 transition for /ka/ is unnecessary for the discrimination of unvoiced stop consonants, contradictory to a widely accepted argument that the F2 transition is critical for the recognition of stop consonants [43, 97].

The group of voiced stop consonants /ba, da, ga/ and the unvoiced stop consonants /pa, ta, ga/ have similar feature patterns with the main difference being the delay between the *voicing* (i.e., the burst release) and the start of the sonorant portion of the speech sound. We shall next show how the voiced stops /ba, da, ga/ can be modified, again through speech cue manipulations. Figure 5.3(a) depicts the AI-gram of /ba/ from talker m111 at 12 dB SNR of white noise, which was perceived robustly by the listeners as a /ba/. After removing the perceptual cue of /ba/ (block 1) and boosting the mid-frequency burst (block 2) by a factor of 4 (12 dB), it turns into a noise-robust /ga/. Figure 5.3(b) shows the AI-gram of /da/ from talker f103 at 14 dB SNR, which contains a typical high-frequency /da/ burst (block 1) and an interfering mid-frequency /ga/ burst (block 2). Just as in Fig. 5.2 where /ka/ was converted to /ta/ or /pa/, the /da/ sound may be converted into a /ga/ by removing the high-frequency burst (block 1) and scaling up the lower frequency burst (block 2) to create a fully audible mid-frequency



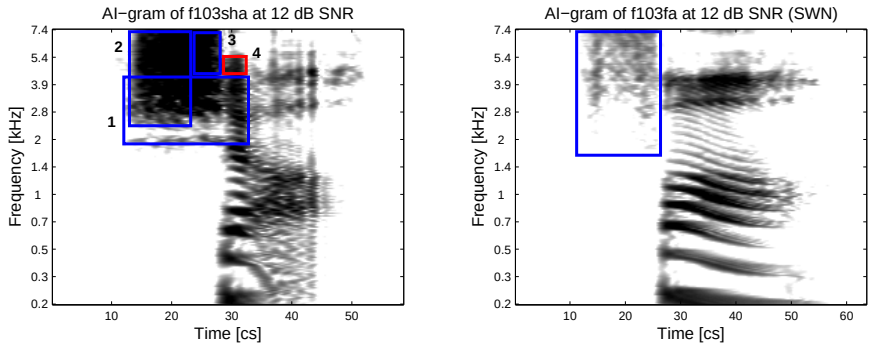
(a) /ba/ → /ga/

(b) /da/ → /ga/

(c) /ga/ → /da/

Figure 5.3: Manipulation of voiced stop consonants /ba, da, ga/. (a) /ba/ from talker f103 morphs into /ga/ when the /ba/ cue in block 1 is replaced by a /ga/ cue in block 2. (Example: ba2ga.) (b) /da/ from talker f103 is heard as a natural /ga/, after removing the high-frequency burst (block 1) and boosting the mid-frequency burst (block 1) by a factor of 5 (14 dB). (Example: da2ga.) (c) Removal of the mid-frequency burst (block 1) causes the original sound /ga/ from talker f103 to morph into a /da/. Boosting the high-frequency burst (block 2) makes the sound a clear /da/. (Example: ga2da.)

burst. The reverse conversion (from /ga/ to /da/) is illustrated in Fig. 5.3(c). After removing the mid-frequency /ga/ cue (block 1), as highlighted by the blue rectangular box, the listeners robustly report /da/. Under some SNR conditions (when the mid-frequency boost is removed and there is insufficient high-frequency residual energy for the labeling of a /da/), a 12 dB boost of the 4 kHz region is required to robustly convert the sound to /da/.



(a) /ʃa/ → /sa/ → /ʧa/ → /za/ → /ða/ (b) /fa/ → /ba/

Figure 5.4: Manipulation of fricatives /ʃa, fa/. (a) The original sound /ʃa/ from talker f103 is converted into a /sa/ when the bandwidth of the noise-like cue is cut from 2–4 kHz (removing block 1). The sound is universally reported as /ʧa/ when the duration is shortened from its natural duration of 15 cs (from 13-28 cs) down to 6 cs (from 22-28 cs) (removing block 2). Combining the two processes (removing block 1 and 2) turns the sound into a /za/. Finally, when all three blocks are taken out, the sound is heard as a /ða/, and boosting the high-frequency residual (block 4) makes the /ða/ clearer. (Example: `Sa2cha2sa2za2Da`.) (b) The original sound /fa/ from talker f103 turns into a /ba/ when the whole fricative cue (highlighted by the blue box) is deleted. (Example: `fa2ba`.)

Fricatives

Unlike the stop consonants, represented by a compact initial burst, the fricatives are characterized by a noise-like cue with varied duration and bandwidth. Cutting the speech cues in bandwidth and duration, we can also convert the fricatives from one into the other. Starting with /ʃa/ from talker f103 (Fig. 5.4(a)), the original sound is heard by all listeners as a solid /ʃa/. The perceptual cue ranges from 13 to 28 cs in time and about 2 to 8 kHz in frequency. For example, cutting the bandwidth in half (remove block 1) morphs the sound into a robust /sa/. Shrinking the duration by 2/3 (remove

block 2) transforms the sound into a /ʧa/. Combining both modifications (remove block 1 and 2) causes most listeners to report /za/. Removing the whole noise patch (remove block 1, 2 and 3) results in /ða/, which can be made robust by amplifying the residual high-frequency burst (highlighted in block 4).

Consonants /fa/ and /va/ are highly confused with /ba/ when the fricative parts of the two sounds are masked. Figure 5.4(b) shows an example of a /fa/→/ba/ conversion. The original sound is a /fa/ from talker f103. When the whole fricative part is removed, it morphs into a robust /ba/.

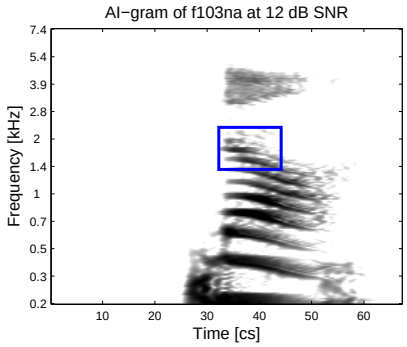
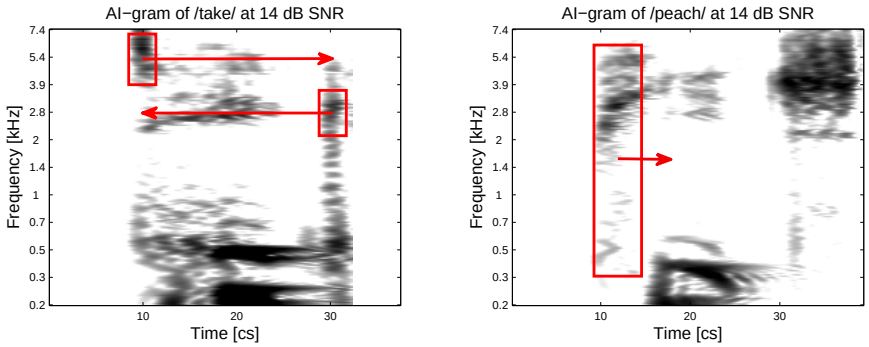


Figure 5.5: AI-gram of /na/ from talker f103. Removing the downward F2 transition turns the /na/ into a /ma/. (Example: na2ma.).

Nasals

The two nasals /ma/ and /na/ share the common feature of a *nasal murmur* and differ from each other in the shape of F2 transition; specifically, /na/ has a prominent downward F2 transition while /ma/ does not. This is because the length of the vocal tract increases with /na/ as the tongue comes off the roof of the mouth, but stays the same length as the lips part, while for /ma/ the tongue remains on the floor of the mouth. Figure 5.5 shows an example of /na/→/ma/ conversion. The original sound is a /na/ from talker f103; when the salient F2 transition is removed, it turns into a /ma/ for which some people can still prime /na/. We have found that it is generally difficult to manipulate the speech cue and turn a /ma/ into a convincing /na/, or vice versa, because the spectral patterns of the two sounds are quite different. The very low-frequency "nasal murmur" does not seem to be critical to the perception of the

nasal class, but is simply characteristic. Namely, it does not seem to be a noise-robust cue used by listeners to label a sound as "nasal."



(a) /take/ → /kate/ (b) /peach/ → /beach/

Figure 5.6: Manipulation of words extracted from continuous speech. (a) A word /take/ morphs into /kate/ when the high-frequency /t/ cue is switched with the mid-frequency /k/ cue. (Example: take2kate.) (b) A word /peach/ turns into /beach/ when the duration between the /p/ burst and the onset of sonorance is reduced from 60 ms to 0 ms. (Example: peach2beach.)

5.3.3 Words

A major difference between words and nonsense syllables is that words are meaningful. This semantic constraint can have a major impact on the perceptual integration of speech cues. In the previous section, we showed that the percept of nonsense CV syllables can be changed through the manipulation of speech cues. A key question is: *Does the same technique apply to words or sentences containing coarticulation and context?*

To explore this question, we chose several words from our speech database and applied our speech-feature modification method. Figure 5.6 shows two such examples, the words /take/ and /peach/, extracted from a sentence. As we see in Fig. 5.6(a), /t/ and /k/ are characterized by a high-frequency burst at the beginning and a mid-frequency burst in the end, respectively. Switching the two cues turns the verb /take/ into the noun /kate/. In Fig. 5.6(b), once the duration between the /p/ burst and the onset of sonorance is removed, /peach/ is reported as /beach/.

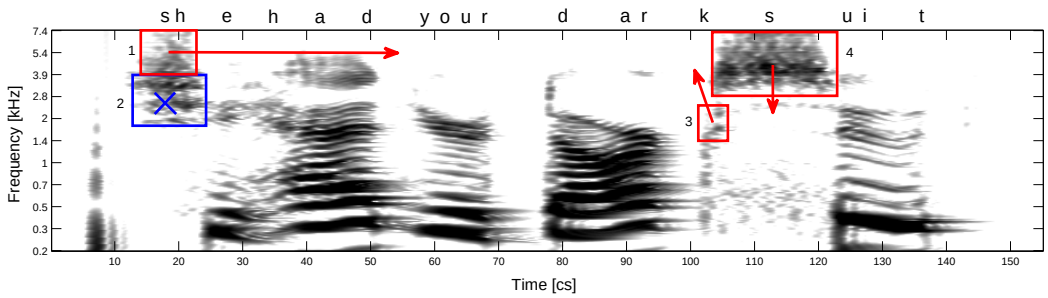


Figure 5.7: Manipulation of speech cues converts a TIMIT sentence /*she had your dark suit*/ into a meaningful new sentence /*he has your dart shoot*/. Step 1: convert /she/ into /he/ by removing the fricative part of /she/ (delete block 1 and 2). Step 2: to convert /had/ into /has/, a /s/ feature is created after /had/ by shifting the upper half of /ʃ/ feature (block 1) to $t = 55$ cs. Step 3: convert /dark/ into /dart/ by shifting the mid-frequency burst (block 3) upward. Step 4: convert /suit/ into a /shoot/ by shifting the /s/ cue (block 4) downward to 2–4 kHz. (Example: `she_had_your_dark_suit`.)

5.3.4 Sentences

The same technique of feature-based speech modification works for natural meaningful sentences, as illustrated in Fig. 5.7 which shows the AI-gram of the sentence /*she had your dark suit*/ at 14 dB SNR (the phones are labeled at the top). Removing the fricative cue around $t = 20$ cs (delete block 1 and 2) morphs the word /she/ into a /he/. Notice that the upper part of the /ʃ/ at 4-8 kHz (block 1) can be used as the perceptual cue of /s/; shifting it from $t = 20$ cs to $t = 55$ cs causes the word /had/ to morph to /has/. Next, we move the mid-frequency /k/ burst in the word /dark/ upward to 4 kHz, which converts the word /dark/ into /dart/. Finally, we changed the /s/ cue in the word /suit/ to be a /ʃ/ cue by shifting it downward from 4-8 kHz to 2-4 kHz, which morphs /suit/ to /shoot/. Thus, the modified sentence becomes /*he has your dart shoot*/. It is relatively easy to change the percept of most sounds once the consonant cues are identified. Interestingly, meaningful sentences may easily be morphed into nonsense by modifying a single event. For example, we can turn the /d/ in /dark/ to a /b/ by zeroing out the frequency component above 1.4 kHz from 75 to 85 cs; then the whole sentence becomes /*she has your bark suit*/, which does not make any sense.

The above examples of sentence modification clearly indicate that speech perception

is dependent on the speech cues. Context information is useful once the listeners can hear the speech cues. A sentence may have key words and accessory words. Similarly, the acoustic cues of continuous speech may be classified into two types: *critical* and *accessory* cues. The critical cues are the irreplaceable units that are critical for perception of the sentence, while the accessory cues are the redundant units recoverable from the critical cues and the associated context information.

Given *a priori* knowledge of specific speech cues, we can change the percept of natural speech through the manipulation of speech cues in CV syllables, words and sentences.

A potential use of this technique is in speech enhancement. In the next section, we will show that speech sounds can be made more robust to noise by enhancing the speech cues.

5.4 Feature-Based Speech Enhancement

People with hearing loss always complain about the difficulty of hearing speech in noisy environments. Depending on the type and degree of hearing loss, it is commonly reported that a hearing-impaired (HI) listener may require a more favorable signal-to-noise ratio (SNR) than normal-hearing (NH) listeners to achieve the same level of performance for speech perception. An alternative hypothesis is that the audibility of speech cues is the key issue. State-of-the-art hearing aids have low functionality in noisy speech because they amplify the entire signal without taking into account the specific features of the speech sounds. If they could automatically detect the onset of the speech cues and selectively enhance them to bring them into the HI's audibility range, they might work better in noise.

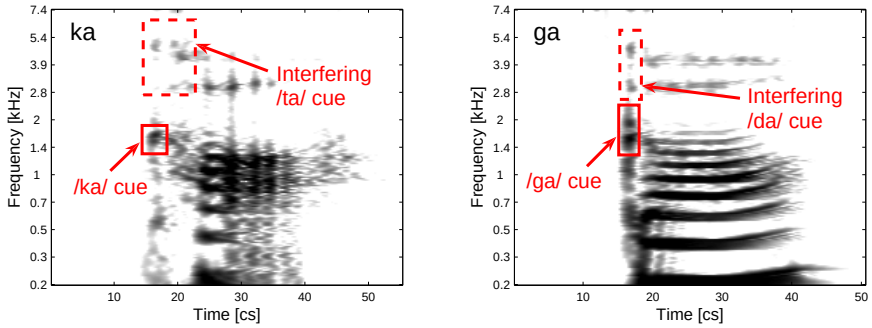
Over the past years, various single-channel noise-reduction techniques have been proposed to increase the SNR [10, 11]. For example, Time-Frequency Gain Manipulation [12] improves the total SNR by assigning larger gains to the time-frequency components with less noise and lower gains to those with more noise.

Since the manipulation is based on the distribution of random noise rather than on prior knowledge about the speech cues, none of these methods have been shown effective in improving speech intelligibility [13]. As a consequence, many HI listeners can hear the amplified noisy speech, but still cannot understand it. To improve HI speech perception,

it is necessary to know the basic elements of speech perception, and according to the AI, it is the SNR in critical bands that counts; while the SNR in a critical band cannot be enhanced, it might be possible to improve the onset dynamics. Furthermore, detailed understanding of the exact sounds causing difficulty to the HI ear has proven to be valuable.

Recently we conducted a hearing-impaired speech perception experiment. One 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, due to a cochlear dead region [6] from 2 to 3.5 kHz, which totally blocks the perceptual cues of /ka/ and /ga/. In contrast, her right ear has no identified cochlear dead regions, and AS can hear these two sounds, but with low accuracy. Confusion analysis indicates that more than 80% of the /ka/s in the left ear were misinterpreted as /ta/, while about 60% of the /ga/s were reported as /da/. A plausible explanation is that the left ear picks up the interfering high-frequency cues, which promote the /ka/→/ta/ and /ga/→/da/ confusions.

To test the idea of feature-base speech enhancement, we conducted a small speech perception experiment on stop consonants to determine the potential benefit of speech cues for speech enhancement. To reduce the "bias" toward /ta/ and /da/, the utterances were modified so that the high-frequency interfering cue was removed, while the mid-frequency perceptual cue was amplified, as depicted in Fig. 5.8.



(a) super /ka/ (b) super /ga/

Figure 5.8: Enhanced /ka/s and /ga/s were created by removing the high-frequency interfering cues (dashed boxes) to promote /ta/→/ka/ responses and /ga/→/da/ confusions, and then boosting the mid-frequency bursts, critical for /ka/ and /ga/ identification.

5.4.1 Methods

The speech stimuli include /pa, ta, ka, ba, da, ga/ and several enhanced super /ka/s and super /ga/s having the mid-frequency /ka/ and /ga/ cue amplified by 1 (no gain), 2 (6 dB gain) and 4 (12 dB gain) respectively. The speech stimuli were chosen from the University of Pennsylvania's Linguistic Data Consortium (LDC) LDC2005S22 "Articulation Index Corpus" such that each nonsense CV syllable has 6 talkers, half male and half female. The speech tokens were fully randomized across talkers, conditions and consonants. A Matlab program was created to control of the procedure. Following each presentation, the subject responded to the stimulus by clicking on the button labeled with the CV. In case the speech is totally unintelligible due to the noise, the subject was instructed to click a "Noise Only" button. The speech stimuli were played at the most comfortable level (MCL) of the listener with no spectral modification for the NH listeners and "NAL-R" enhancement for the HI listener.

5.4.2 Results

To obtain a baseline performance, the speech perception experiment was given to three normal hearing listeners before we tested HI subject AS. For the normal hearing listeners, the speech stimuli were presented to both ears simultaneously under two SNR conditions, -9 and -3 dB SNR, using speech-weighted noise (SWN).

Normal Hearing

Results of the speech perception experiment indicate that boosting the mid-frequency /ka/ and /ga/ cue significantly increases the recognition scores for normal hearing listeners. Table 5.1 shows the confusion matrix of the normal hearing listeners. Each row of the table represents the number of responses made by the listeners when the sound on the left-most column is presented. At -9 dB SNR, removing the interfering high frequency cue from /ka/ reduces the /ta/ confusion from 8 (row 3, col 2) to 5 (row 4, col 2). Enhancing the mid-frequency cue for the target sound by 12 dB increases the number of correct responses from 13 (row 3, col 3) for the original sound /ka/ to 27 (row 6, col 3) for the modified sound $\text{ka}_{t\times 0,\, k\times 4}$. Similar results are observed for /ga/, for

Table 5.1: Normal hearing (NH) listeners.

	-9 dB SNR (SWN)						-3 dB SNR (SWN)					
	pa	ta	ka	ba	da	ga	pa	ta	ka	ba	da	ga
pa	19	7	1	6	5	1	46	1	2	5		
ta	3	42	2		2	1		51	2			1
ka	12	**8**	**13**	3	5	3	6	**3**	**39**	4		1
$\text{ka}_{t\times 0}$	22	**5**	**4**	5		3	22	**4**	**16**	4	1	5
$\text{ka}_{t\times 0,\,k\times 2}$	7	**2**	**14**	2	1	6	6	**1**	**42**		1	4
$\text{ka}_{t\times 0,\,k\times 4}$	3	**1**	**27**	1	2	9	4	**2**	**42**		1	4
ba	4	1	3	8	7	5	8		1	31	6	1
da	5	11	2	3	25	1	1	1	1	1	44	3
ga	4	2	3	7	**16**	**12**	2	3	2	1	**16**	**26**
$\text{ga}_{d\times 0}$	4	3	2	8	**4**	**16**	1			8	**8**	**33**
$\text{ga}_{d\times 0,\,g\times 2}$	1	1	11	3	**10**	**20**				1	**5**	**42**
$\text{ga}_{d\times 0,\,g\times 4}$	1		9	4	**3**	**26**			1		**5**	**48**

$*$ $t, d \times 0$ means removing the interfering /ta/ or /da/ cue;
$k, g \times N$ means amplifying /ka/ or /ga/ cue by a gain factor of N.

which the number of correct responses is 12 (row 9, col 6) for the original sound versus 27 (row 12, col 6) for the enhanced sound $\text{ga}_{d\times 0,\,g\times 4}$. When the SNR increases from -9 to -3, the advantage of feature manipulation is still large for /ga/ with the number of correct responses being 26 (row 9, col 12) for the original sound versus 48 (row 12, col 12) for the enhanced sound /ga/ ($\text{ga}_{d\times 0,\,g\times 4}$); the benefit of speech enhancement becomes minimal for /ka/ as the performance saturates.

Hearing Impaired

Results of the speech perception experiment indicate that feature manipulation significantly changes the nature of speech communication for HI subject AS in quiet, as shown in Table 5.2. Each row contains the number of responses when the CV in the left-most column is presented. It is shown that for the left ear AS can hardly hear the /ka/ sound, with 20 out of 30 /ka/s being reported as /ta/, and 8 out of 30 /ka/s are misinterpreted as /pa/ (row 3). After removing the interfering high-frequency /ta/ cue from /ka/, the /ta/ confusion (column 2) drops dramatically from 20 to 8 (row 3 vs. row 4), while the /pa/ confusion increases from 8 to 16 (row 3 vs. row 4). Due to the impact of the cochlear dead region in the left ear, which blocks the mid-frequency /ka/ cue, feature boosting has a minor effect on /ka/ perception as the number of correct

Table 5.2: Hearing-impaired listener AS in quiet.

	Left Ear						Right Ear					
	pa	ta	ka	ba	da	ga	pa	ta	ka	ba	da	ga
pa	29			1			30					
ta		30						30				
ka	8	**20**	**2**				3	**7**	**17**		1	1
$\text{ka}_{t\times 0}$	16	**8**	**1**		1		1	**4**	**19**	1	3	
$\text{ka}_{t\times 0,\,k\times 2}$	16	**11**	**3**					**10**	**18**			2
$\text{ka}_{t\times 0,\,k\times 4}$	18	**6**	**5**	1				**5**	**25**			
ba				19	1	2				19	3	1
da				6	21	2					30	
ga		3			**7**	**16**		1			**26**	**3**
$\text{ga}_{d\times 0}$		4		2	**4**	**14**		1			**21**	**8**
$\text{ga}_{d\times 0,\,g\times 2}$		4			**6**	**16**		2			**19**	**9**
$\text{ga}_{d\times 0,\,g\times 4}$	1	5			**4**	**20**		3	1		**21**	**5**

$*$ $t, d \times 0$ means removing the interfering /ta/ or /da/ cue;
$k, g \times N$ means amplifying /ka/ or /ga/ cue by a factor of N.

responses increases from 2 to 5, i.e., 10% when the acoustic cue of /ka/ is amplified by a factor of 4 or 12 dB ($\text{ka}_{t\times 0,\,k\times 4}$). Similar results were observed for /ga/ except that the percent correctness is much higher than that of /ka/.

Unlike the left ear, the right ear of AS has more difficulty in identifying /ga/ as compared to /ka/. Of the 30 /ga/s, 26 were misinterpreted as /da/. Removing the interfering /da/ cue in the high-frequency region and boosting the mid-frequency burst reduces the /da/ given /ga/ ($d|g$) confusion from 26 to 19 and increases the correct responses ($g|g$) from 3 to 9, when the acoustic cue of /ga/ is enhanced by 6 dB ($\text{ga}_{d\times 0,\,g\times 2}$). In addition, this feature manipulation increases the $k|k$ responses from 17 to 25. Since the right ear does not have a $t|k$ confusion, removing the interfering cue has little effect.

To investigate the effect of noise on the perception of the feature-enhanced speech, we also tested subject AS with noisy speech sounds. To generate the noisy speech stimulus, the original clean speech was enhanced in the feature regions and then mixed with white noise at 12 dB SNR. Table 5.3 lists the results of the test. Again the feature-enhanced /ka/s and /ga/s have a significantly higher number of correct responses, usually 5 or 6 out of 30 presentations, than the unmodified sounds. However, due to the existence of white noise, removing interfering cues only has a minor effect in reducing the abnormal /ka/→/ta/ confusion in the left ear and the /ga/→/da/ confusion in the right ear. A

Table 5.3: Hearing-impaired listener AS at 12 dB SNR.

	Left Ear						Right Ear					
	pa	ta	ka	ba	da	ga	pa	ta	ka	ba	da	ga
pa	6	8					14	2			1	
ta		26						29	1			
ka	5	**23**	**1**		1			**21**	**6**		1	
$\text{ka}_{t\times 0}$	1	**20**	**2**		1	1	1	**10**	**12**		5	
$\text{ka}_{t\times 0,\,k\times 2}$	3	**17**	**6**		1	3		**16**	**11**	1	1	1
$\text{ka}_{t\times 0,\,k\times 4}$	6	**19**	**4**				3	**19**	**8**			
ba				7	14		1			13	8	
da					26	2					30	
ga		5			**10**	**13**		1			**28**	**1**
$\text{ga}_{d\times 0}$				1	**8**	**18**					**27**	**1**
$\text{ga}_{d\times 0,\,g\times 2}$		3			**14**	**10**					**23**	**7**
$\text{ga}_{d\times 0,\,g\times 4}$		7			**16**	**7**		2			**25**	**3**

$*$ $t, d \times 0$ means removing the interfering /ta/ or /da/ cue;
$k, g \times N$ means amplifying /ka/ or /ga/ cue by N times.

possible explanation is that subject AS has picked up some of the high-frequency noise onsets amplified by the NAL-R and used them as the speech cues.

5.5 Summary and Discussion

Speech perception critically depends on the reception of these speech cues. To process natural speech, it is necessary to have a direct way of determining the cues from natural speech sounds. Using the combined approach of the AI-gram that predicts speech audibility and the 3DDS that measures the contribution of sub-speech components to perception [110], we have identified the speech cues for many initial consonants and manipulated them in natural speech. The following are our major findings:

- Many natural speech sounds, especially stop consonants, contain conflicting cues that are characteristic of competing speech sounds, which could complicate the training of automatic speech recognition software.

- Through the manipulation of the conflicting cues, most often a tiny spot on the spectrogram, the target sound can be convincingly converted into its competing sounds, as demonstrated by the selected examples in this chapter.

- When a hearing-impaired listener confuses a speech sound with its competing sounds, it is because the perceptual cue of the target sound gets masked by the shift of hearing threshold.

- Speech cue manipulation has potential in speech enhancement. A speech sound can be made more robust to noise by boosting the defining speech cue, or the perceptual confusions can be reduced by removing the interfering cue.

The success of feature-based speech processing is largely dependent on the accuracy of identified speech cues. A small change in speech feature may lead to a big difference in perception. For example, the speech stimuli generated by speech synthesizers, such as *pattern playback*, are generally low quality and barely intelligible, because the assumptions about the spectrum patterns are either incomplete or inaccurate. To be more accurate, we need a more realistic cochlear model having upward spread of masking and forward masking.

Automating the detection of the onsets and sorting out multiple talkers, for example, will be a challenge. Presently, there is no way of solving this problem. That humans are naturally good at this task should give us some respite. We would like to approach human performance, based on learning more about the statistical distributions of these critical cues.

CHAPTER 6

CONCLUSION

6.1 Summary

This project aims to gain more insight into the nature of human speech perception and understand why hearing-impaired people have difficulty with noisy speech, so that more advanced signal processing techniques can be developed to help those people. The working hypothesis is that speech sounds are encoded by time-varying spectral patterns called acoustic cues; the processing and detection of these acoustic cues leads to *events*, the psychological correlates of the acoustic cues on the basilar membrane; a hearing-impaired listener may have problems understanding speech simply because he/she cannot hear certain sounds, since the events are missing, due to either the hearing loss, or the masking effect introduced by the noise.

A systematic psychoacoustics method, the 3D Deep Search (3DDS), was developed to identify the perceptual cues of basic speech sounds. The idea is to assess the importance of various speech components from the change of the recognition score due to masking and temporal truncations. For a particular consonant sound, the 3D approach uses three independent experiments to measure these key features of speech perception. Speech sounds are truncated in time, high/low-pass filtered, and masked with white noise, before being presented to a group of normal hearing listeners. When an acoustic cue is essential for speech perception, masking it critically modifies the speech sound and dramatically reduces the recognition score. 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 full band into multiple bands of equal length along the BM of different frequency bands. The third experiment assesses the *event strength* by masking

the speech at various signal-to-noise ratios.

Based on the results of psychophysical studies on isolated CV speech sounds, supplemented by a computational model (AI-gram) that visualizes the audibility of acoustic cues as they propagate on the basilar membrane (BM), we have successfully identified the perceptual cues (events) that are the basic perceptual units for speech recognition. Figure 6.1 provides a summary of the time-frequency map of the perceptual cues for plosives and fricatives preceding vowel /a/. Due to the fundamentally different types of speech cues, the stops are rarely confused with the fricatives. The stop consonants are characterized by a compact burst, caused by the sudden release of pressure in the oral cavity. The voiced and unvoiced stops differ mainly in the duration between the burst and the start of sonorance. Among the sub-group of unvoiced stop consonants, /ta/ is labeled by a high-frequency burst above 4 kHz, /ka/ is defined as a mid-frequency burst from 1.4–2 kHz, whereas /pa/ is represented by a low-frequency burst from 0.7–1 kHz. The three voiced stop consonants /da/, /ga/ and /ba/ have similar frequency patterns. The fricatives are characterized by a patch of wide-band noise created by the turbulent airflow through lips and teeth. Duration and frequency range are critical parameters for fricatives. A voiced fricative usually has a considerably shorter duration than its unvoiced counterpart. The consonant events are consistent across different talkers, even though the parameters, such as timing, frequency and strength, may slightly change within a given range.

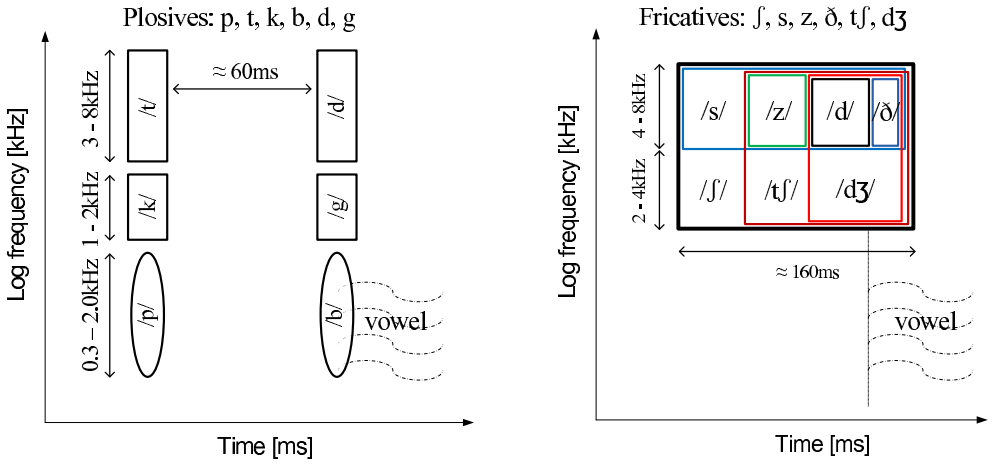


Figure 6.1: A schematic drawing of the perceptual cues for initial consonants preceding vowel /a/, in terms of time-frequency allocation.

Despite its success in feature identification, the 3DDS method is imperfect in several aspects. A major limitation comes from the AI-gram, the what-you-see-is-what-you-hear model that simulates auditory peripheral processing. It is based on a linear model which does not account for cochlear compression, forward masking, upward masking and other neural nonlinear responses. As a consequence, it over-predicts the audibility of speech sounds under certain circumstances. The question is: How do we know that the identified speech cues are accurate and reliable? A software named Beren is developed for the verification of perceptual cues, which usually involves zeroing out the feature area and listening to the modified speech.

In the speech community, it is widely believed that speech contains redundant cues. However, in our study we found that natural speech often contains conflicting cues that are characteristic of competing speech sounds. Through the manipulation of these cues, usually a tiny spot on the AI-gram, we can convert one phone into another phone, or turn a weak speech sound into a strong one. A speech modification software, named KunLun, is developed for the manipulation of speech cues in words and sentences. Plenty of examples have been created and uploaded to the following web site http://hear.ai.uiuc.edu/wiki/Files/VideoDemos. These speech modification examples clearly indicate that speech perception is critically dependent on the speech cues. Context information is useful once the listeners can hear the speech cues. Given a priori knowledge of specific speech cues, we can change the percept of natural speech through the manipulation of speech cues in nonsense syllables, words and sentences.

Next we investigated the impact of sensorineural hearing loss on consonant identification by combining the information about speech cues and hearing loss. In addition to conventional pure tone audiometry (PTA) test, threshold equalized noise (TEN) test and psychoacoustic tuning curve (PTC) test are applied to diagnose possible cochlear dead regions. Two elderly subjects, AS with a moderate loss and DC with a mild-to-severe sloping high-frequency loss, volunteered for the pilot study. Results of speech perception test reveal that AS has no problem with /ta/, and has little difficulty with /pa, sa, da, za/, but never reports /ka/ and /ga/ due to a big cochlear dead region from 2–3 kHz in the left year, which blocks the perceptual cues for /ka/ and /ga/. In contrast, her right ear hears these sounds. Although NAL-R improves the average per-

ception score from 0.42 to 0.53 under quiet conditions, it provides little help for noisy speech (a 3 percent increase at 12 dB SNR); /ka/ and /ga/ remain unintelligible with NAL-R. The other subject, DC, can hear all 16 consonants tested in both ears with the assistance of NAL-R. However, it only improves the recognition scores of low and mid-frequency sounds such as /pa, ba/ and /ka, ga/. It degrades the high-frequency sounds such as /ta/ and /da/. The investigation was extended to many more hearing-impaired subjects during the past half year, and it is still going on at the moment.

A major problem with the hearing impaired study is how to analyze the data, specifically, how to quantify the effect of hearing loss and masking noise on the perception of speech cues, assuming that a speech sound is intelligible if and only if the dominant cue is audible. The only tool available is the speech banana, which is a qualitative tool supposed to work only under quiet conditions. The information of speech cues is based on the formant data of Swedish vowels and consonants measured by Fant during the 1940s. Due to the lack of accurate information about speech cues, most studies can only look at the perceptual score of speech on average, or draw some general conclusions about the correlation between the configuration of hearing loss and the confusion patterns in speech perception without touching the detail. Based on the theory of simultaneous masking and a method proposed by Fletcher for the calculation of effective hearing threshold in noise, we derived the extended speech banana, which integrates the information of speech cues, the configuration of hearing loss, and masking noise. Given the speech intensity, signal-to-noise ratio and pure tone audiogram, it predicts the audibility of speech cues, which can be used to estimate the probability of correctness (P_c) in perceiving a sound. The accuracy of the extended speech banana can be evaluated by comparing the predicted Pc to the actual perceptual scores. Experimental results show that the extended speech banana works well for mild flat hearing loss. For subject AS, who has a big cochlear dead region in the left ear, and for subject DC, who has a severely unbalanced high-frequency loss, the prediction of the extended speech banana is nowhere close to the actual data, suggesting that audibility may not be the only factor that accounts for the disability in speech perception.

Motivated by the success of manipulating speech cues in natural speech, we conducted a small speech perception experiment on subject AS to determine the potential

benefit of feature-based speech enhancement. Confusion analysis indicates that more than 80% of the /ka/s in the left ear were misinterpreted as /ta/, while about 60% of the /ga/s were reported as /da/, because of the interfering high-frequency cues. To reduce the /ka/→/ta/ and /ga/→/da/ confusions, we created some super /ka/s and /ga/s for which the high-frequency interfering cues have been removed. Experimental results indicate that feature manipulation significantly increases the efficiency of speech communication. The /ka/→/ta/ and /ga/→/da/ confusions are much lower for the super /ka/s and /ga/s, as compared to the ordinary ones.

6.2 Contributions, Limitations and Implications

The main contributions and limitations of this study can be summarized as follows:

- A novel psychoacoustic method, named 3D Deep Search (3DDS), is developed to explore the perceptual cues of consonant sounds from natural speech. Compared to the conventional method of synthetic speech, which requires prior knowledge about the speech cues to be identified, the 3DDS method is more practical and reliable. It takes into consideration the natural variance of speech sounds that are beyond the reach of synthetic speech.

- Many natural speech sounds, especially stop consonants, contain conflicting cues that are characteristic of competing sounds. The production of these is likely due to the physical limit of the articulatory organs. Through the manipulation of conflicting cues, most often a tiny spot on the spectrogram, the target sound can be convincingly converted into its competing sounds, as demonstrated by the selected examples. A speech sound can be made more robust to noise by boosting the defining speech cue, or the perceptual confusions can be reduced by removing the interfering cue.

- A quantitative tool named the extended speech banana is derived for the evaluation of hearing-impaired speech perception. Given the speech intensity, signal-to-noise ratio and pure tone audiogram, it predicts the audibility of speech cues. Assuming that a speech sound is intelligible if and only the dominant cue is audible, the extended speech banana can be used to predict the perceptual score

of individual sounds. Results show that it works well for mild flat hearing loss and fails for listeners with cochlear dead region and extremely unbalanced hearing loss, suggesting that audibility, as characterized by the pure tone audiogram, may not be the only significant factor for hearing impaired speech perception.

- Results of hearing impaired speech perception tests indicate that different types of hearing loss, such as flat, sloping and cochlear dead region, have a distinct impact on consonant identification. It is generally true that a hearing-impaired listener cannot hear a sound because the dominant cue that defines the sound is distorted or inaudible due to the hearing loss or masking noise. Under certain circumstances, the hearing impaired listener may learn to use a set of minor cues that are ignored by the average normal hearing listeners because of the existence of the dominant cue.

- Using the high/low-pass data for feature identification, we verified that the multi-band product rule of frequency integration, an empirical formula justified by the two properties about speech and hearing—flat distribution of speech information across the frequency and the independence of critical bands in terms of speech perception—is statistically true for consonants on average. It may also apply to subgroups of consonant sounds, such as stops and fricatives, that are characterized by a flat distribution of speech cues along the frequency. It fails for individual consonants, as expected.

REFERENCES

[1] R. Plomp and A. Mimpen, "Improving the reliability of testing the speech reception threshold for sentences," *Audiology*, vol. 18, no. 1, pp. 43–52, Feb. 1979.

[2] R. Plomp and A. Mimpen, "Speech reception threshold for sentences as a function of age and noise level," *J. Acoust. Soc. Am.*, vol. 66, no. 5, pp. 1333–1342, Nov. 1979.

[3] F. Li and J. Allen, "Identification of perceptual cues for consonant sounds and the influence of sensorineural hearing loss on speech perception," in *Int. Sym. on Hearing*, Salamanca, Spain, 2009, oral presentation.

[4] Y. Yoon and J. Allen, "Signal-to-noise ratio loss and consonant perception in hearing impaired under noisy environment," in *Abstr. Int. Hearing Aid Research Conf.*, Lake Tahoe, CA, 2006.

[5] S. Phatak, Y. Yoon, D. Gooler, and J. Allen, "Consonant loss profiles for hearing impaired listeners," *J. Acoust. Soc. Am.*, 2009, accepted with revision.

[6] B. C. J. Moore, "Dead regions in the cochlea: Conceptual foundations, diagnosis, and clinical applications," *Ear and Hearing*, vol. 25, no. 2, pp. 98–116, Apr. 2004.

[7] E. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[8] B. G. Shinn-Cunningham, "Why hearing impairment may degrade selective attention," in *Inter. Sym. Auditory and Audiological Research*, Denmark, 2007, oral presentation.

[9] S. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1220–1233, Aug. 2008.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech and Sig. Processing*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[11] H. Levitt, "Noise reduction in hearing aids: A review," *J. Rehab. Res. and Development*, vol. 38, pp. 111–121, Jan. 2001.

[12] M. Anzalone, L. Calandruccio, K. Doherty, and L. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, Oct. 2006.

[13] R. Bentler and L. Chiou, "Digital noise reduction: An overview," *Trends in Amplification*, vol. 10, no. 2, pp. 67–82, Jun. 2006.

[14] H. Fletcher and R. Galt, "The perception of speech and its relation to telephony," *J. Acoust. Soc. Am.*, vol. 22, pp. 89–151, 1950.

[15] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, pp. 90–119, 1947.

[16] S. E. Blumstein and K. N. Stevens, "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," *J. Acoust. Soc. Am.*, vol. 66, no. 4, pp. 1001–1017, Oct. 1979.

[17] J. L. McClelland and J. L. Elman, "The trace model of speech perception," *Cognitive Psychology*, vol. 18, pp. 1–86, 1986.

[18] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *J. of Neurophysiology*, vol. 42, no. 4, pp. 1083–1107, 1979.

[19] B. Delgutte, "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers," *J. Acoust. Soc. Am.*, vol. 63, no. 3, pp. 843–857, 1980.

[20] S. A. Shamma, "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve," *J. Acoust. Soc. Am.*, vol. 78, no. 5, pp. 1612–1621, 1985.

[21] H. Stevens and R. E. Wickesberg, "Ensemble responses of the auditory nerve to normal and whispered stop consonants," *Hearing Res.*, vol. 131, pp. 47–62, 1999.

[22] R. Jakobson, C. Gunnar, M. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, Massachusetts: The MIT Press, 1961.

[23] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York, NY: Harper and Row Publishers, 1968.

[24] K. Stevens and S. Blumstein, *The Search for Invariant Acoustic Correlates of Phonetic Features*, P. Eimas and J. Miller, Eds. Erlbaum, Hillsdale, NJ: Lawrence Erlbaum Associates, 1981.

[25] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am.*, vol. 64, no. 5, pp. 1358–1369, Nov. 1978.

[26] A. S. Bregman and J. Campell, "Primary auditory stream segregation and perception of order in rapid sequences of tones," *J. Exp. Psych*, vol. 89, no. 2, pp. 244–249, 1971.

[27] A. S. Bregman, *Auditory Scene Analysis*. Cambridge 39, Massachusetts: MIT Press, 1990.

[28] R. Remez, P. Rubin, S. Berns, J. Pardo, and J. Lang, "On the perceptual organization of speech," *Psychol. Review*, vol. 101, no. 1, pp. 129–156, 1994.

[29] A. Liberman, F. Cooper, D. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code," *Psychol. Review*, vol. 74, no. 6, pp. 431–61, Nov. 1967.

[30] A. Liberman and I. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[31] P. Kuhl and J. Miller, "Speech perception by the chinchila: Voiced-voiceless distinction in alveolar plosive consonants," *Science*, vol. 190, no. 4209, pp. 69–72, Oct. 1975.

[32] P. Kuhl and J. Miller, "Speech perception by the chinchila: Identification functions for synthetic VOT stimuli," *J. Acoust. Soc. Am.*, vol. 63, no. 3, pp. 905–917, Mar. 1978.

[33] P. Reed, P. Howell, S. Sackin, L. Pizzimenti, and S. Rosen, "Speech perception in rats: Use of duration and rise time cues in labeling of affricate/fricative sounds," *J. Exp. Analy. Behavior*, vol. 80, no. 2, pp. 205–215, Sep. 2003.

[34] C. A. Fowler, "Segmentation of coarticulated speech in perception," *Perception and Psychophysics*, vol. 36, no. 4, pp. 359–368, 1984.

[35] W. Marslen-Wilson and L. Tyler, "The temporal structure of spoken language understanding," *Cognition*, vol. 8, no. 1, pp. 1–71, Mar. 1980.

[36] W. D. Marslen-Wilson, "Functional parallelism in spoken word-recognition," *Cognition*, vol. 25, pp. 71–102, 1987.

[37] D. Massaro and G. Oden, "Evaluation and integration of acoustic features in speech perception," *J. Acoust. Soc. Am.*, vol. 67, pp. 996–1013, 1980.

[38] D. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1987.

[39] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1872–3000, Apr. 2002.

[40] R. K. Potter, G. A. Kopp, and H. G. Kopp, *Visible Speech*. New York: Dover Publications, Inc, 1966.

[41] G. Fant, *Speech Sounds and Features*. Cambridge, Massachusetts: The MIT Press, 1973.

[42] F. Cooper, P. Delattre, A. Liberman, J. Borst, and L. Gerstman, "Some experiments on the perception of synthetic speech sounds," *J. Acoust. Soc. Am.*, vol. 24, no. 6, pp. 597–606, Nov. 1952.

[43] P. Delattre, A. Liberman, and F. Cooper, "Acoustic loci and translational cues for consonants," *J. Acoust. Soc. Am.*, vol. 24, no. 4, pp. 769–773, Jul. 1955.

[44] S. E. Blumstein, K. N. Stevens, and G. N. Nigro, "Property detectors for bursts and transitions in speech perceptions," *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1301–1313, May 1977.

[45] G. Hughes and M. Halle, "Spectral properties of fricative consonants," vol. 28, no. 2, pp. 303–310, Mar. 1956.

[46] J. Heinz and K. Stevens, "On the perception of voiceless fricative consonants," *J. Acoust. Soc. Am.*, vol. 33, no. 5, pp. 589–596, May 1961.

[47] A. Malécot, "Acoustic cues for nasal consonants: An experimental study involving a tape-splicing technique," *J. Acoust. Soc. Am.*, vol. 32, no. 2, pp. 274–284, Jun. 1956.

[48] A. Liberman, "Some results of research on speech perception," *J. Acoust. Soc. Am.*, vol. 29, no. 1, pp. 117–123, Jan. 1957.

[49] D. Recasens, "Place cues for nasal consonants with special reference to Catalan," *J. Acoust. Soc. Am.*, vol. 73, no. 4, pp. 1346–1353, 1983.

[50] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech*. New York: Springer-Verlag, 1993.

[51] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 971–995, 1980.

[52] G. Miller, G. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," *J. Exp. Psych*, vol. 41, no. 5, pp. 329–335, May 1951.

[53] R. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392–393, Jan 1970.

[54] H. Savin and T. Bever, "The nonperceptual reality of the phoneme," *J. Verbal Learning and Verbal Behavior*, vol. 9, pp. 295–302, Jun. 1970.

[55] N. Kazanina, C. Phillips, and W. Idsardi, "The influence of meaning on the perception of speech sounds," *Proc. Nat. Acad. Sci.*, vol. 103, no. 30, pp. 11 381–11 386, July 2006.

[56] R. Remez, P. Rubin, D. Pisoni, and T. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, no. 4497, pp. 947–949, May 1981.

[57] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 2, pp. 1053–1064, Feb 1994.

[58] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, vol. 95, no. 5, pp. 2670–2680, May 1994.

[59] H. Dudley, "The vocoder," *Bell Labs Rec.*, vol. 17, pp. 122–126, 1939.

[60] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, Oct. 1995.

[61] F.-G. Zeng, G. Stickney, Y. Kong, M. Vongphe, A. Bhargave, C. Wei, and K. Cao, "Speech recognition with amplitude and frequency modulations," *Proc. Nat. Acad. Sci.*, pp. 2293–2298, Feb. 2005.

[62] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

[63] P. Ladefoged, *A Course in Phonetics*. Boston, MA: Heinle & Heinle, 1975.

[64] G. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 338–352, Mar. 1955.

[65] R. Ahmed and S. Agrawal, "Significant features in the perception of (Hindi) consonants," *J. Acoust. Soc. Am.*, vol. 45, no. 3, pp. 758–763, May 1968.

[66] S. Singh, D. Woods, and G. Becker, "Perceptual structure of 22 prevocalic English consonants," *J. Acoust. Soc. Am.*, vol. 52, no. 6, pp. 1698–1713, May 1972.

[67] M. D. Wang and R. C. Bilger, "Consonant confusions in noise: A study of perceptual features," *J. Acoust. Soc. Am.*, vol. 54, pp. 1248–1266, May 1973.

[68] M. E. Hayden, E. Kirstein, and S. Singh, "Role of distinctive features in dichotic perception of 21 English consonants," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 1039–1046, Apr. 1979.

[69] J. B. Allen, "Nonlinear cochlear signal processing and masking in speech perception," in *Springer Handbook on Speech Processing and Speech Communication*, J. Benesty and M. Sondhi, Eds. Heidelberg Germany: Springer, 2008, ch. 3, pp. 1–36.

[70] H. Fletcher, *Speech and Hearing in Comminication*, ASA ed., J. B. Allen, Ed. Woodbury, NY: Acoustical Society of America, 1995.

[71] J. B. Allen, "How do humans process and recognize speech?" *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[72] C. E. Shannon, "The mathematical theory of communication," *Bell System Tech. J.*, vol. 27, pp. 379–423 (parts I, II), 623–656 (part III), 1948.

[73] *American National Standard Methods for the Calculation of the Articulation Index*. New York, NY: American National Standards Institute, 1969, ANSI S3.5-1969.

[74] *Methods for Calculation of the Speech Intelligibility Index (SII-97)*. New York, NY: American National Standards Institute, 1997, ANSI S3.5-1997.

[75] L. L. Beranek, "The design of speech communication systems," *Proc. IRE*, vol. 35, no. 9, pp. 880–890, Sep. 1947.

[76] K. D. Kryter, "Methods for the calculation and use of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1689–1697, Nov. 1962.

[77] H. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[78] C. V. Pavlovic, "Use of the articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," *J. Acoust. Soc. Am.*, vol. 75, pp. 1253–1258, Nov. 1984.

[79] C. V. Pavlovic, G. A. Studebaker, and R. L. Sherbecoe, "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," *J. Acoust. Soc. Am.*, vol. 80, no. 1, pp. 50–57, Jul. 1986.

[80] G. A. Studebaker, C. V. Pavlovic, and R. L. Sherbecoe, "A frequency importance function for continuous discourse," *J. Acoust. Soc. Am.*, vol. 81, no. 4, pp. 1130–1138, Apr. 1987.

[81] V. Duggirala, G. A. Studebaker, C. V. Pavlovic, and R. L. Sherbecoe, "Frequency importance functions for a feature recognition test material," *J. Acoust. Soc. Am.*, vol. 83, no. 6, pp. 2372–2382, Jun. 1988.

[82] J. B. Allen, *Articulation and Intelligibility.* San Rafael, CA: Morgan and Claypool Publishers, 2005.

[83] K. D. Kryter, "Validation of the articulation index," *J. Acoust. Soc. Am.*, vol. 34, no. 11, pp. 1698–1702, Nov. 1962.

[84] K. W. Grant and L. D. Braida, "Evaluating the articulation index for auditory-visual input," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2952–2960, Jun. 1991.

[85] R. P. Lippmann, "Accurate consonant perception without mid-frequency speech energy," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 66–69, Jan. 1996.

[86] H. Müsch and S. Buus, "Using statistical decision theory to predict speech intelligibility I. Model structure," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 2896–2909, Jun. 2001.

[87] H. Müsch and S. Buus, "Using statistical decision theory to predict speech intelligibility II. Measurement and prediction of consonant-descrimination performance," *J. Acoust. Soc. Am.*, vol. 109, no. 6, pp. 2910–2920, Jun. 2001.

[88] H. Steeneken and T. Houtgast, "Mutual dependence of octave-band weights in predicting speech intelligibility," *Speech Communication*, vol. 28, no. 2, pp. 109–123, Jun. 1999.

[89] D. Ronan, A. K. Dix, S. Shah, and L. D. Braida, "Integration across frequency bands for consonant identification," *J. Acoust. Soc. Am.*, vol. 116, no. 3, pp. 1749–1762, Sep. 2004.

[90] T. Ching, H. Dillon, and D. Byrne, "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency," *J. Acoust. Soc. Am.*, vol. 103, no. 2, pp. 1128–1140, Feb. 1998.

[91] S. Phatak and J. Allen, "Consonant and vowel confusions in speech-weighted noise," *J. Acoust. Soc. Am.*, vol. 121, no. 4, pp. 2312–2326, Apr. 2007.

[92] D. D. Greenwood, "A cochlear frequency-position function for several species – 29 years later," *J. Acoust. Soc. Am.*, vol. 87, pp. 2592–2605, Jun. 1990.

[93] B. Lobdell and J. Allen, "A model of the vu (volume-unit) meter, with speech applications," *J. Acoust. Soc. Am.*, vol. 121, pp. 279–285, Oct. 2007.

[94] M. S. Régnier and J. B. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2801–2814, May 2008.

[95] J. B. Allen, "Harvey Fletcher's role in the creation of communication acoustics," *J. Acoust. Soc. Am.*, vol. 99, no. 4, pp. 1825–1839, Apr. 1996.

[96] J. B. Allen, "Consonant recognition and the articulation index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2212–2223, Apr. 2005.

[97] S. E. Blumstein and K. N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Am.*, vol. 67, no. 2, pp. 648–266, Feb. 1980.

[98] A. Alwan, "Modeling speech perception in noise: The stop consonants as a case study," PhD disseration in Electrical and Computer Engineering, MIT, Cambridge, Massachusetts, 1992.

[99] V. Hazan and S. Rosen, "Individual variability in the perception of cues to place contrasts in initial stops," *Perception and Psychophysics*, vol. 49, no. 2, pp. 187–200, 1991.

[100] B. E. Lobdell, "Models of human phone transcription in noise based on intelligibility predictors," PhD dissertation in Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Beckman Institute, Urbana, IL, 2009.

[101] S. Furui, "On the role of spectral transition for speech perception," *J. Acoust. Soc. Am.*, vol. 80, no. 4, pp. 1016–1025, Oct. 1986.

[102] J. B. Allen, "Short time spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Trans. Acoust. Speech and Sig. Processing*, vol. 25, no. 3, pp. 235–238, Jun. 1977.

[103] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, Nov. 1977.

[104] J. B. Allen, M. Régnier, S. Phatak, and F. Li, "Nonlinear cochlear signal processing and phoneme perception," in *Proceedings of the 10th Mechanics of Hearing Workshop*, N. P. Cooper and D. T. Kemp, Eds. Singapore: World Scientific Publishing Co., 2009, pp. 95–107.

[105] C. E. Shannon, "Communication in the presence of noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, Jan. 1949.

[106] H. Duifhuis, "Level effects in psychophysical two-tone suppression," *J. Acoust. Soc. Am.*, vol. 67, no. 3, pp. 914–927, 1980.

[107] M. Zilany and I. Bruce, "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *J. Acoust. Soc. Am.*, vol. 120, pp. 1446–1466, Nov. 1987.

[108] B. Moore, M. Huss, D. A. Vickers, B. R. Glasberg, and J. I. Alcántara, "A test for the diagnosis of dead regions in the cochlear," *British J. Audiology*, vol. 34, no. 4, pp. 205–224, Aug. 2000.

[109] B. Moore and J. I. Alcántara, "The use of psychophysical tuning curves to explore dead regions in the cochlea," *Ear and Hearing*, vol. 22, no. 4, pp. 268–278, Aug. 2001.

[110] J. B. Allen and F. Li, "Speech perception and cochlear signal processing," *IEEE Signal Processing Magzine*, vol. 26, no. 4, pp. 73–77, Jul. 2009.

[111] R. Bilger and M. Wang, "Consonant confusions in patients with senseoryneural loss," *J. Speech and Hearing Res.*, vol. 19, no. 4, pp. 718–748, Dec. 1976.

[112] M. D. Wang, C. M. Reed, and R. C. Bilger, "A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions," *J. Speech and Hearing Research*, vol. 21, pp. 5–36, Mar. 1978.

[113] D. A. Fabry and D. Van Tasell, "Masked and filtered simulation of hearing loss: effects on consonant recognition," *J. Speech and Hearing Research*, vol. 29, pp. 170–178, Jun. 1986.

[114] J. R. Dubno and A. B. Schaefer, "Comparison of frequency selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 91, no. 4, pp. 2110–2121, 1992.

[115] P. Zurek and L. Delhorne, "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," *J. Acoust. Soc. Am.*, vol. 82, no. 5, pp. 1548–1559, Nov. 1987.

[116] W. A. Dreschler and R. Plomp, "Relation between psychophysical data and speech perception for hearing-impaired subjects," *J. Acoust. Soc. Am.*, vol. 68, no. 6, pp. 1608–1615, 1980.

[117] R. Plomp, "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," *J. Speech and Hearing Research*, vol. 29, pp. 146–154, Jun. 1986.

[118] J. M. Festen and R. Plomp, "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *J. Acoust. Soc. Am.*, vol. 88, no. 4, pp. 1725–1736, 1990.

[119] G. Fant, "Phonetics and phonology in the last 50 years," in *From Sound to Sense*, no. B, MIT, Jun. 2004, pp. 20–41.

[120] F. Li and J. B. Allen, "Multiband product rule and consonant identification in white noise," *J. Acoust. Soc. Am.*, vol. 126, no. 1, pp. 347–353, Jul. 2009.

[121] F. Li, A. Menon, and J. B. Allen, "Perceptual cues in natural speech for 6 stop consonants," *J. Acoust. Soc. Am.*, 2009, submitted.

[122] B. Moore, *Cochlear Hearing Loss*. London, UK: Whurr Publishers Ltd., 1998.

[123] B. Moore, "Dead regions in the cochlea: Diagnosis, perceptual consequences, and implications for the fitting of hearing aids," *Trends in Amplification*, vol. 5, no. 1, pp. 1–35, 2001.

[124] L. Rabiner, "The power of speech," *Science*, vol. 301, pp. 1494–1495, Sep. 2003.

[125] S. Dusan and L. Rabiner, "Can automatic speech recognition learn more from human speech perception?" in *Trends in Speech Technology*, C. Bunleanu, Ed. Bucharest, Romania: Romanian Academic Publisher, 2005, pp. 21–36.

[126] H. Hermansky, "Should recongizers have ears?" *Speech Communication*, vol. 25, pp. 3–27, Aug. 1998.

[127] J. T. Huang and M. Hasegawa-Johnson, "Maximum mutual information estimation with unlabeled data for phonetic classification," in *Proc. Interspeech*, 2008.

[128] D. Kewley-Port, "Time-varying features as correlates of place of articulation in stop consonants," *J. Acoust. Soc. Am.*, vol. 73, no. 1, pp. 322–335, Jan. 1983.

[129] B. Delgutte, "Auditory neural processing of speech," in *The Handbook of Phonetic Sciences*, W. Hardcastle and J. Laver, Eds. Oxford, UK: Blackwell, 1997, pp. 507–538.

[130] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, ASA ed. New York: Academic Press Inc., 1965.