TECHNIQUES FOR UNDERSTANDING HEARING-IMPAIRED PERCEPTION OF CONSONANT CUES

BY

ANDREA CAROLINA TREVINO

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, 2013

Urbana, Illinois

Doctoral Committee:

Associate Professor Jont B. Allen, Chair
Professor Mark Hasegawa-Johnson
Professor Stephen E. Levinson
Professor Peggy B. Nelson, University of Minnesota

ABSTRACT

We examine the cues used for consonant perception and the systematic behavior of normal and hearing-impaired listeners. All stimuli were presented as isolated consonant-vowel tokens, using the vowel /ɑ/. Use of low-context stimuli, such as consonants, aids in minimizing the influence of some variable cognitive abilities (e.g., use of context, memory) across listeners, and focuses on differences in the processing or interpretation of the existing acoustic consonant cues.

In a previous study on stop consonants, the 3D Deep Search (3DDS) method for the exploration of the necessary and sufficient cues for normal-hearing speech perception was introduced. Here, this method is used to isolate and analyze the perceptual cues of the naturally produced American English fricatives /ʃ, ʒ, s, z, f, v, θ, ð/ in time, frequency, and intensity. The 3DDS analysis labels the perceptual cues of sibilant fricatives /ʃa, ʒa, sa, za/ as a sustained frication noise preceding the vowel onset, with the acoustic cue for both /sa, za/ located between 3.8–7 kHz, and the acoustic cue for both /ʃa, ʒa/ located between 2–4 kHz. The /ʃa, ʒa/ utterances were also found to contain frication components above 4 kHz in natural speech that are unnecessary for correct perception, but can cause listeners to correspondingly hear /sa, za/ when the dominant cue between 2–4 kHz is removed by filtering; such cues are denoted "conflicting cues". While unvoiced fricatives were observed to generally have a longer frication period than their voiced counterparts, duration of frication was found to be an unreliable cue for the differentiation of voiced from unvoiced fricatives. The wideband amplitude-modulation of the F2 and F3 formants at the pitch frequency F0 was found to be a defining cue for voicing. Similar to previous results with stop consonants, the robustness of fricative consonants to noise was found to be significantly correlated to the intensity of the acoustic cues that were isolated with the 3DDS method.

The consonant recognition of 17 ears with sensorineural hearing loss is evaluated for fourteen consonants /p, t, k, f, s, ʃ, b, d, ɡ, v, z, ʒ, m, n/+/ɑ/, under four speech-weighted noise conditions (0, 6, 12 [dB] SNR, quiet). For a single listener, we find that high errors can exist for a small subset of test stimuli, while performance for the majority of test stimuli can remain at ceiling. We show that hearing-impaired perception can vary across multiple tokens of the same consonant, in both noise-robustness and confusion groups. Within-consonant differences in noise-robustness are related to natural variations in intensity of the consonant cue region. Within-consonant differences in confusion groups entail that an average over multiple tokens of the same consonant results in a larger confusion group than for a single consonant token, causing the listener to appear to behave in a less systematic way. At the token level, hearing-impaired listeners are relatively consistent in their low-noise confusions; confusion groups are restricted to fewer than three confusions, on average. For each consonant token, the same confusion group is consistently observed across a population of hearing-impaired listeners. Quantifying these token differences provides insight into hearing-impaired perception of speech under noisy conditions and characterizes each listener's hearing impairment.

Auditory training programs are currently being explored as a method of improving hearing-impaired speech perception; precise knowledge of a patient's individual differences in speech perception allows for a more accurately prescribed training program. Re-mapping or variations in the weighting of acoustic cues, due to auditory plasticity, can be examined with the detailed confusion analyses that we have developed. Although the tested tokens are noise-robust and unambiguous for normal-hearing listeners, the subtle natural variations in signal properties can lead to systematic within-consonant differences for hearing-impaired listeners. At the individual token level, a k-means clustering analysis of the confusion data shows that hearing-impaired listeners fall into similar confusion-based groups. Many of the token-dependent confusions that define these groups can also be observed for normal-hearing listeners, under higher noise levels or filtering conditions. These hearing-impaired listener groups correspond to different acoustic-cue weighting schemes, highlighting where auditory training should be most effective.

To my parents and sister, for always knowing I would make it this far

ACKNOWLEDGMENTS

I would like to thank my Ph.D. advisor Jont Allen, my Master's degree advisor Todd Coleman, along with Tom Quatieri and Bob Wickesberg for their mentoring, guidance, and research advice throughout the Ph.D. process. Their passion for research is infectious, and has inspired me to attempt to continue in the field as I move forward.

Many friends helped keep me sane along the way, including, Tony Mangognia, Justin Koepke, Jason Laska, Jason and Heidi Lamb, Abhinauv Kapoor, David Jun, Brian Stieber, Miri Kim, Sarah Holton, Helen Hwang, Tracy Flood, Christine Yang, Dave Cohen, Dan Fisher, Tommy Gehrels, Erik Johnson, Sarah Robinson, Laura Sloofman, Emily Sprague, Markita Landry, Vishal Kottadiel, Michael Friedman, and the Church of the Infinite Paradigm. I have to thank Anthony Christodoulou for being the perfect partner in crime.

Finally, I would like to thank my family, I couldn't have done it without their support. They have always helped me to remember who I am and believe that I have the ability to push through anything that life throws at me.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

3DDS	3-Dimensional Deep Search
AI	Articulation Index
AM	Amplitude Modulation
ANOVA	Analysis of Variance
CP	Confusion Pattern
CV	Consonant Vowel
HI	Hearing-Impaired
HL	Hearing Loss
HSR	Human Speech Recognition
INDSCAL	Individual Differences in Scaling
MCL	Most Comfortable Level
NH	Normal-Hearing
SNR	Signal to Noise Ratio
SPL	Sound Pressure Level
SWN	Speech-Weighted Noise
WN	White Noise

CHAPTER 1

INTRODUCTION

The general goal of this research is to better understand how hearing-impaired (HI) listeners perceive speech, by focusing on the perception of consonants. In order to explore the cues that HI listeners use to recognize consonants, the cues that normal-hearing (NH) listeners use must first be characterized.

The work of the Human Speech Recognition (HSR) research group at the University of Illinois Urbana-Champaign, particularly by Phatak (2007), Régnier and Allen (2008) and Li (2010), has made great strides toward characterizing the NH perceptual cues on an individual token basis. Previous publications of the HSR research group have focused on stop consonants, their work is extended here to the fricative consonants (see Chapter 3).

Seventeen HI ears with slight-to-severe hearing loss were tested on a consonant recognition task, in both quiet and speech-weighted noise (SWN). Consonant-vowel stimuli, presented in isolation, were chosen as the stimuli to minimize the influence of some variable cognitive abilities (e.g., use of context, memory) across listeners, therefore allowing for a focus on the processing and perception of existing consonant cues by HI listeners. Our analysis focuses on examining the individual differences across HI ears and different stimuli. The methodologies developed by the HSR research group are then used to investigate possible sources of these individual differences.

1.1 Consonant Cues

When isolating the cues used for speech perception, a widely recognized key problem for analyzing natural speech is the large variability introduced by the speaker (e.g., pitch, rate). Following the 1930-1940 development of the speech "vocoder" at Bell Labs, speech synthesis has been a hallmark of speech perception research. Beginning at Haskins Laboratories in the 1950s (Cooper et al. 1952; Delattre et al. 1955; Bell et al. 1961), almost all of the classical studies have used vocoder speech (Shannon et al. 1995) or speech synthesis methods (Hughes and Halle 1956; Heinz and Stevens 1961; Blumstein et al. 1977; Stevens and Blumstein 1978), as a way of controlling the variability of speech cues when performing perceptual experiments. A major disadvantage of this method is that one must first make assumptions about the nature of perceptual cues in order to synthesize the target speech stimuli; depending on the accuracy of these assumptions, this can potentially lead to listeners using different cues for recognition than they would for natural speech. Synthesized speech can often sound unnatural or have low baseline intelligibility (Delattre et al. 1955; Remez et al. 1981). Later studies analyzed the spectrum of natural speech (Soli 1981; Baum and Blumstein 1987; Behrens and Blumstein 1988; Shadle and Mair 1996; Jongman et al. 2000) and identified the acoustic cues sufficient for sound identification/discrimination, but without verifying them against human psychoacoustic data. While the results characterize the variability of natural speech, it remains uncertain whether those cues are indeed necessary and sufficient for speech perception.

To determine the cues for speech perception, the HSR group has developed a new methodology named *3-Dimensional Deep Search* (3DDS) (Li et al. 2010) that analyzes the perceptual contributions of naturally produced speech components based on the results of three independent psychoacoustic experiments with NH listeners. This is paired with a time-frequency representation that predicts audibility, the AI-gram (Régnier and Allen 2008; Lobdell 2009; Lobdell et al. 2011), to observe which acoustic cues remain audible as masking noise is introduced.

What are the necessary and sufficient perceptual cues of fricative consonants? We explore this question by using the 3DDS method to analyze perceptual data from a larger past study which gathered data for 16 consonants, including 6 plosives, 8 fricatives and 2 nasals, with 6 talkers per consonant (Phatak et al. 2008; Li et al. 2010). Isolating the spectral region that contains the necessary and sufficient perceptual cues is equivalent to stating that normal-hearing listeners can correctly perceive the target consonant if (sufficient) and only if (necessary) the cues contained in that region are present. The results for a similar analysis on stop consonants are discussed in Li et al. (2010). Our second component generalizes the 3DDS analysis to the American English fricatives /ʃ, ʒ, s, z, f, v/.

1.2 Hearing-Impaired Consonant Perception

Given that the primary purpose of wearing a hearing aid is to improve speech perception, it follows that a speech test should be able to provide one of the most useful measures of hearing impairment. Yet, speech has not been found to be a useful diagnostic tool for fitting hearing aids (Walden et al. 1983; Dobie 2011). Pure-tone thresholds remain the primary prescriptive measure for hearing aid fitting (Humes et al. 1991; Dillon 2001) despite the common clinical observation that HI ears can have similar pure-tone thresholds but differ in their speech perception abilities (Skinner 1976; Skinner and Miller 1983; Kamm et al. 1985; Smoorenburg 1992; Roeser et al. 2007; Halpin and Rauch 2009; Walden and Montgomery 1975; Killion and Niquette 2000). Differences in speech perception abilities are most commonly examined using the average score and/or speech recognition threshold (SRT); we examine the variability of speech perception abilities across individual consonant tokens. A significant impediment to research in developing speech-based measures is the large amount of natural variability in speech and the difficulty in characterizing the perceptually relevant cues. When the acoustic properties of the perceptual cues in a speech test are not characterized, the conclusions that may be drawn are limited.

The work of Boothroyd and Nittrouer (1988) formulated the relationship between correct perception of low-context speech segments (e.g., phonemes) and high-context segments (e.g., words) in NH ears. Follow-up studies by Bronkhorst et al. (1993, 2002) greatly extended this work. These studies demonstrate that an individual's ability to decode high-context speech depends critically on his or her low-context error, a conclusion first addressed by Fletcher et al. (1996). When HI listeners report that they can "hear speech but have trouble understanding it," it may be due to a small group of specific phonemes being incorrectly recognized. These observations affirm the utility of studies of hearing impairment that use low-context speech segments.

We have found that, in both quiet and low-noise conditions, errors can be concentrated on a small, ear-dependent subset of the test stimuli. Thus, an ear that seems "almost normal" in terms of an average error measure can, in actuality, have high error with a small, specific group of sounds. In addition, ears with similar degrees and configurations of hearing loss and similar average consonant errors can have very different individual consonants that fall into error.

Multiple tokens of the same consonant, by different talkers or within different contexts, are often considered as multiple measures of the same effect. Contrary to this approach, the consonant cue literature has documented, in detail, the variability of the cues that are present in naturally produced speech (Baum and Blumstein 1987; Dorman et al. 1977; Herd et al. 2010; Jongman et al. 2000; Kurowski and Blumstein 1987; Li et al. 2010, 2012). This variability is quantified by analysis of the acoustical properties of each consonant token and can be observed across speech samples that are unambiguous and robust to noise for NH listeners. The question remains: does the natural variability within tokens of the same consonant lead to differences in HI perception?

We find that HI perceptual differences exist across multiple tokens of a single consonant (which show no recognition differences for NH listeners). We refer to perceptual differences across multiple tokens of the same consonant as *within-consonant* differences. The HI within-consonant differences are observed in terms of both robustness to noise and/or confusion groups. Tokens that show HI within-consonant differences in terms of noise robustness do not necessarily also have differences in confusion groups, nor vice versa.

Within-consonant differences in noise robustness are observed over all the HI subjects. Previous studies have shown that, for individual consonant tokens, the intensity of each necessary cue region is correlated to the NH robustness to noise (Régnier and Allen 2008; Li et al. 2010, 2012). We test if natural variations in the intensity of the acoustic cue region, which affect NH perception at low SNRs, would be observed similarly in the HI results, at higher SNRs. Although a significant correlation is observed, HI within-consonant noise-robustness differences in this dissertation are only partially explained by the natural variations in the intensity of the necessary consonant cue region. In order to further examine if the variability in the acoustic properties can lead to differences in HI perception, the confusion groups of individual tokens are also analyzed.

We observe that each token of a consonant has a unique subgroup of possible confusions, and that these confusion groups can be different across tokens of the same consonant. Although there is often overlap in the confusion groups across multiple tokens of the same consonant, systematic within-consonant differences are common. The averaged responses of HI ears to multiple tokens of a single consonant can often appear to be random guesses drawn from a large confusion group. Some of this randomness is an artifact of averaging across tokens; smaller confusion groups are observed when HI subjects are examined at the token level.

When examined at the token level, we have also observed that the token-specific confusion groups are shared across different HI ears, implying that similar acoustic cues are being used across different HI listeners when making confusions. Although, for HI ears, the SNR at which errors are first made can vary widely, similarities in confusions, once an error is made, are observed for all consonant tokens. Thus, we can conclude that the subtle variability in the acoustical properties, that do not affect NH recognition, are the source of the systematic within-consonant differences in confusion groups for HI listeners.

In order to explore the extent of these similarities in consonant confusions, the k-means clustering algorithm, based on the Hellinger distance, is used to analyze the confusion matrix data. The k-means clustering approach to consonant confusion matrix analysis is a novel application, and allows for analysis of confusion matrix data without averaging across tokens, HI ears, or SNRs. We find that the number of statistically significant token-dependent clusters of the HI data is small (≤ 4); this result, paired with the angle between elements of each cluster, quantifies the extent of the similarity across HI ears.

To investigate the possible sources of the common token-specific confusions that are observed across HI ears, we examine the confusions for NH ears when the speech is degraded by high/low-pass filtering or noise masking. The data from the NH high/low-pass filtering experiments shows the locations in frequency of conflicting cues, which can cause confusions when the primary cue region is masked or attenuated. The white noise (WN) masking data for NH listeners can show similar confusion to those of HI listeners at lower levels of SWN. This similarity across different types of noise masking may be due to the audiometric configurations of the HI ears in our study; all but one of the HI ears in the study have sloping high-frequency hearing loss, which more-severely attenuates consonant cues at the higher frequencies.

When testing HI ears, the selection of the tokens for a perceptual experiment is critically important. Multiple tokens of a single consonant, having acoustic cues that vary naturally in terms of intensity, frequency, and/or temporal cues, can result in different measures of hearing impairment. Each token of a consonant may then be considered as a sensitive probe that can provide fine-grained information about an individual's hearing impairment. The existing natural variability of speech may be used to advantage, but only once it has been controlled for.

CHAPTER 2

BACKGROUND: SPEECH PERCEPTION IN NH AND HI EARS

2.1 Necessary Cues for Consonant Perception

2.1.1 Development of the AI-Gram

Starting in the 1920s, Fletcher and his colleagues used masking noise along with high- and low-pass filtered high-entropy "nonsense" syllables to study the contribution of different frequency bands to speech perception, as a function of SNR (Fletcher and Galt 1950; French and Steinberg 1947; Allen 1994). These classic studies led to the articulation index (AI) model of speech intelligibility (American National Standards Institute 1969).

Based on the AI, Lobdell and Allen developed a computational model denoted the *AI-gram* that simulates the effect of noise masking on audibility (Lobdell 2009; Lobdell et al. 2011). The AI-gram is a time-domain implementation of the AI model of speech intelligibility and Fletcher's critical-band auditory model (i.e., Fletcher's SNR model of signal detection). Given a speech sound and masking noise, the AI-gram simulates the effect of noise masking and produces an image that predicts the audible speech components along the time and frequency axes.

2.1.2 3-Dimensional Deep Search (3DDS)

The objective of 3DDS is to measure the significance of speech subcomponents on perception in three dimensions: time, frequency and signal-to-noise ratio (SNR). In Miller and Nicely (1955); Wang and Bilger (1973); Allen (2005), noise masking was used to study consonant confusions. In 1986, Furui used time-truncation experiments to analyze the essential time-waveform components for speech perception. All of these techniques are merged for the

3DDS methodology, which uses three independent listening experiments and the AI-gram to evaluate the contribution of speech components to consonant perception.

To isolate the perceptual cue of a consonant-vowel token, the 3DDS method is composed of three independent psychoacoustic experiments that modify the speech as a function of time, frequency and SNR (see Fig. 2.1). The naming paradigm for each experiment (TR07, HL07, MN05) is set up such that the two-digit suffix indicates the year when the experiment was performed. The first experiment (TR07) uses truncation in order to find the location in time or minimum possible duration of the perceptual cue region (Li et al. 2010). The second experiment (HL07) is designed to isolate the perceptual cue region in frequency by high- or low-pass filtering the speech at 10 cutoff frequencies that span from 0.25–8 [kHz] (Li et al. 2010). A third experiment (MN05) assesses the masked threshold (i.e., perceptual robustness to noise) of the speech cue region, by masking the speech with WN at various SNRs (Phatak et al. 2008).

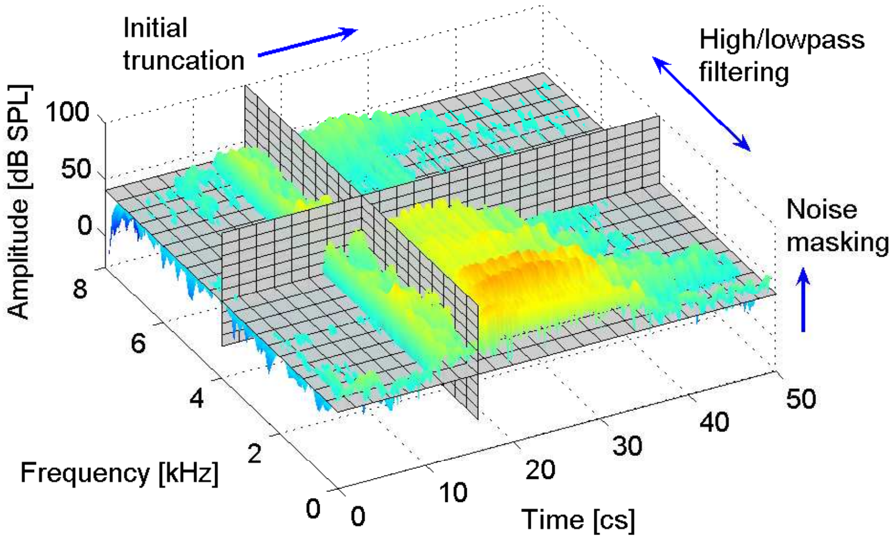


Figure 2.1: Schematic diagram of 3DDS to characterize the contribution of speech subcomponents to perception as a function of time, frequency and intensity (figure from Li et al. (2010)).

2.1.3 3DDS Stop Consonant Cue Findings

The 3DDS method has been used to explore the perceptual cues of stop consonants (Li et al. 2010). The time-frequency regions that contain the necessary cues for perception of /p, t, k, b, g, d/ are illustrated in Fig. 2.2. The intensity of the necessary cue region has been found to be correlated to the robustness to noise (Régnier and Allen 2008; Li et al. 2010). Natural fluctuations in the intensity of the stop consonant cue regions were shown to explain the large variations in the AI (Singh and Allen 2012).

It was discovered that natural speech sounds often contain *conflicting cue regions* that lead to confusions, when the target-consonant cue region is removed by filtering or masking noise. Through the manipulation of these spectral conflicting cue regions, one consonant can be *morphed* into another or a perceptually weak consonant can be converted into a strong one (Li and Allen 2011; Kapoor and Allen 2012).

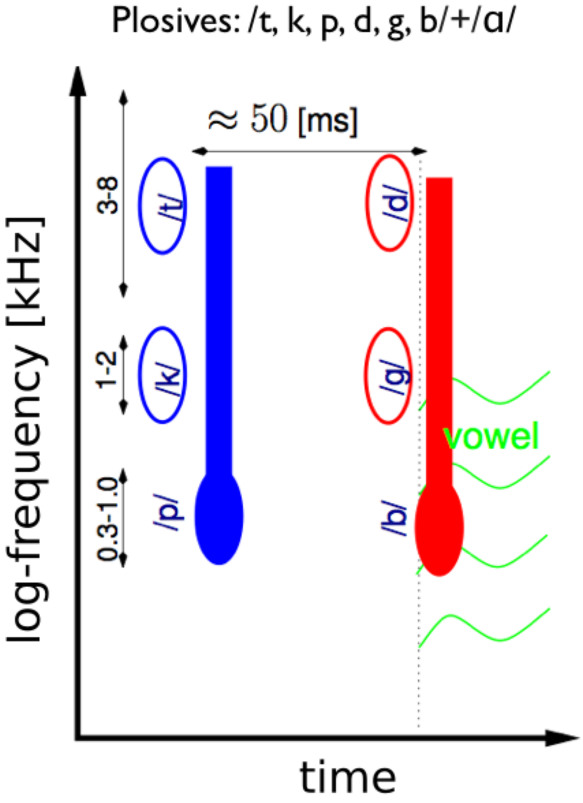


Figure 2.2: Cartoon displaying the time-frequency regions which contain the necessary consonant cues for perception of stop consonants (Li et al. 2010).

2.1.4 Past Studies on Fricative Cues

Fricative consonants are a major source of perceptual error under noisy conditions, thus they are of special interest. This is true for clearly articulated speech (Miller and Nicely 1955) as well as natural speech (Phatak et al. (2008), Fig. 1). These studies have shown that, at 12 [dB] SNR in WN, the non-sibilant labial and dental fricatives /f, v, θ, ð/ are involved in more than half of the confusions. In contrast, the sibilant alveolars /s, z/ and postalveolars /ʃ, ʒ/ are seldom confused with any other consonants at the same noise level.

Fricative consonants are produced by forcing air through a narrow constriction of the vocal tract above the glottis (Stevens et al. 1992). The study by Miller and Nicely (1955) observed that the frication noise of the voiced consonants is modulated by the fundamental frequency F0. Miller and Nicely (1955) also note that the frication regions of /s, ʃ, z, ʒ/ are of longer duration than /f, v, θ, ð/. A consistent difference between "voiced" and "unvoiced" fricatives is the presence of energy below 0.7 [kHz] (Hughes and Halle 1956) as well as the average duration of the frication (Baum and Blumstein 1987; Stevens et al. 1992). Stevens concluded that listeners based their voicing judgments of intervocalic fricatives on the time interval duration for which there was no glottal vibration. If this time interval was greater than 6 [cs], the fricative was typically judged as unvoiced (Stevens et al. 1992). When reporting time in our study, the unit centiseconds [cs] is used, as 1 [cs] is a natural time interval in speech perception. For example, an F0 of 100 [Hz] has a period of 1 [cs], while relevant perceptual times are typically $\geq$ 1 [cs]. The minimal duration of the frication noise is approximately 3 [cs] for /z/ and 5 [cs] for /f, s, v/ (Jongman 1989). The consonants /θ, ð/ are identified with reasonable accuracy only when at full duration (i.e., no time-truncation) (Jongman 1989). Although the mean duration of unvoiced fricatives is generally longer than that of the voiced fricatives, the distribution of the two overlap considerably (Baum and Blumstein 1987).

A number of studies have concluded that /s, z/ are characterized by a strong concentration of frication energy around 4–5 [kHz], while /ʃ, ʒ/, pronounced with a longer anterior cavity, have a spectral peak around 2–3 [kHz] (Miller and Nicely 1955; Hughes and Halle 1956; Heinz and Stevens 1961; Jongman et al. 2000). Harris (1958), used consonant-vowel tokens to investigate the relative importance of cues in the frication noise vs. the following vocalic portion. The tokens were modified by swapping the vocalic portions of different fricatives. Harris found that the cues that discriminated place of articulation for /s, ʃ, z, ʒ/ were in the frication noise portion, while the place of articulation of /f, θ/ was perceived based on the vocalic portion, although both were perceived as /θ/ when the frication noise was paired with the vocalic portion of /s, ʃ/. Similarly, /ð/ and /v/ were confused with each other when their vocalic portions were swapped. For the voiced fricatives, Harris (1958) notes that the segmentation of frication and vocalic portions may be imprecise, leading to variable results. A later study used hybrid speech to find that both frication noise and formant transitions can be used for the distinction of /s/ and /ʃ/ (Whalen 1991). Synthesized stimuli with resonant frequencies around 6.5 to 8 [kHz] usually yielded /f/ and /θ/ responses, with /f/ being distinguished from /θ/ on the basis of the second formant transition in the following vowel (Heinz and Stevens 1961). In contrast, analysis of natural speech from twenty talkers indicates that both /f, v/ and /θ, ð/ display a relatively flat spectrum without any dominating spectral peak (Jongman et al. 2000). In a consonant-vowel context, the effects of anticipatory coarticulation on the spectral characteristics of /ʃ, ʒ, s, z/ were smallest when the fricatives were followed by the vowel /ɑ/ (Soli 1981).

Other acoustic cues such as the amplitude of fricative noise (Behrens and Blumstein 1988; Hedrick and Ohde 1993) and spectral moments, including skewness and kurtosis, (Shadle and Mair 1996) have been shown to have minimal perceptual significance. Clearly articulated fricatives have perceptual cues shifted toward the high-frequency region (Maniwa et al. 2008).

To summarize the findings of these many past studies: For the sibilant fricatives /ʃ, ʒ, s, z/, the place of articulation is encoded by the spectral distribution of the frication noise. Naturally produced voicing can be identified by the presence of a low-frequency (<0.7 [kHz]) component, F0 modulations of the frication noise, as well as a longer original duration of frication. The relative perceptual roles of these three characteristics of voiced fricatives remains unclear. No conclusive picture is available for the non-sibilant fricatives /f, v, θ, ð/. These many findings result from studies conducted without noise.

2.2 Hearing-Impaired Consonant Perception

Consonants comprise approximately 58.5% of conversational speech (Mines et al. 1978). While the relative importance of consonants and vowels for HI speech perception remains uncertain (Hood and Poole 1977; Burkle et al. 2004), here, we concentrate on HI consonant perception. Many past works have examined HI consonant recognition using naturally produced speech, including Lawrence and Byers (1969); Bilger and Wang (1976); Owens (1978); Wang et al. (1978); Dubno and Dirks (1982); Boothroyd (1984); Fabry and Van Tasell (1986); Dreschler (1986); Gordon-Salant (1987); Zurek and Delhorne (1987). Overall, the effects of hearing impairment on speech perception are more severe in the presence of noise (Dubno and Dirks 1982; Dreschler 1986). In these past studies, data analysis is performed using either an average measure (over all consonants) or with consonants grouped by distinctive features. Speech measures derived from an average are useful tools for screening and classifying those with a hearing impairment, however, they have not proven useful as detailed prescriptive measures (Taylor 2006; Killion and Gudmundsen 2005).

Owens (1978) developed the *California Confusion Test* to examine the consonant confusions of HI listeners, using speech in a CVC context and in only the quiet condition. This test presents a single CVC with four multiple-choice response options. One of the findings during the development of this test is that the listeners would respond differently to different re-recordings of the same CVCs, this supported the findings of Kreul et al. (1969) which found that "only the actual recordings of the spoken lists ... can be considered to be test material". The results of the Confusion Test showed that, similar to NH listeners, the consonant-specific confusion groups of the HI listeners contained only two or three other consonants, although this may have been influenced by the limited, multiple-choice nature of the task. In addition, the confusion groups were similar across a wide array of pure-tone configurations.

Dubno and Dirks (1982) examined the consonant confusions of HI listeners with flat, gradual, and steeply sloping hearing loss. For the analysis, the responses of the ears in each hearing loss group were averaged together, despite a 20 dB standard deviation in the hearing loss within each of the groups. The consonants are presented in a Consonant-Vowel (CV) and Vowel-Consonant context at 90 dB SPL, with a 20 dB SNR. For analysis, the consonants are grouped by voicing, manner, place and by vowel; in general, the strongest effect is that steep-sloping loss has the most error across the majority of groupings. The ears with gradually sloping loss have a slightly better performance than those with flat loss. A detailed analysis of the consonant confusions shows that certain confusions are most common for certain audiometric groups. There is no analysis of the individual HI ears, so it is unknown if some of the higher probabilities are due to one or two ears with highly consistent confusions. Notably, all confusion groups seemed to be common to the three types of hearing loss, with different confusion weightings for each type.

Dubno et al. (1984) compared mild hearing impairment in subjects separated by age groups. Two groups of NH listeners and two groups of HI listeners were separated into groups of <44 years and >65 years in age. The NH listeners all had thresholds below 20 [dB] between 0.5–4 [kHz] with error bars of only 5 [dB], indicating a homogeneity across the listeners. The HI listeners, on the other hand, all had thresholds above 20 [dB] from 0.5–4 [kHz] and the individual listener thresholds varied across a range of at least 20 [dB] in terms of standard error, indicating a wide amount of variability among the HI subjects. All HI listeners in each age category were treated as a homogenous group for the analysis (i.e., averaged together). The experiment measured the signal-to-babble ratio necessary for 50% performance on three speech-recognition tasks, at three different speech levels. For spondee recognition, a difference of <3 dB was found between the NH and HI groups of both age categories. For high-predictability sentences, this difference remained small, reaching a maximum of 5 [dB] at the lowest (56 [dB] sound pressure level (SPL)) speech level. Only when low-predictability sentences were used does the difference in necessary signal-to-babble ratio for 50% performance between the NH and HI listeners become as large as 10 [dB] at the lowest speech level. The AI was then measured and compared to the results; the AI was found to not correlate with the signal-to-babble ratio necessary for 50% speech perception. A number of significant conclusions are drawn from this data. First, since the average audiograms of the young and elderly listeners were “identical” and yet a clear age disadvantage for noisy speech processing was observed for the elderly, then the audiometric findings did not fully predict the speech performance in noise. These age effects were observed for both the NH and HI groups, thus an elderly listener with a normal audiogram could still have difficulty perceiving speech in noise. Age effects were not observed in the quiet conditions, highlighting the necessity of using noise when investigating hearing impairment. The low-context stimuli best captured the processing difficulties of both the HI and NH elderly subjects. The deficit in performance of both the HI and NH elderly groups was observed even at the highest (88 [dB]) speech levels, suggesting that there is some "distortion"/suprathreshold effect that cannot be accounted for by audibility. The results of Patterson et al. (1982) support that this distortion is caused at the periphery and is not due to a central processing disfunction.

Gordon-Salant (1987) looked at the consonant confusions of elderly NH and HI listeners with both gradual and sharply sloping hearing loss. All speech was presented at 6 dB signal-to-babble ratio noise and in a CV format. The speech stimuli were broken into groups based on the vowel, manner, place, voicing, or level; significant differences were tested for NH, HI gradual-sloping, and HI steeply sloping loss. In general, the HI listeners had worse consonant perception performance than the NH listeners. In some of the cases the gradual-sloping loss corresponded to better performance than the steeply sloping loss. An INDSCAL analysis showed the common consonant confusions across a number of dimensions, including manner, place and voicing. Examples of common consonant confusion groups were /m, n/ and /f, v/. An ANOVA of the INDSCAL weightings showed that none of the weightings for the three subject groups were significantly different, indicating that the same confusion groups were observed across all listeners. These findings contrast in some respect with the results of Dubno and Dirks (1982), which showed differences in the nonsense syllable perception of gradual vs. steeply sloping audiometric configurations; one possible explanation is the inclusion of young listeners in the Dubno and Dirks (1982) study.

Comparisons between the consonant recognition errors of HI listeners vs. NH listeners with simulated hearing losses (noise and/or filtering applied) has shown some agreement in both errors and confusions (Wang et al. 1978; Fabry and Van Tasell 1986; Zurek and Delhorne 1987). Zurek and Delhorne (1987) tested the consonant perception of both HI and noise-masked NH listeners. The noise was shaped for the NH listeners such that their noise-masked thresholds matched the pure-tone thresholds of individual HI ears. The matching was implemented over the range of 0.125–4 [kHz], hearing loss over 4 [kHz] was not considered. All of the HI ears had moderate to severe hearing loss, with thresholds that reached 70 [dB] within the 0.125–4 [kHz] frequency range. When this noise matching was implemented, the average CV score, over 72 test tokens, approximated the perception of the corresponding HI ears. The majority of HI ears had a <70% probability correct at even the quiet condition, with four out of six audiometric configurations showing an average performance <50% in the quiet condition. These results showed that matching NH ears to HI audiometric measures can result in a similar degrees of averaged error.

CHAPTER 3

FRICATIVE CONSONANT CUES

3.1 Methods

The basic methodologies of the three perceptual experiments are given next. For additional detail about the experimental methods, refer to Li et al. (2010); Phatak et al. (2008).

3.1.1 Subjects

In total, 48 listeners were enrolled over three studies, of which 12 participated in experiment HL07, 12 participated in experiment TR07 (one participated in both), and 24 participated in experiment MN05. All listeners self-reported no history of speech or hearing disorder. To guarantee that no listeners with hearing loss or other problems were included in this study, any listener with significantly lower performance was excluded from further testing (see Phatak et al. (2008) for details). The first or primary language of all of the listeners is American English, with all but two having been born in the United States. No significant differences were observed in the consonant scores or confusions of these two listeners, and hence their responses were included. Listeners were paid for their participation. International Review Board approval was obtained prior to the experiment.

3.1.2 Speech Stimuli

As in the study by Miller and Nicely (1955), sixteen isolated consonant-vowels (CV)s: /p, t, k, b, d, g, s, ʃ, f, v, θ, ð, z, ʒ, m, n/+/ɑ/ (no carrier phrase) were chosen from the University of Pennsylvania's Linguistic Data Consortium database (LDC-2005S22, a.k.a. *the Fletcher AI corpus*) as the common test material for the three experiments. The speech sounds were sampled at 16 [kHz]. Experiment MN05 used 18 talkers for each CV. When extending the psychoacoustic database with HL07 and TR07, 6 tokens per CV (half male and half female) were selected. In order to explore how cue characteristics contribute to noise robustness, the 6 tokens were selected such that they were representative of the CVs in terms of confusion patterns and score, based on the results of MN05. Specifically, one-third high-scoring tokens were selected, and one-third low-scoring tokens were selected. Thus, a total of 96 tokens were used (16 CVs × 6 tokens per CV), 48 of which were fricatives and are reported on in this chapter. Sounds were presented diotically (both ears) through Sennheiser HD-280 PRO circumaural headphones, adjusted in level at the listener's *Most Comfortable Level* (MCL) for CV tokens in 12 [dB] of WN, i.e., $\approx$ 70–75 [dB] SPL. Subjects were allowed to change the sound intensity during the experiment, which was noted in the log files. All experiments were conducted in a single-walled Industrial Acoustics Company sound-proof booth, situated in a lab with no windows, with the lab outer door shut.

3.1.3 Conditions

Experiment TR07 assesses the temporal distribution of speech information (Li et al. 2010). For each token, the initial truncation time is set before the beginning of the consonant and the final truncation time is set after the end of the consonant-vowel transition. The truncation times were chosen such that the duration of the consonant was divided into frames of 0.5, 1, or 2 [cs]. An adaptive strategy was adopted for the calculation of the sample points. The basic idea is to assign more points where the speech perception scores change rapidly (Furui 1986). Starting from the end of the consonant-vowel transition and moving backward in time, the scheme allocates eight truncation times (frames) of 0.5 [cs], then twelve frames of 1 [cs], and finally as many 2 [cs] frames as needed until the onset of the consonant is reached. White noise was added following truncation at an SNR of 12 [dB] (based on the unmodified speech sound), matching the control condition of the filtering experiment (HL07).

Experiment HL07 investigates the distribution of speech information in frequency. It is composed of 19 filtering conditions, namely one full-band condition (0.25–8 [kHz]), nine high-pass and nine low-pass conditions. The full-band frequency range was divided into 12 bands, each having an equal distance along the human basilar membrane. The cutoff frequencies were calculated using Greenwood's inverse cochlear map function. The common high- and low-pass cutoff frequencies were 3678, 2826, 2164, 1649, 1250, 939, and 697 [Hz]. To this we added the cutoff frequencies 6185, 4775 (high-pass) and 509, 363 [Hz] (low-pass). All speech samples were high-pass filtered above 250 [Hz] based on estimates of the frequency importance region observed by Fletcher (Allen 1994). The filters were implemented as sixth-order elliptic filters having a stop-band attenuation of 60 [dB]. White noise (12 [dB] SNR) was added to the modified speech in order to mask out any residual cues that might still be audible. Note that for most CVs, 12 [dB] SNR does not reduce the probability of correct perception (Phatak et al. 2008).

Experiment MN05 (a.k.a. MN16R) measures the strength of the perceptual cue in terms of robustness to white masking noise. Besides a quiet condition, speech sounds were masked at eight different SNRs: -21, -18, -15, -12, -6, 0, 6, and 12 [dB] (Phatak et al. 2008).

We define the probability of correct detection of the target consonant as P_c. The cutoff frequencies of experiment HL07 are denoted f^H (high-pass) and f^L (low-pass). The SNR_{90} is defined as the SNR at which the target consonant has a probability of correct detection of 90% ($\mathrm{P}_c(\mathrm{SNR}) = 0.9$).

All three experiments include a common control condition, i.e., full-bandwidth, full-duration speech at 12 [dB] SNR. The recognition scores for this control condition were verified to be consistent across the three experiments.

3.1.4 Experimental Procedure

The three experiments (TR07, HL07, MN05) used nearly identical experimental procedures. A MATLAB® program was written for the stimulus presentation and data collection. A mandatory practice session, with feedback, was given at the beginning of each experiment. Speech tokens were randomized across talkers, conditions and tokens. Following each stimulus presentation, listeners responded by clicking on the button that was labeled with the CV that they perceived. In case the CV was completely masked by the noise, the listener was instructed to click a “Noise Only” button. Frequent (e.g., 20 minute) breaks were encouraged to prevent test fatigue. Subjects were allowed to repeat each token up to three times. The waveform was played via a SoundBlaster 24 bit sound card in a PC Intel computer, running MATLAB® via Ubuntu Linux.

3.1.5 3DDS Procedure

Each of the three experiments provides estimates of different aspects of the necessary perceptual cue region: the critical temporal information, the frequency range, and the intensity. As the listener P_c curves are roughly monotonic (with small amounts of jitter due to random listener error), linear interpolation was used in the analysis of the MN05 and TR07 data. The minimum duration of frication or location in time of the perceptual cue is determined by the truncation time at which the P_c drops below a threshold of 90%. The perceptual cue robustness to noise is defined by the SNR at which the P_c falls below the 90% threshold (SNR_{90}).

For the high- and low-pass filtering experiments, the upper and lower frequency boundaries of the perceptual cue region are determined from the two frequencies at which the P_c drops below a threshold of 75%. This lower threshold was chosen due to the low number of trials in HL07 (N = 12), requiring a lower threshold for significant errors. In addition, probit fits for each token data set were calculated using the glmfit() MATLAB function, in order to provide a better estimate of the threshold frequencies. A critical band set the limit for the minimum bandwidth recorded.

When the speech token has a low P_c even in the quiet, full-band, and full-duration condition, a cue region cannot be isolated by the 3DDS method since the listeners will not show correct perception at any condition.

3.2 Results

Next, we demonstrate how the perceptual cues of fricative consonants are isolated by the 3DDS method. For each consonant, a single representative token (out of the six tokens) for each CV is presented in a figure and analyzed in detail. The results of experiments TR07, HL07, and MN05 are depicted as *confusion patterns* (CP)s; a CP displays the probability of hearing all possible responses r, given the spoken consonant s, as the conditions for a single experiment vary (Allen 2005). More precisely, the CPs $\mathrm{P}_{r|s}(t)$, $\mathrm{P}_{r|s}(f)$, and $\mathrm{P}_{r|s}(\mathrm{SNR})$ are shown for experiments TR07, HL07, and MN05, respectively, in Figs. 3.1–3.3. Confusions with probability <0.2 and "Noise Only" responses are not shown in the CPs in order to more clearly display the primary confusions.

The figures are organized into three pairs /ʃ, ʒ/, /s, z/, and /f, v/ (Fig. 3.1–3.3) to highlight both the similarities and differences. The paired unvoiced-voiced fricatives are displayed in subfigures (a) and (b), respectively. Each of the subfigures contains five panels labeled with a boxed number in the upper left or right corner. Panel 1 shows the AI-gram of the full-bandwidth, full-duration CV at 18 [dB] SNR with the 3DDS-isolated spectral cue region highlighted by a small rectangular box; the range of truncation times is marked by a large frame. Panel 2, aligned with panel 1 along the time axes, shows the listener CP as a function of truncation time, a star and vertical dotted line marks the truncation time at which P_c drops below 90%, a second vertical dotted line marks the end of the frication region. The line marked with "a" in panel 2 shows the probability of responding "Vowel Only". Panel 3, rotated by 90° clockwise and aligned with panel 1 along the frequency axes, illustrates the CP when listeners hear a high- or low-pass filtered sound as a function of the nine cutoff frequencies, with dashed lines indicating high-pass responses and solid lines indicating low-pass responses. Panel 4 shows the CPs when the token is masked by WN at six different SNRs, with the SNR_{90} marked by a star. The AI-grams, at each tested SNR, are displayed in panel 5. For the AI-grams of panel 5, only the region within the range of truncation times is shown.

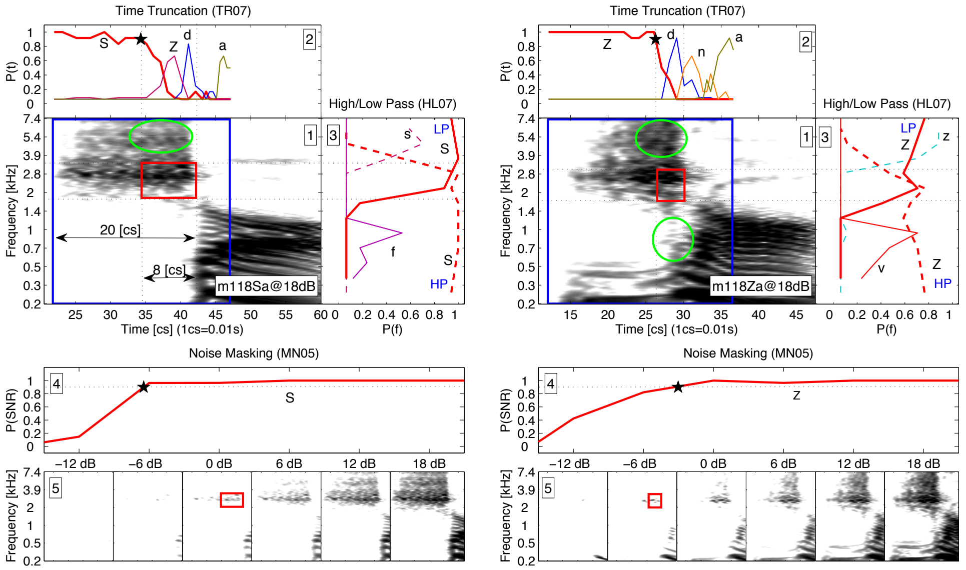
Figures are referenced by the figure number, subplot letter, and panel number; for example, Fig. 3.1 a.5 is a reference to Fig. 3.1, subplot (a), panel 5.

3.2.1 /ʃɑ/ and /ʒɑ/

The results of the perceptual experiments for /ʃɑ/ and /ʒɑ/, for talker m118, are shown in Fig. 3.1.

The AI-gram for /ʃɑ/ from talker m118 (Fig. 3.1 a.1) shows a wide-bandwidth, sustained frication noise above 2 [kHz] in the consonant region. The truncation experiment (TR07, Fig. 3.1 a.2) shows that when the duration of the frication is truncated from the original ≈20 [cs] to ≈8 [cs], the listeners begin to show confusions with /ʒɑ/. Once that the frication region is truncated to ≈1 [cs] (a burst), the listeners report /dɑ/ 80% of the time. The results from the filtering experiment (HL07, Fig. 3.1 a.3) show that the perceptual cue region lies between 1.7–3.6 [kHz]. When the token is low-pass filtered below f^L <1 [kHz], confusions with /fɑ/ emerge. High-pass filtering above f^H >3.9 [kHz] causes confusions with /sɑ/. The low-pass filtering responses indicate that this energy above 3.9 [kHz] is unnecessary for correct perception, indicating that this high-frequency frication contains conflicting cues. The noise masking experiment (MN05, Fig. 3.1 a.4) shows a sharp drop in the P_c, from 96% to 15% between -6 and -12 [dB] SNR. Examining the AI-gram across SNRs (Fig. 3.1 a.5), we see a predicted loss of audibility of the isolated cue region between 0 and -6 [dB] SNR.

For the voiced /ʒɑ/ from talker m118, the AI-gram (Fig. 3.1 b.1) for the consonant region contains a wide-bandwidth, sustained frication region above 1.6 [kHz] (similar to /ʃɑ/) along with a coincident voicing below 0.6 [kHz]. The truncation experiment (TR07, Fig. 3.1 b.2) shows that the original ≈16 [cs] duration of the voiced frication can be shortened to ≈3.5 [cs] before listeners report non-target consonants. When the voiced frication is truncated to <2 [cs], listeners primarily report /dɑ/. Once the frication region is removed by truncation, leaving just the vowel onset, listeners primarily report /nɑ/ and do not report "Vowel Only" until the vowel onset is also removed. The results of filtering (HL07, Fig. 3.1 b.3) show that the necessary perceptual cue region for /ʒɑ/ lies between 1.7–2.9 [kHz]. When the primary cue region is removed by a low-pass filter at f^L ≈1 [kHz], listeners primarily report /vɑ/. When the token is filtered above f^H > 3.2 [kHz], listeners primarily report /zɑ/. The frication region above 3.2 [kHz] contains conflicting cues for /zɑ/, the P_c does not significantly change when this high-frequency frication noise is removed by low-pass filtering. The



(a)

(b)

Figure 3.1: Isolated perceptual cue regions for /ʃɑ/ and /ʒɑ/, denoted by S and Z respectively. Each token has five numbered panels: (1) the AI-gram with the highlighted perceptual cue (rectangle), hypothetical conflicting cue region (ellipse), and the range of truncation times (large frame); (2) CP of TR07, aligned to panel 1 along the time axes, "a" represents the IPA vowel /ɑ/; (3) CP of HL07, rotated by 90° clockwise and aligned to panel 1 along the frequency axes; (4) CP of MN05; and (5) AI-grams across the different tested SNRs (large frame region) showing the change of speech audibility as noise level increases.

results of the noise masking experiment (MN05, Fig. 3.1 b.4) show that the P_c begins to drop at 0 [dB] SNR. The AI-gram across SNRs (Fig. 3.1 b.5) predicts full loss of audibility of the identified frication region below -6 [dB] SNR; correspondingly, the P_c falls from 82% at -6 [dB] SNR to 42% at -12 [dB] SNR.

The lower HL07 performance for /ʒɑ/ from talker m118 at the full-band case ($P_c = 80\%$ in HL07 vs. 100% at the corresponding MN05 and TR07 control conditions for different listener populations) is due to a percentage of listeners (3 out of 17) that had consistent difficulty discriminating /ʒɑ/ across all of the tokens, despite little to no error in the full-band perception of other high to medium scoring fricatives in our data set. For these listeners, since /ʒ/ does not have a dedicated representation in written English (e.g., "sh" for /ʃ/), we conjecture that these low HL07 scores are the result of an insufficient practice session.

Somewhat unexpectedly, the perception of the target sound /ʒɑ/ does not become confused with /ʃɑ/ when the low-frequency (< 0.7 [kHz]) voicing is removed by a high-pass filter (Fig. 3.1 b.3). Further analysis revealed that the frication regions of voiced sibilants are amplitude modulated by F0, the fundamental frequency of vocal vibration, retaining a sufficient amount of voicing information for correct perception even when the low-frequency voicing is removed (Fig. 3.1 b.3). A detailed summary on the perceptual cues for discrimination of voicing can be found in the Section 3.3.

Other /ʃɑ, ʒɑ/ tokens: 3DDS analysis of the five other /ʃɑ/ tokens (talkers m115, m111, f103, f109, f106) show similar cue region results to those for the token from talker m118. The truncation data shows that the P_c drops below 90% once that the consonant frication regions are truncated to a duration of < 9 [cs], on average. The frequency range of the cue region was found to be between 1.6–3.6 [kHz] for the total group of tokens, with variability within that range across different talkers. All six /ʃɑ/ tokens resulted in a P_c > 90% for the noise masking experiment at 12 [dB] SNR, with a perceptual robustness to noise that varied from -7 to 5 [dB], as quantified by the SNR_{90}.

Of the remaining five /ʒɑ/ tokens, four (talkers f103, m107, m114, and m117) show cue region results similar to the token from talker m118. The truncation results show that the consonant is perceived correctly until the frication region is shortened to <3.5 [cs], on average. The overall frequency range of the cue region lies between 1.3–3.6 [kHz], with variability within this range across talkers. The noise masking experiment shows that the robustness to noise of the perceptual cues, as measured by the SNR_{90}, falls in a 12 [dB] range, from -7 to 5 [dB] SNR. One /ʒɑ/ token (talker f108) has a P_c of only 40% at the quiet, full-band, full-duration condition. Analysis of the AI-gram of this token revealed a frication with similar frequency distribution and duration to the other /ʒɑ/ tokens, but with a barely audible low-frequency voicing (as estimated by the AI-gram), indicating weak modulation of the high-frequency frication noise and predicting a high likelihood of confusion with /ʃɑ/; these voicing confusions are observed in the listener responses (50% /ʃɑ/ in quiet).

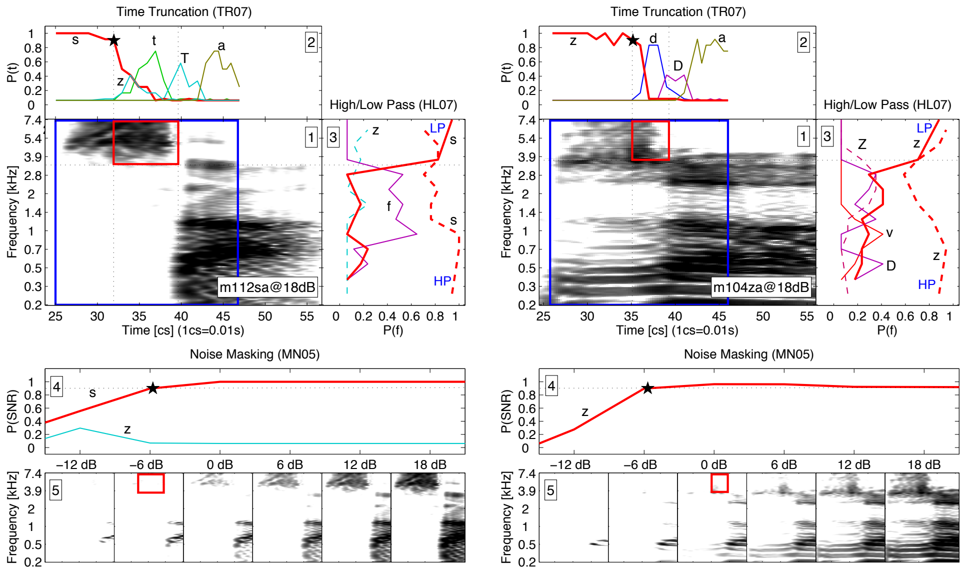
Summary cue region results are provided in Table 3.1 and illustrated in Fig. 3.4.

3.2.2 /sɑ/ and /zɑ/

The results of the perceptual experiments for /sɑ/ (talker m112) and /zɑ/ (talker m104) are shown in Fig. 3.2.

The AI-gram for /sɑ/ from talker m112 (Fig. 3.2 a.1) displays a high-frequency, sustained frication noise above 3.2 [kHz] in the consonant region. The truncation results (Fig. 3.2 a.2) show that once the duration of the frication is truncated from the original ≈14 [cs] to <7.5 [cs], the P_c begins to drop, with a high percentage of listeners reporting /tɑ/ once the frication is shortened below 4 [cs]. The results of the filtering experiment (Fig. 3.2 a.3) show that the cue region lies above 3.7 [kHz]. Low-pass filtering of the token at $0.9 \leq f^L \leq 2.8$ causes confusions with /fɑ/. The noise masking experiment (Fig. 3.2 a.4) shows a drop in the P_c after 0 [dB] SNR, with an SNR_{90} of approximately -6 [dB] SNR. This is consistent with the AI-grams across SNRs (Fig. 3.2 a.5), which predict a faint but still audible frication within the 3DDS-isolated cue region at -6 [dB] SNR.

For the voiced /zɑ/ from talker m104, the AI-gram (Fig. 3.2 b.1) displays a sustained frication region above 2.3 [kHz] and coincident voicing mainly below 0.7 [kHz], in the consonant region. The time truncation results (Fig. 3.2 b.2) show that once the duration of the frication is truncated from the original ≈14.5 [cs] to ≤4 [cs], listeners begin to report non-target consonants. Once the frication is truncated to ≤3 [cs], listeners primarily report /dɑ/. The results of the filtering experiment (Fig. 3.2 b.3) show that the cue region for this /zɑ/ token falls between 3.8–8 [kHz]. No strong confusions emerge when the spectral cue region is removed by filtering, instead the listeners chose the "Noise Only" response. The noise masking experiment (Fig. 3.2 b.4) shows an abrupt drop in the P_c below -6 [dB] SNR. The AI-grams across SNRs (Fig. 3.2 b.5) predict that a small amount of the isolated cue region is still audible at 0 [dB] SNR and is completely masked by noise at -12 [dB] SNR.



(a)

(b)

Figure 3.2: Isolated perceptual cue regions for /sɑ/ and /zɑ/. Each token has five numbered panels: (1) the AI-gram with the highlighted perceptual cue (rectangle) and the range of truncation times (large frame); (2) CP of TR07, aligned to panel 1 along the time axes; (3) CP of HL07, rotated by 90° clockwise and aligned to panel 1 along the frequency axes; (4) CP of MN05; and (5) AI-grams across the different tested SNRs (large frame region) showing the change of speech audibility as noise level increases. S, Z, T, D and "a" represent the IPA symbols /ʃ, ʒ, θ, ð, ɑ/ respectively.

Other /sɑ, zɑ/ tokens: Of the remaining five /sɑ/ tokens, three of them (talkers f108, f109 and f113) have similar cue regions as the token from talker m112. The time truncation data for these tokens showed a drop in the P_c once the frication region was truncated below 9 ± 1.5 [cs]. The frequency ranges of the perceptual cue regions are between 3.7–8 [kHz], with

Table 3.1: Summary of 3DDS results for /ʃa, ʒa, sa, za/. Minimum duration defined from the truncation time at which perception drops below 90% to end of frication. Tokens ordered based on SNR_{90} value. Only tokens with conclusive 3DDS estimates from all three experiments are listed.

CV	Talker	Min Dur [cs]	Freq [kHz]	SNR_{90}
/ʃɑ/	m118	8	1.7–3.6	-7
	m115	9	1.6–2.2	-2
	m111	8.5	1.7–2.6	-1
	f103	8	1.7–2.9	-1
	f109	11	2.2–2.8	0
	f106	9	1.7–2.5	5
/ʒɑ/	f103	4	1.9–3.6	-7
	m118	3.5	1.7–2.9	-3
	m114	3	1.3–2.7	-1
	m117	3	2.2–2.8	3
	m107	2.5	1.6–2.8	5
/sɑ/	f109	7.5	5.8–8.0	-10
	m112	7.5	3.7–8.0	-6
	f113	10.5	4.9–8.0	-5
	f108	9	4.2–8.0	-2
/zɑ/	m120	5	3.3–8.0	-10
	f105	4	5.3–8.0	-7
	m104	4	3.8–8.0	-6
	f108	4.5	2.0–8.0	-1
	m118	5.5	4.4–8.0	-1

token-specific variability within this region. The SNR_{90} measurements, fall across an 8 [dB] range, from -10 to -2 [dB] SNR. The remaining two /sɑ/ tokens (talkers m111 and m117), selected for their low perceptual scores, were primarily confused with /zɑ/ in quiet. Further analysis showed barely audible voicing cues in the token from talker m111 and a low-level, short-duration (4 [cs]) frication from talker m117.

Of the five remaining /zɑ/ tokens, four (talkers f105, f108, m118, and m120) contain similar perceptual cue regions to those of talker m104. The truncation data for these tokens showed a sharp drop in the P_c when the frication was shortened to 5 ± 1 [cs] and the frequency ranges of the cue regions fell between 2–8 [kHz]. The noise masking experiment resulted in SNR_{90} measurements within a 9 [dB] range, from -10 to -1 [dB] SNR. One token (talker f109) showed strong (≈50%) confusions with /ðɑ/ in quiet, yet the P_c rose to 100% once the first quarter of the consonant region was removed by truncation. Further investigation showed an energy burst before the onset of frication, creating a conflicting cue region, which led to this /ðɑ/ confusion. This conflicting cue region led to inconclusive filtering and noise-masking results for this token.

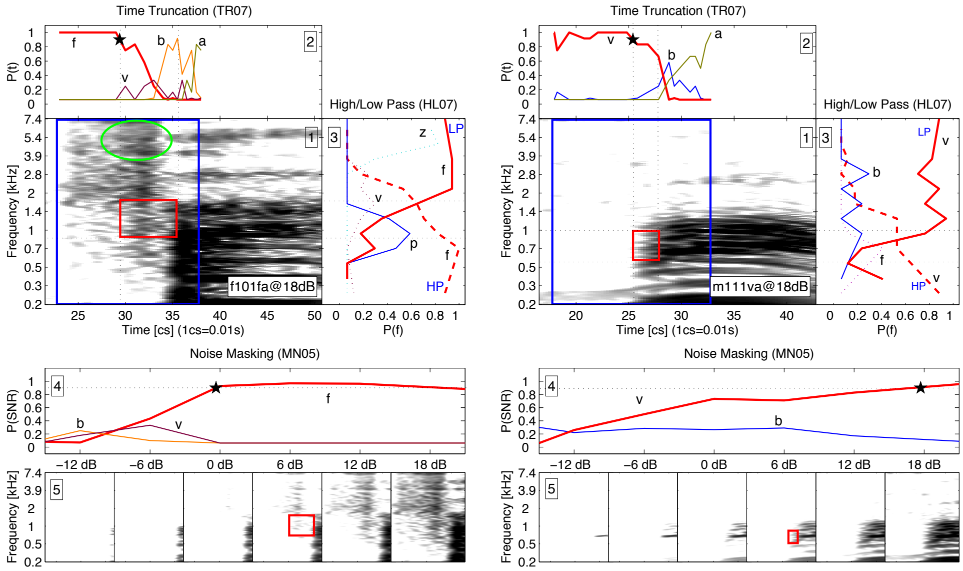
Summary cue region results are provided in Table 3.1 and illustrated in Fig. 3.4.

3.2.3 /fɑ/ and /vɑ/

The results of the perceptual experiments for /fɑ/ (talker f101) and /vɑ/ (talker m111) are shown in Fig. 3.3.

The AI-gram for /fɑ/ from talker f101 (Fig. 3.3 a.1) displays a wide-band, sustained frication noise that spans 0.3–7.4 [kHz], in the consonant region. The truncation results (Fig. 3.3 a.2) show that once the frication is shortened from the original ≈12 [cs] to <6.5 [cs], listeners report non-target consonants (primarily /vɑ/). Truncating the entire frication region, while leaving the vowel onset intact, results in a large proportion of /bɑ/ responses (>80%). The filtering experiment (Fig. 3.3 a.3) shows that the frequency range of the cue region, despite the wide-bandwidth of the full frication region, lies between 0.7–1.7 [kHz]. High-pass filtering at $f^H \geq 3.9$ [kHz] causes listeners to primarily report /zɑ/, indicating that the frication energy above this frequency contains a conflicting cue. Low-pass filtering at $0.7 \leq f^L \leq 1.3$ [kHz] results in listeners reporting /pɑ/, indicating that the full frequency range of the cue region is necessary for perception of this /fɑ/ token. The noise masking experiment (Fig. 3.3 a.4) shows an SNR_{90} of -1 [dB]. The AI-grams across SNRs (Fig. 3.3 a.5) predict a loss of audibility between 6 and 0 [dB] SNR.

For the voiced /vɑ/ from talker m111, the AI-gram (Fig. 3.3 b.1) displays a faint wide-band frication and a coincident low-frequency (<0.3 [kHz]) voicing in the consonant region. The vertical dotted line marking the end of the frication region (at ≈ 27 [cs]) indicates that the frication is briefly sustained into the onset of the vowel (as determined from the time waveform). The time truncation results (Fig. 3.3 b.2) show that once that the frication region is shortened from the original ≈11 [cs] to <2 [cs], listeners report confusions with /bɑ/. The results of the filtering experiment (Fig. 3.3 b.3) show that the perceptual cue region lies between 0.6–0.9 [kHz]. Filtering out the isolated cue region leads to “Noise Only” responses. The noise masking experiment (Fig. 3.3 b.4) suggests that the perceptual cue for this token can be masked by low levels (12 [dB] SNR) of WN. The AI-gram across SNRs (Fig. 3.3 b.5) predicts that some of the isolated cue remains audible up to 6 [dB] SNR but is shortened in duration to < 2 [cs]. In agreement with the results of the truncation experiment, the primary confusion at this 6 [dB] SNR condition is with /bɑ/.



(a)

(b)

Figure 3.3: Isolated perceptual cue regions for /fɑ/ and /vɑ/. Each token has five numbered panels: (1) the AI-gram with the highlighted perceptual cue (rectangle), hypothetical conflicting cue region (ellipse), and the range of truncation times (large frame); (2) CP of TR07, aligned to panel 1 along the time axes, “a” represents the IPA symbol /ɑ/; (3) CP of HL07, rotated by 90° clockwise and aligned to panel 1 along the frequency axes; (4) CP of MN05; and (5) AI-grams across the different tested SNRs (large frame region) showing the change of speech audibility as noise level increases.

Other /fɑ, vɑ/ tokens: The /fɑ/ token from talker m111 contains a similar spectral cue region to that of talker f101. A wide-band frication in the slightly lower frequency range of 0.6–0.9 [kHz], a minimum frication duration for correct recognition of 4.5 [cs], and an SNR$_{90}$ of 10 [dB] summarize the

3DDS findings for the cue region of this token. The low-performance /fɑ/ token from talker m117 showed a $P_c < 90\%$ correct even in the quiet, unmodified condition, leading to inconclusive 3DDS results. The three remaining low- and medium-performance /fɑ/ tokens (talkers f103, f105, and m112) have low-level but audible frication regions. These low-level but audible regions were erroneously removed in the preparation of the waveforms for the experiments, leading to a $P_c < 90\%$ at the control condition. As a result, we were unable to isolate the perceptual cues for these three /fɑ/ tokens. We only present results unaffected by this inappropriate consonant signal processing.

The /vɑ/ token from talker f108 is defined by a spectral cue region that is almost the same as the one observed for talker m111. This token has a minimum frication duration of 1.5 [cs] for correct perception, a frequency range of 0.7–0.95 [kHz], and an SNR_{90} of 3 [dB] SNR. Of the remaining /vɑ/ tokens, three (talkers f105, m104, and m120) are composed of consonant cues that are partially masked by the 12 [dB] of noise used at the control condition, leading to inconclusive 3DDS results. One mislabeled /vɑ/ token in the data set (talker f103) is primarily reported as a /fɑ/ in quiet and low noise levels.

Summary cue region results are provided in Table 3.2 and illustrated in Fig. 3.4. Since 3DDS results are available for only two tokens of /vɑ, fɑ/, these results are not considered to be widely generalizable.

Table 3.2: Summary of 3DDS results for /fɑ,vɑ/. Minimum duration defined from the truncation time at which perception drops below 90% to end of frication.

CV	Talker	Min Dur [cs]	Freq [kHz]	SNR_{90}
/fɑ/	f101	6.5	0.9–1.7	-1
	m111	4.5	0.6–0.9	10
/vɑ/	f108	1.5	0.7–0.95	3
	m111	2	0.6–0.9	18

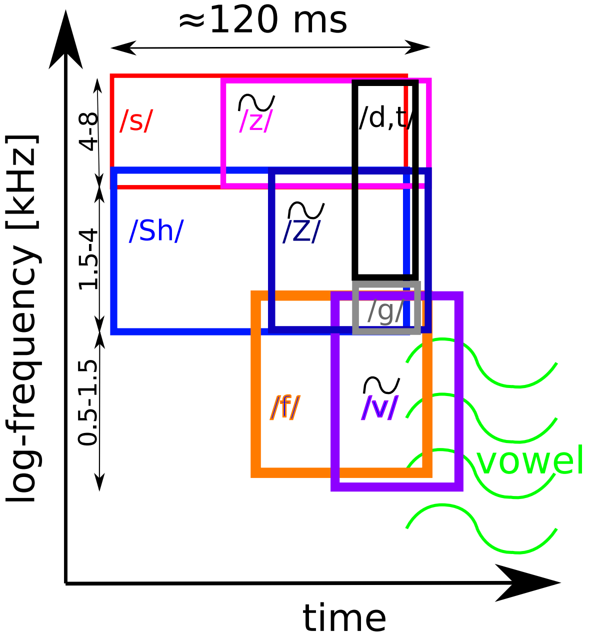


Figure 3.4: Cartoon displaying the time-frequency regions which contain the necessary consonant cues for perception of fricative consonants. Regions are determined from the data in Tables 3.1 and 3.2. A tilde "∼" indicates that the frication noise is modulated by F0. Cue regions for stop consonants with similar spectral shapes (/t, d, g/) are included for reference.

3.3 Discussion and Conclusions

This study generalizes the 3DDS psychoacoustic method to fricative American English consonants. This method allows us to examine the effects of highly variable natural speech components on human perception. We have also identified several natural confusions that are observed for these fricatives under different modification conditions.

3.3.1 Discriminating Cues for the Place of Articulation

For all of the fricatives in this study, the 3DDS-isolated cue regions were within the frication region. This is consistent with the observations of Harris (1958) for /s, ʃ, z, ʒ/+/i, e, o, u/. The alveolar consonants /sɑ, zɑ/ have isolated cue regions in the sustained frication, no lower than 2 [kHz]. The palato-alveolar consonants /ʃɑ, ʒɑ/ have isolated cue regions in the sustained frication between 1.3–3.6 [kHz]. For the non-sibilant labiodentals /fɑ, vɑ/, a band of frication between the frequency range of 0.6–1.7 [kHz] is isolated as the cue region. Frication noise at higher frequencies than the isolated cue regions was present in all tokens of /ʃɑ, ʒɑ, fɑ, vɑ/, thus, the necessary perceptual cue for fricative place of articulation is the frequency of the lowest bound (i.e., the frequency of the lower edge) of the band of frication noise.

3.3.2 Discrimination of Voicing

For stop consonants, it is evident that timing information, such as voice-onset time, defined as the duration between the release of burst and the onset of voicing, is critical for the discrimination of voiced consonants /b, d, g/ from their unvoiced counterparts /p, t, k/ (Liberman et al. 1958; Li and Allen 2011). In Section 3.2, we observed that unvoiced sibilants /ʃɑ, sɑ/ tend to have a longer frication region than their voiced counterparts /ʒɑ, zɑ/. Most studies note that the duration of unvoiced fricatives is generally longer than that of the voiced fricatives (Baum and Blumstein 1987; Stevens et al. 1992; Jongman et al. 2000), but the natural distributions of the two categories overlap considerably (Baum and Blumstein 1987).

The results of the truncation experiment (TR07) show that even when the unvoiced fricatives are truncated to the shorter average original duration of voiced fricatives, listeners can still correctly perceive the unvoiced fricatives. Only when the unvoiced /ʃɑ, sɑ/ are deeply time-truncated to a duration similar to the minimum possible duration of /ʒɑ, zɑ/ (2.5–5.5 [cs]), do listeners report some confusions with the corresponding voiced fricative. Similarly, when the frication for the unvoiced /fɑ/ is truncated to $\leq$3 [cs], but not completely removed, weak (<40%) confusions with the voiced fricative /vɑ/ are reported. Together these observations suggest that, although duration may play a role when the signal is sufficiently degraded, it is not the discriminating cue for fricative voicing.

An acoustic analysis reveals that voiced sibilants contain a salient F0 amplitude modulation (AM) introduced by the glottal vibration. It is observed that the frication portions of the voiced fricatives /ʒɑ, zɑ, vɑ/ are modulated by F0, while the frication portions of the unvoiced fricatives /ʃɑ, sɑ, fɑ/ are not. The relevant question is then whether or not the AM of the frication is the primary cue for voicing.

The perceptual data suggests that the AM of the frication at F0 is sufficient for correct perception of voicing. In the filtering experiment (HL07), the P_c for voiced fricative /zɑ/ tokens with isolated cue regions remains > 80% even when the original signal is high-pass filtered at 1.3 [kHz]. Similarly, the P_c of all /ʒɑ/ tokens does not drop below the full-band level until the signal is high-pass filtered above 0.9 [kHz]. These consistent observations confirm that the low-frequency voicing (including the voice bar) is not a necessary cue. Thus, the modulation present in the high-frequency frication is a necessary cue for reliable perception of voiced fricatives.

3.3.3 Role of Duration

For all tokens in this study, the ranges of minimum frication durations were identified, as summarized in Tables 3.1 and 3.2. When the frication is truncated beyond these minimum durations, but not completely removed, strong (>50%) confusions with plosives emerge. The plosive confusions for severely truncated /zɑ, ʒɑ/ tokens are /dɑ, gɑ/. For /sɑ, ʃa/ the confusions are /dɑ, gɑ, tɑ/. For both /fɑ, vɑ/, only significant /bɑ/ confusions are observed when the frication is truncated beyond the minimum duration. Thus, when the frication is truncated to <3 [cs], the spectral information that remains is perceived primarily as a voiced plosive. The durational cue, encoded in the sustained frication region, is a necessary cue for identification of fricatives.

3.3.4 Conflicting Cue Regions

The masking of speech by noise, and the resulting confusions, are of key importance to understanding speech communication. In a previous study (Li et al. 2010), it was discovered that naturally produced stop consonants often contain acoustic components that are not necessary for correct perception, but can cause listeners to confuse the target sound with competing sounds when the primary cue region for the target consonant is removed. A speech component that contains cues for a non-target consonant and are also not necessary for correct perception of the target consonant is defined as a *conflicting cue region*. For instance, in Fig. 4a of Li et al. (2010), the /kɑ/ token from talker f103, with a mid-frequency burst centered at 1.6 [kHz] isolated as the cue region, also contains a high-frequency burst energy above 4 [kHz] and a low-frequency burst energy below 1 [kHz] that contain perceptual cues for /tɑ/ and /pɑ/, respectively. Once the mid-frequency burst cue region is removed, the /kɑ/ is confused with /tɑ/ or /pɑ/ (Li and Allen 2011); selective amplification of the conflicting cue regions can lead to complete morphing of the token into a consistently perceived /tɑ/ or /pɑ/ (Kapoor and Allen 2012).

Fricative consonants can also contain conflicting cue regions, specifically, all /ʃɑ/ and most of the /ʒɑ/ tokens that we examined contain conflicting cue regions in the frication above 4 [kHz]. When the frication <4 [kHz] is removed by filtering from /ʃɑ, ʒɑ/, the listeners report the non-target consonants /sɑ, zɑ/, respectively. The non-sibilant fricative /fɑ/ from talker f101 (Fig. 3.3 a) also contains a high-frequency frication noise above 3 [kHz] that leads most listeners to report /zɑ/ in the absence of the mid-frequency /fɑ/ cue region. Similarly, the /fɑ/ token from talker m111 contains a high-frequency conflicting cue region for /zɑ/.

These conflicting cue regions can have a significant impact on speech perception, especially when the primary cue region is masked under noisy circumstances. Based on our perceptual data, we hypothesize that many of the most frequent confusion patterns (Miller and Nicely 1955; Phatak et al. 2008), e.g., /p/⇔/t/⇔/k/, /b/⇔/d/⇔/g/, /ʃ/⇔/s/, /ʒ/⇔/z/, /f/⇔/v/⇔/b/, and /m/⇔/n/, are explained by the existence of such conflicting cues. Thus, the most efficient way to reduce confusions in speech perception is to either increase the strength of the primary spectral cue region and/or remove the conflicting cue region(s).

3.3.5 3DDS Method for Isolating the Perceptual Cue Region

The 3DDS method has proven to be effective in locating the spectral regions that contain the necessary and conflicting cues for both plosives (Li et al. 2010) and fricatives in natural speech. No single cue was found to be sufficient for the perception of a fricative, instead the combination of all necessary cues, contained within a subset of the frication region, were together sufficient for perception. The 3DDS method allows one to analyze the perceptual effects of speech components without assumptions about the time-frequency location or type of perceptual cues.

The 3DDS method finds the single spectral region that contains a set of cues that are sufficient for perception. This region can contain multiple acoustic elements and variable cues. In this study, a sufficient set of cues was isolated in a subset of the frication regions of the tokens. Once the spectral region is isolated, further analysis is needed to determine the discriminating acoustic properties.

In spite of its success, the 3DDS method has limitations. We conclude that a requirement on 3DDS is that all test tokens be perceived correctly at the control condition of the three experiments. Tokens that do not meet this requirement have 3DDS results where no spectral region contains sufficient perceptual cues. Our assumptions about the distribution of error which led to the sampling of one-third high, medium, and low-scoring tokens for experiments TR07 and HL07 were wrong. A recent study has shown that the majority of CV tokens are high-scoring for normal-hearing listeners in low-noise conditions (Singh and Allen 2012). A final assumption of this study, which is supported by the results of Singh and Allen (2012), is that the normal hearing listeners are using the same cues for perception. It is possible that children (Nittrouer 2002) and people with hearing loss (Zeng and Turner

1990) may rely on different or additional cues for consonant perception.

Analysis of conflicting cue regions can also be applied to investigations of noise-robustness and natural confusions. Finally, although not the purpose of this study, a larger database of tokens per consonant would be necessary in order to properly investigate the variability of perceptual cues across talkers.

3.3.6 Conclusions

1. The 3DDS method (Li et al. 2010) is extended to fricative consonants, thus providing a novel technique for isolating the cue regions in natural tokens.

2. The analysis of the results goes beyond target-consonant cue observations by examining the possible effects of conflicting cue regions. Conflicting cue regions in natural speech frequently explain the consonant confusions that arise under noisy or limited bandwidth conditions.

3. It has been previously observed that voiced fricatives exhibit modulations. Our novel result is that even when the entire low-frequency spectral region is removed by filtering, the high-frequency modulations are a sufficient cue for voicing.

4. In our study, the fact that voiceless fricatives are longer on average than their voiced counterparts does not affect perception; when voiceless fricatives are truncated to the same original durations as their voiced counterparts, the error remains at zero. The presence of modulations is the necessary discriminating cue for voiced vs. voiceless.

CHAPTER 4

HEARING-IMPAIRED CONSONANT RECOGNITION

4.1 Methods

The methodologies of HI Experiment 2, as well as the techniques used in the analysis of the HI consonant recognition data, are detailed in this section.

4.1.1 Subjects

Nine HI subjects with sensorineural hearing loss were recruited for this study from the Urbana-Champaign, IL community. Both ears were tested for all listeners but one, resulting in data for 17 individual ears. All subjects reported American English as their first language and were paid to participate. IRB approval was obtained prior to the experiment. Typanometric measures showed no middle-ear pathologies (type A tympanogram). The ages of eight HI subjects ranged from 65 to 84; one HI subject (14R) was 25 years old. Based on the pure-tone thresholds, all ears had >20 [dB] of hearing loss (HL) for at least one frequency in the range 0.25–4 [kHz].

4.1.2 Audiometric Measurements

The majority of the ears in our study have slight-to-moderate hearing loss with high-frequency sloping configurations (see Table 4.1). One HI ear (14R), has an inverted high-frequency loss, with the most hearing loss <2 [kHz] and a threshold within the normal range at 8 [kHz]. The audiometric configuration of low-frequency flat loss with high-frequency sloping loss can be modeled as a piecewise linear function of the form

Table 4.1: The 17 HI ears are ordered by the average of the left and right ear h_0 values (Eq. 4.1). The model parameters estimate the flat low-frequency loss h_0 [dB], the frequency at which sloping loss begins f_0 [kHz], and the sloping high-frequency loss s_0 [dB/octave]. RMS error ϵ of the model fits is reported in [dB]. The three-tone (0.5, 1, 2 kHz) Pure Tone Average (PTA) [dB HL], age of the listener, and MCL for each ear are included. The mean and standard deviation (μ, σ) for all values are reported in the bottom row (ear 14R excluded). Note that the HI subject ordering would be the same if based on average PTA across the left and right ears.

HI ear	h_0	f_0	s_0	RMS ϵ	PTA	Age	MCL
44L	9	1	10	11	10	65	82
44R	13	1	7	7	15	65	78
46L	11	1.5	20	9	8	67	82
46R	18	3	27	7	17	67	82
40L	22	2	20	5	22	79	80
40R	18	1	11	5	23	79	80
36L	19	1	7	8	27	72	68
36R	25	1	10	4	28	72	70
30L	28	1.5	22	3	30	66	80
30R	25	1.5	27	5	27	66	80
32L	30	1	9	3	35	74	79
32R	27	1.5	14	3	27	74	77
34L	34	3	50	6	32	84	84
34R	26	1.5	26	4	28	84	82
01L	44	4	33	2	45	82	83
01R	47	3	41	4	47	82	82
14R	72	2	-37	3	73	25	89
(μ, σ)	(25, 11)	(2, 0.9)	(21, 13)	(5, 2)	(26, 11)	(74, 7)	(79, 4)

$$h = \begin{cases} h_0 & \text{if } f \leq f_0 \\ h_0 + s_0(log_2(f/f_0)) & \text{if } f > f_0 \end{cases} \tag{4.1}$$

where h is the hearing loss in [dB] and f is frequency in [kHz]. The parameter f_0 estimates the frequency at which the sloping loss begins; h_0 estimates the low-frequency ($f \leq f_0$) flat loss in [dB]; s_0 estimates the slope of the high-frequency loss in [dB/octave]. The parameters are fit to minimize the root-mean-square (RMS) error ϵ, in [dB]. The resulting parameter and RMS ϵ values for each model fit are reported in Table 4.1. The linear fits for the HI ears with the lowest (01L) and highest (44L) RMS error are shown, overlaid on the raw data, in Fig. 4.2, along with the overlaid linear fits for all 17 HI ears. Note that the ordinate is flipped, therefore positive slopes s_0 slant downward. Further details on the computation of the parameter values are discussed in Appendix A.

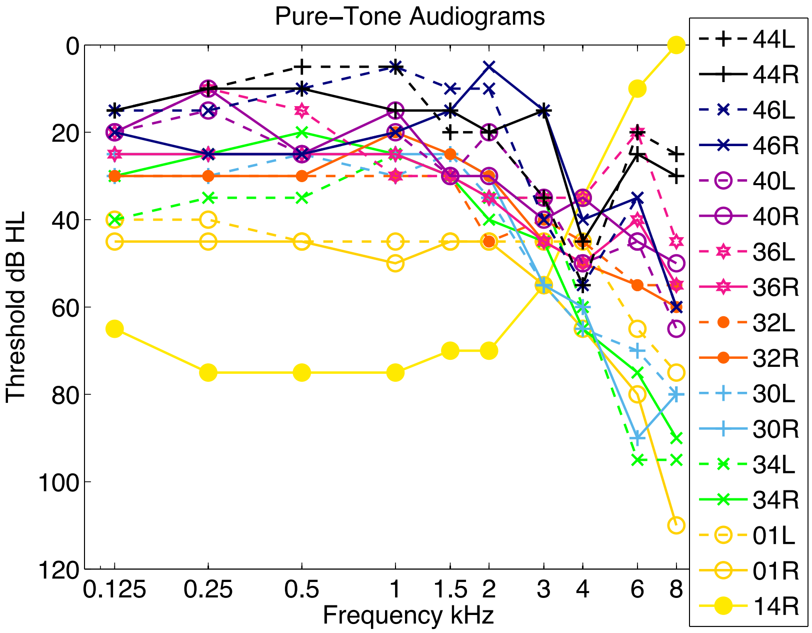
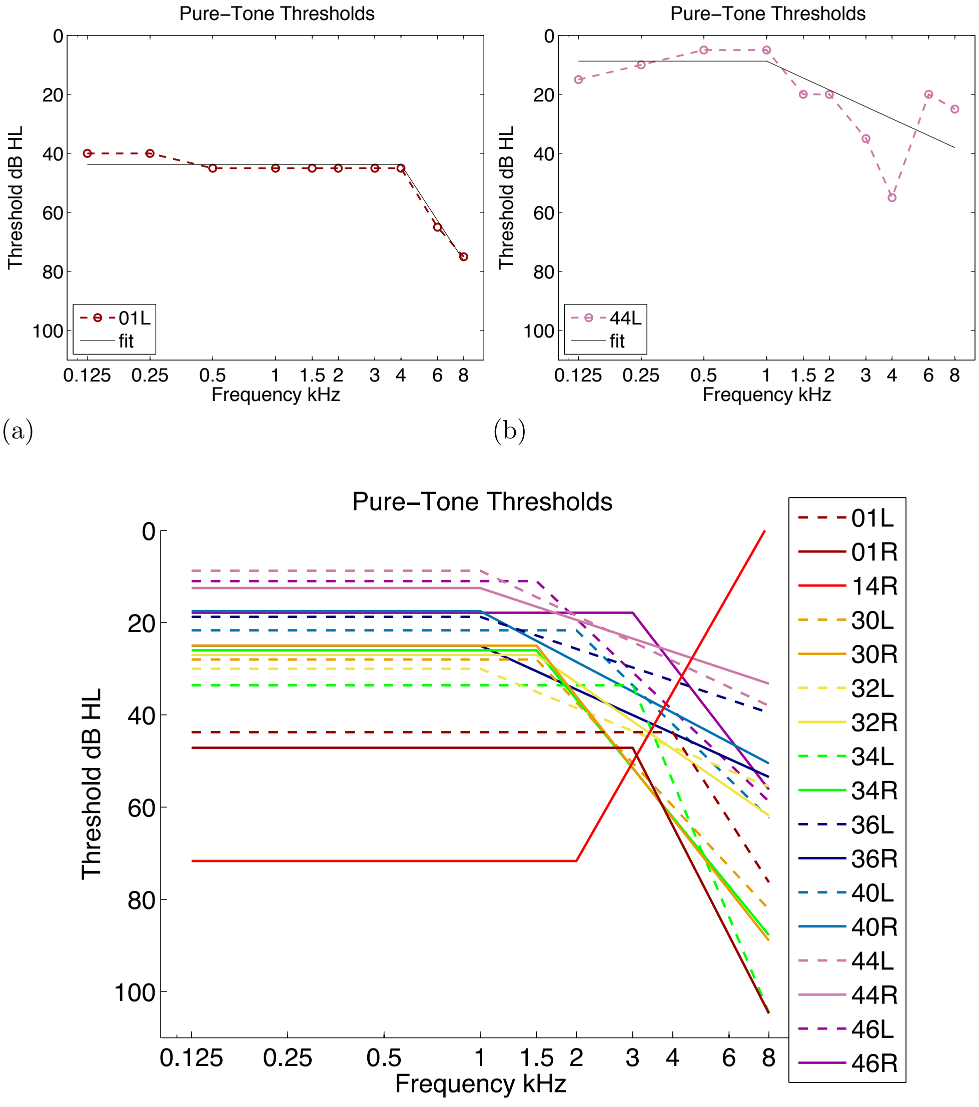


Figure 4.1: Pure-tone thresholds for the 17 HI ears included in this study. Right ears (R) are shown as solid lines, left ears (L) are shown as dashed lines.



(a)

(b)

(c)

Figure 4.2: Pure-tone thresholds and the three-parameter linear fits for (a) 01L and (b) 44L; these plots represent the lowest and highest RMS ϵ, respectively. (c) The linear fits for all 17 HI ears.

4.1.3 Speech Materials

All stimuli used in this study were selected from the Linguistic Data Consortium Database (LDC-2005S22) (Fousek et al. 2004). Speech was sampled at 16 [kHz]. Fourteen naturally spoken American English consonants (/p, t, k, f, s, ʃ, b, d, g, v, z, ʒ, m, n/+/ɑ/) were used as the test stimuli. Each consonant was spoken in an isolated (i.e., no carrier phrase) consonant-vowel (CV) context, with the vowel /ɑ/. Speech samples from six female talkers and five male talkers were used (see Table 4.2), with two talkers (one male, one female) for each consonant, resulting in a total of 28 test tokens (14 consonants x 2 talkers = 28 tokens). The term *token* is used throughout this dissertation to refer to a single CV speech sample from one talker. The 28 test tokens were chosen based on their performance in noise for NH listeners; all tokens used were “zero-error” at or above −2 [dB] SNR in SWN (Phatak and Allen 2007). The term “zero-error” is used to indicate that a maximum of one error was observed over a population of 24 NH listeners at quiet and -2 [dB] SNR, with 36 ± 4 combined trials per token (Singh and Allen 2012). One token of /fɑ/ (male token, m112), that was damaged in the pre-processing, has not been included in this analysis.

The stimuli were presented at the MCL for each individual HI ear. For the majority of the HI ears the MCL was approximately 80 ± 4 [dB] SPL (see Table 4.1). The only listeners that did not choose a MCL within this range were subjects 36L/R and 14R.

The LDC-2005S22 Database labels for the test tokens, along with the NH SNR_{90} values for SWN, are listed in Table 4.2. All SNR_{90} values are calculated by linear interpolation between measurements taken at -22, -20, -16, -10, and -2 [dB].

4.1.4 Experimental Procedure

The speech was presented at 4 SNRs (0, 6, 12 [dB] and quiet) using SWN, generated as described by Phatak and Allen (2007). Presentations were randomized over consonant, talker, and SNR. For each HI ear, the experiment was performed in two phases. The first phase presented each consonant eight times (four per token) at each of the four SNRs, resulting in 32 presentations per consonant (4 presentations x 2 tokens x 4 SNRs). The second phase

Table 4.2: For each consonant-vowel token (CV), the male (m) and female (f) talker labels are listed, along with the corresponding NH SNR_{90} values, in [dB], for SWN. These values are calculated from the noise-masking data of Phatak and Allen (2007). The /fɑ/ from talker m112 is marked with an * to indicate that this token was not included in the HI data analysis.

CV	M Talker	SNR_{90}	F Talker	SNR_{90}
bɑ	m112	-2	f101	-10
dɑ	m118	-7	f105	-13
fɑ	m112*	-5*	f109	-12
gɑ	m111	-12	f109	-3
kɑ	m111	-13	f103	-11
mɑ	m118	-14	f103	-11
nɑ	m118	-4	f101	-7
pɑ	m118	-14	f103	-17
sɑ	m120	-10	f103	-13
ʃɑ	m118	-16	f103	-15
tɑ	m112	-17	f108	-14
vɑ	m118	-3	f101	-10
ʒɑ	m107	-7	f105	-17
zɑ	m118	-17	f106	-18

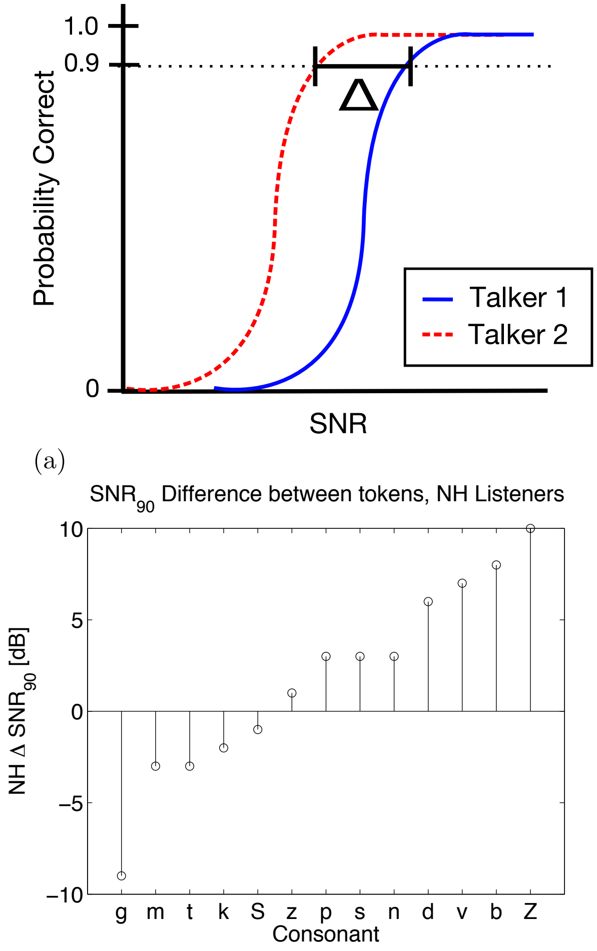
used an adaptive scheme to increase the number of presentations, and thus the statistical power of the test. This adaptive scheme determined the number of additional presentations for each token; the number of phase two presentations ranged from 1–6 at each SNR, with increased presentations assigned to conditions that produced the most error. Thus, the total number presentations of each consonant (over the two phases and 4 SNRs) ranged from $N = 40$–80 for each HI ear. The Vysochanskiï–Petunin inequality (Vysochanskij and Petunin 1980) was used to verify that the number of trials were sufficient to determine correct perception within a 95% confidence interval (Singh and Allen 2012).

The experiment was implemented as a MATLAB graphical user interface. All of the data-collection sessions were conducted with the subject seated in a single-walled, sound-proof booth, with the door of the outer lab closed. The speech was presented monoaurally via an Etymotic ER–3 insert earphone. The contralateral ear was not masked or occluded. The subjects set their MCL for speech before testing began. Subjects were allowed to adjust the sound level at any time during the experiment; such changes were automatically recorded in the log file. Based on the log files, no subject chose to adjust the sound level during the course of the experiment. A practice session, with different tokens from those in the test data set, was run first in order to familiarize the subject with the testing paradigm. Feedback was presented during the practice session, after each response by the subject. The remaining sessions presented the randomized test speech tokens. After hearing a single presentation of a token, the patient would choose from the 14 possible consonant responses by clicking one of 14 CV-labeled buttons on the graphical user interface, with the additional option of up to two token repetitions, to improve accuracy; after three repetitions of the same token, the subject had to select a response to continue. Short breaks were encouraged to reduce the effects of test fatigue. Additional experimental details are provided in Han (2011).

4.1.5 Characterizing Tokens with NH Psychoacoustic Data

Psychoacoustic data from classical masking, filtering and time truncation experiments can be used to characterize the consonant cues of each token in terms of intensity, frequency, and temporal properties. NH listener psychoacoustic data for the 28 test tokens (14 consonants) used in the present experiment, were collected by Phatak and Allen (2007); Li (2010). High/low-pass filtering and time-truncation data allows one to identify, in each naturally variable token, the spectral time-frequency region that contains the acoustic components that are necessary for correct perception, we refer to this as the *necessary cue region* (Li et al. 2010, 2012).

A key metric of each token's robustness to noise is the SNR_{90}, defined as the full-bandwidth SNR at which the probability of NH correct recognition for that individual token drops below saturation to 90%. The lower the SNR_{90}, the more robust a token is to noise. For NH listeners, this psychoacoustic measure has been found to be significantly correlated to the physical intensity of the necessary consonant cue region, with tokens that have more intense cue regions having lower SNR_{90} values, in both WN and SWN (Régnier and Allen 2008; Li et al. 2010, 2012). As discussed in Section 4.1.3, the NH SNR_{90} values for the selected test tokens are below the worst noise condition that was used to test HI recognition in the present experiment, 0



(a)

(b)

Figure 4.3: (a) Illustration of probability vs. SNR curves for two tokens, with the difference in SNR$_{90}$ values (ΔSNR$_{90}$) indicated. The SNR$_{90}$ is defined as the SNR at which the probability of recognition drops below 90%, while ΔSNR$_{90}$ quantifies the difference in noise-robustness across two tokens. (b) The NH ΔSNR$_{90}$ for all consonants in this experiment, as computed from NH perceptual data (Table 4.2) in the presence of SWN (Phatak and Allen 2007). A positive NH ΔSNR$_{90}$ indicates that the female token (talker 2) is more robust to noise, while a negative value indicates that the male token (talker 1) is more robust to noise. The capital "S" and "Z" labels refer to /ʃ/ and /ʒ/, respectively.

[dB] SNR (see Table 4.2). Due to natural variability of cue region intensity, the SNR_{90} values for a large number of tokens are approximately Gaussian distributed (Singh and Allen 2012).

It follows from these findings that, for two tokens of the same consonant, the difference between the NH SNR_{90} values is proportional to the difference in intensity of the necessary acoustic cue regions. Since tokens of the same consonant have perceptual cues within a similar frequency range, the NH $\Delta\mathrm{SNR}_{90}$ can be used to relate the true audibility of their necessary cue regions.

For each consonant, the SNR_{90} of the token from the male talker was subtracted from that of the female talker; this measure is illustrated in Fig. 4.3 (a) with Δ marking the difference between the two SNR_{90} values. These differences are reported for each pair of consonant tokens in Fig. 4.3 (b), with the consonants sorted along the abscissa by monotonically increasing NH $\Delta\mathrm{SNR}_{90}$ values. This plot shows that for /g/, the male token is more robust to noise by 9 [dB], while for /ʒ/, the female token is more robust to noise by 10 [dB]. Of the selected tokens, there are small differences in the noise robustness ($\leq \pm 3$ [dB]) of eight consonants, /m, t, k, ʃ, z, n, p, s/. The NH $\Delta\mathrm{SNR}_{90}$ values are controlled by the selection of the experimental tokens.

Although the NH SNR_{90} was controlled in the design of the experiment, the effect of NH $\Delta\mathrm{SNR}_{90}$ on HI perception was unknown and this measure was allowed to vary from -9 to $+10$ [dB]. The NH SNR_{90} values reported throughout Chapter 4 are calculated from the SWN masking data of Phatak and Allen (2007), in order to provide a noise-matched comparison to the HI responses in SWN.

4.1.6 Hearing-Impaired Average Error

For each ear, the traditional metric of average consonant error at a particular SNR, $\overline{P_e}(s)$, is computed as

$$\overline{P_e}(s) = \frac{1}{28} \sum_{\mathrm{C}=1}^{14} P_e^M(C,s) + P_e^F(C,s) \tag{4.2}$$

where $P_e(C, s)$ is the probability of error for consonant C at SNR s. M and F indicate the tokens from talkers 1 (male) and 2 (female), respectively.

4.1.7 Hearing-Impaired Relative Noise-Robustness

The average error difference, $\overline{\Delta P_e}$, for a given consonant is formulated as

$$\overline{\Delta P_e} = \frac{1}{n(S)} \sum_{s \in S} P_e^M(s) - P_e^F(s) \tag{4.3}$$

$$S = \{s \in \{0, 6, 12, Quiet\} : s \leq s^*\}$$

where s^* is the highest SNR at which more than one error is observed for either of the two tokens and $n(S)$ indicates the number of elements in set S. $\overline{\Delta P_e}$ for each consonant is only computed over the SNRs at which error is observed for at least one of the two tokens in order to better capture the largest observed differences in error. If no error is observed over all SNRs, $\overline{\Delta P_e} \doteq 0$.

4.1.8 Hearing-Impaired Consonant Confusion Analysis (K-Means)

We refer to perceptual differences across multiple tokens of the same consonant as *within-consonant differences*. The variability of naturally spoken acoustic cues can lead to HI within-consonant differences in both error and consonant confusions (Trevino and Allen 2013a,b); therefore, calculations at the token level are necessary in any analysis that attempts to understand how a HI listener is decoding the acoustic cues that are available to them.

The Hellinger distance is a metric for computing the distance between two probability distributions. The probability distributions that we compare in this dissertation are the ones defined by each row of a confusion matrix. In the case of this experiment, there are 14 possible consonant responses. Such a vector of probabilities can be considered as a point in 14-dimensional space, where each dimension corresponds to the probability of each consonant response. Distances between confusion results are computed within this 14-dimensional space; such distances provide a measure of consonant confusion similarity, which can be used to compare HI ears, SNRs, or tokens.

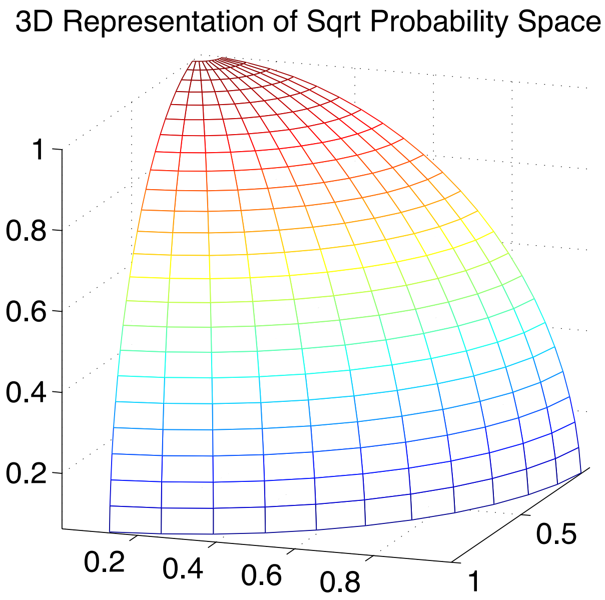


Figure 4.4: Illustration of the spherical space over which the HI response data would be distributed in the case of three possible consonant responses.

We will show that the squared Hellinger distance is equivalent to 1 minus the direction cosine, when computed from the square root of probabilities. This relationship allows us to use widely known algorithms that employ 1 minus the direction cosine, such as spherical k-means clustering, to analyze the data. Let $P_{r|s}(snr, HI)$ be the probability of the consonant response r for a fixed stimulus s, SNR, and HI ear; the probabilities for all possible responses for a fixed stimulus would be a row of the confusion matrix. A data point in the 14-dimensional space, $\mathbf{x}$, is then defined as $x_i = \sqrt{P_{r_i|s}(snr, HI)}$, $i = 1, 2, 3, \ldots 14$. Since the vector is composed of probabilities that sum to 1, the points lie on the unit sphere, $||\mathbf{x}|| = 1$. An illustration of the spherical space over which the data would be distributed, in the case of three possible consonant responses, is shown in Fig. 4.4.

Let $\mathbf{x}, \mathbf{y}$ be two data points in the 14-dimensional space. We define the Euclidean norm as

$$||\mathbf{x}|| = \sqrt{<\mathbf{x}, \mathbf{x}>} = \sqrt{\sum_i x_i^2}$$

and the inner product between vectors $\mathbf{x}$ and $\mathbf{y}$ as

$$< \mathbf{x}, \mathbf{y} >= \sum_i x_i y_i = ||\mathbf{x}||||\mathbf{y}||cos(\Theta_{xy})$$

Then the square of the Hellinger distance, $\frac{1}{\sqrt{2}}||x - y||$, is

$$\begin{aligned} H^2(\mathbf{x}, \mathbf{y}) &= \frac{1}{2}||\mathbf{x} - \mathbf{y}||^2 = \frac{1}{2}(||\mathbf{x}||^2 - 2 < \mathbf{x}, \mathbf{y} > + ||\mathbf{y}||^2) \\ &= 1 - < \mathbf{x}, \mathbf{y} >= 1 - ||\mathbf{x}||||\mathbf{y}||cos(\Theta_{xy}) = 1 - cos(\Theta_{xy}) \end{aligned}$$

Thus, the spherical k-means algorithm, which forms clusters based the measure $1 - cos(\Theta_{xy})$ between points distributed on the unit sphere, also produces clusters that minimize the Hellinger distance. The spherical k-means clustering algorithm is implemented in MATLAB, with the *kmeans*() function. For each token in our implementation, one of the clusters is always composed of the data points where HI listeners correctly perceived the consonant; the remaining clusters are composed of the data with varying degrees of error. Therefore, assuming there are errors, the minimum possible number of clusters, K, for each token is 2.

Additionally, the angle between the HI listener response x and the plane representing the “primary” confusion groups can be calculated. With this implementation, HI listener data that contains varying degrees of the same primary confusions would show zero distance between the points; nonzero distances would indicate the degree of deviation from the primary confusion group.

The k-means algorithm groups HI listener data that is similar in terms of consonant confusions. The size and number of clusters are a function of the diversity of hearing impairment across listeners in the study (i.e., there is no fixed prior that represents all possible groups of subjects), therefore, a k-means implementation which does not assign a prior probability to each cluster models the experimental setup more realistically than a Gaussian Mixture Model (GMM). The G-means algorithm (Hamerly and Elkan 2004) was added to the implementation in order to automatically select the number of means, K, based on an Anderson-Darling test of statistical significance.

4.2 Results

4.2.1 Pure-Tone Thresholds

The pure-tone thresholds for the 17 ears in the study are shown in Fig. 4.1 and, parametrically, in Fig. 4.2. Two audiometric configurations are observed in our subject population. The majority of HI subjects in this study (16 out of 17 ears) have approximately flat loss below 1-3 [kHz] with high-frequency sloping loss; these HI subjects fall within the slight-to-moderate hearing loss range. One ear (14R) has severe hearing loss <2 [kHz], with an inverted high-frequency loss and a pure-tone threshold within the NH range at 8 [kHz].

The HI ears, along with general subject data, are listed in Table 4.1, and are roughly ordered based on hearing loss. Using the piecewise linear function h as a model, the variables h_0, f_0, and s_0 are fit for each ear. All resulting parameter fits are also listed in Table 4.1. For the 16 ears with a high-frequency sloping configuration (i.e., positive s_0), the level of low-frequency flat loss h_0 ranged from 9–47 [dB]. As may be seen from Fig. 4.1 and Table 4.1, the hearing loss across right and left ears is fairly similar (i.e., symmetrical); the maximum measured difference between hearing loss in the right and left ears at any low frequency (0.125–2 [kHz]) is 15 [dB] and at any of the high frequencies (3–8 [kHz]) the maximum measured difference is 35 [dB].

4.2.2 Average Consonant Error

As described in Section 4.1, all tokens in the experiment were selected based on NH-listener results of <3% error at SNRs $\geq$ -2 [dB]. Thus, for the HI ears, a result of zero error at the 0, 6, 12 [dB] SNR and quiet conditions is equivalent to "normal-hearing" performance.

The average consonant error, $\overline{P_e}(s)$, shown in Fig. 4.5, varies widely across HI ears, with four ears within the range of normal performance at low-noise levels (44L/R, 36L, 34L) and three ears reaching 50% error at 0 [dB] SNR (34R, 01L/R). Note that, on a log scale, the $\overline{P_e}(s)$ for a HI ear is approximately linear with respect to s, similar to the error predicted by the articulation index formula, as established by H. Fletcher in 1921 (Allen 1994).

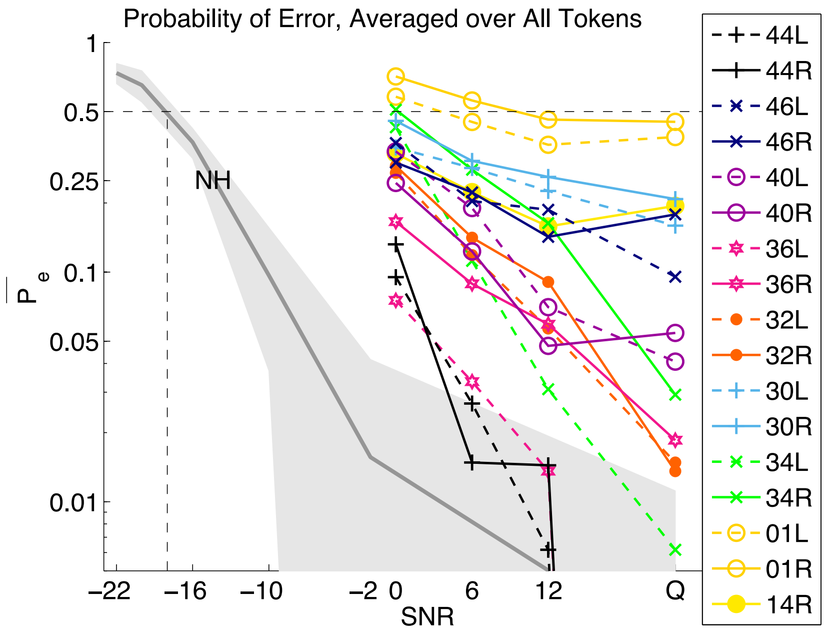


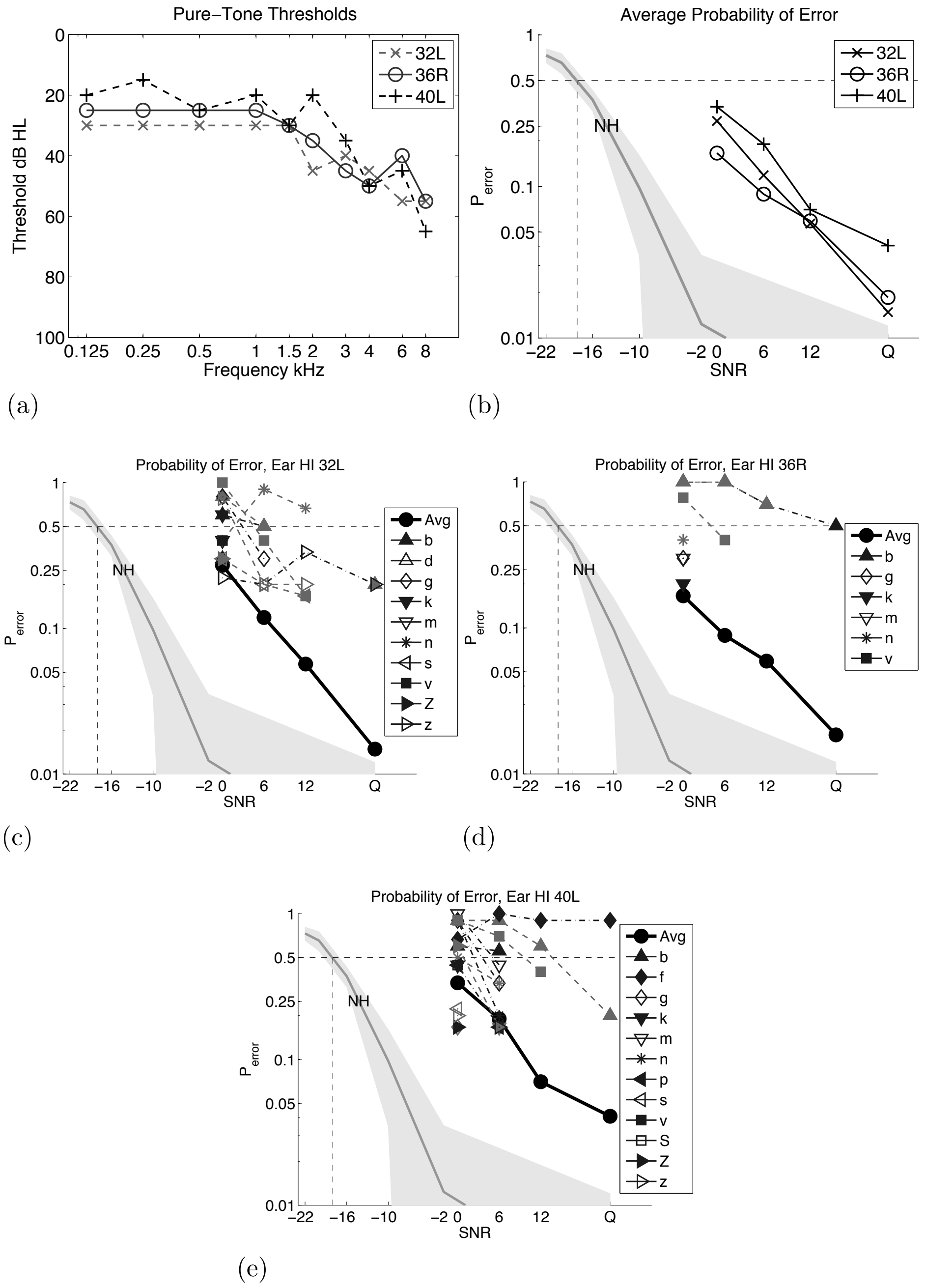
Figure 4.5: Average error over all consonant stimuli for each HI ear as a function of SNR (Eq. 4.2). Right ears (R) are shown as solid lines, left ears (L) as dashed lines. The average NH error (gray solid line) is included for reference, along with a gray error region representing one standard deviation.

4.2.3 Individual Token Error vs. Average Error

The average speech score, across many different stimuli, or the SNR at which the average speech score hits 50% are often used to compare HI subjects. Can HI listeners with similar pure-tone thresholds and similar average speech scores in both quiet and noise have different individual consonant tokens that are perceived in error?

Figure 4.6 (a, b) show data for three HI listeners (one ear per listener) who have similar pure-tone thresholds and average error ($\overline{P_e}(s)$ within 10% in 6, 12 [dB] SNR speech-shaped noise and in quiet). In addition, the average consonant error for all three HI ears is within 10% of NH performance at 12 [dB] SNR and quiet.

Despite the similar pure-tone thresholds and average errors, the individual consonants that are perceived in error are ear-dependent. The error for individual consonant tokens is compared to the average error for three example ears in Fig. 4.6 (c, d, e). Ear 32L shows errors with a large range of consonant tokens, particularly of /b, g, n, v, z/, reaching almost 100%



(a)

(b)

(c)

(d)

(e)

Figure 4.6: (a) The pure-tone thresholds from "similar" 3 HI listeners (data from 1 ear shown). (b) Average consonant error over 27 consonant tokens (14 consonants x 2 talkers, on token /fɑ/ omitted), for individual HI and NH listeners, on a log scale (shaded region for NH listener data denotes one standard deviation). (c-e) Individual and average token errors for HI ears 32L, 36R, and 40L. For the individual consonant tokens, dash-dot black lines indicate a token from a female talker, dashed gray lines indicate a token from a male talker. Random errors ≤1/N are not shown to reduce clutter. The labels "S" and "Z" correspond to /ʃ/ and /ʒ/, respectively. Similar pure-tone thresholds and average errors can have a large underlying individual variability in terms of the individual consonant token errors.

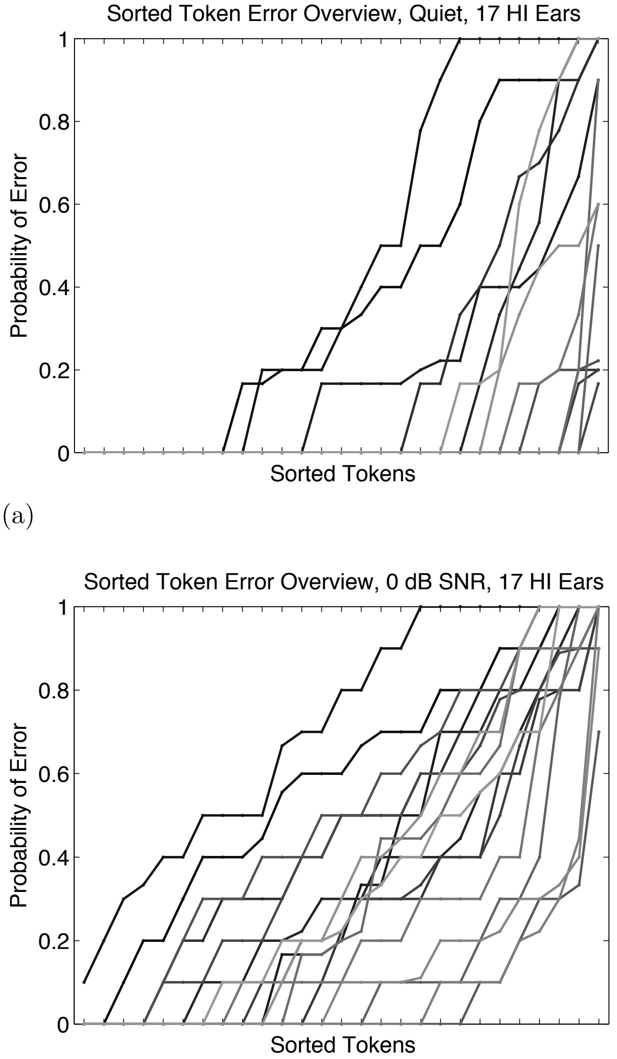
error with the male /n, v, z/ and the female /g/ tokens. Ear 36R shows high error with only the two /b/ tokens and the male /v/. Ear 40L also has difficulty with a wider array of consonants in noise, showing consistent errors with tokens of /b, f, g, n, m, v/. We see that similar pure-tone thresholds and average errors can have large underlying individual variability in terms of individual token errors.

A listener with a low average error can also have underlying large errors for only a small, specific set of stimuli. In terms of the average, ear 36R (Fig. 4.6 (d)) appears to be almost normal, falling within 10% of NH performance at 6, 12 [dB] SNR and quiet. However, when the individual consonant errors are examined, ear 36R actually has very high error for both tokens of /b/ and the female token of /v/; in fact, 36R has 100% error for both /b/ tokens at 0 and 6 [dB] SNR. Thus, an ear that appears "almost normal" in terms of the average speech score can actually have a severe problem with a very specific type of speech sound.

Sorted Error Plots

An overview plot of the distribution of errors across the set of individual consonant tokens allows one to visualize the underlying token errors that are often reported as a single averaged value. Overview plots of the errors across 27 test stimuli (2 talkers x 14 consonants, one /fɑ/ token omitted), are shown in Fig. 4.7, with each line representing the data for a single HI ear. For each individual HI ear, the tokens are sorted along the abscissa to create a monotonically increasing error plot. The sorting order of the tokens is ear-dependent. This monotonically increasing plot clearly shows the nature of the errors; that is, whether errors are widespread over all stimuli or if, instead, high errors are only observed for a specific subset of stimuli.

The distributions of error in quiet (Fig. 4.7 (a)) and in 0 [dB] SWN (Fig. 4.7 (b)) show that high degrees of error can be concentrated to small subsets of the total consonant stimuli. In the quiet condition, all tested HI ears with slight-to-mild hearing loss make no errors for the majority of consonant tokens. Despite the large number of tokens with no errors, the same ears can reach >50% error with the remaining tokens. In noise, as one would expect, higher degrees of error are observed across the test tokens; again, HI ears show a fraction of tokens with high error and others with zero



(a)

(b)

Figure 4.7: Sorted error overviews of the 27 test tokens (14 consonants x 2 talkers, one /fɑ/ token omitted), for (a) 17 HI ears in quiet, and (b) 17 HI ears in 0 [dB] SNR speech-shaped noise. Each line represents the data for an individual HI ear at a fixed SNR. For each individual ear, the tokens are sorted along the abscissa, creating a monotonically increasing error plot. Both with and without noise, errors for each HI ear are concentrated to a small subset of the stimuli. In quiet, all ears with slight-to-mild hearing loss show zero error for the majority of test tokens.

error.

4.2.4 Within-Consonant Differences - Noise Robustness

Next, we examine the variability in the noise robustness of tokens of the same consonant (i.e., within-consonant differences), in detail. The most extreme example of this variability is where one token of a consonant has no error at any tested SNR for a HI ear while the other token of the consonant reaches errors as high as 100%.

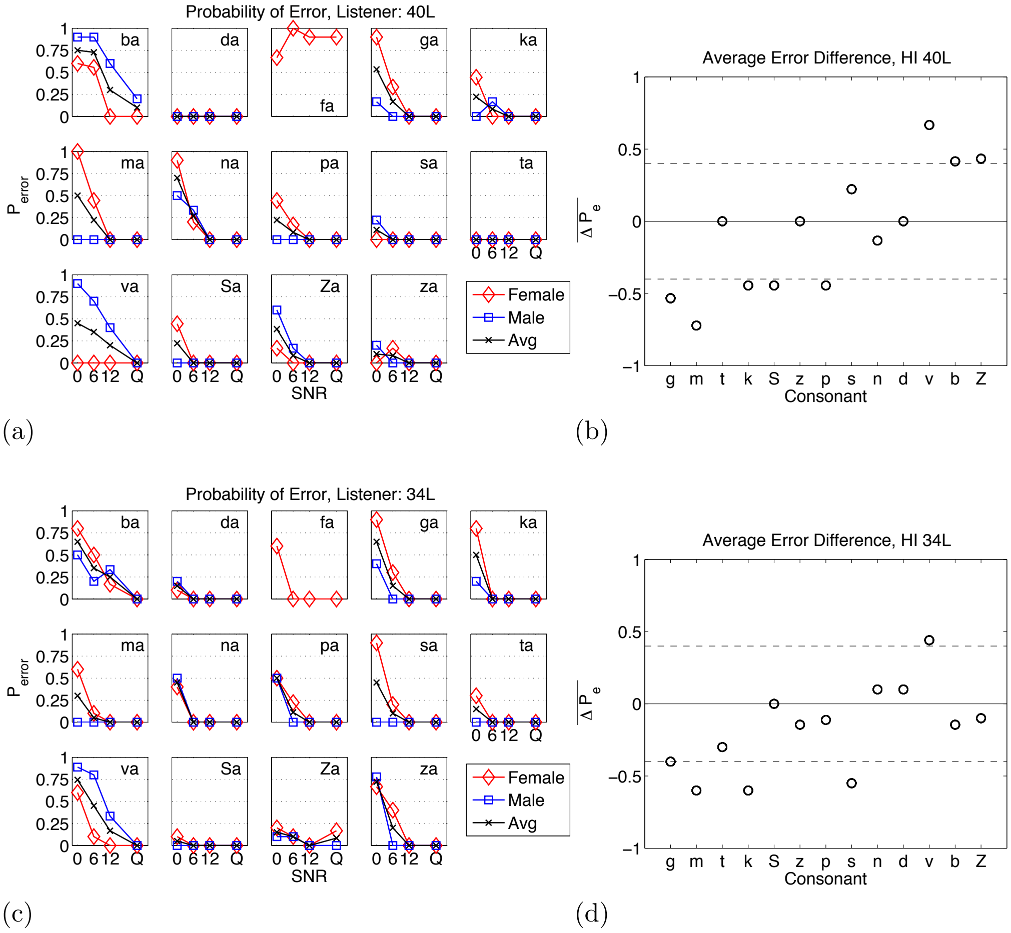


Figure 4.8: Results for HI ears 40L and 34L. (a, c) Consonant recognition error as a function of SNR, each subplot shows the data for one consonant; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b, d) $\overline{\Delta P_e}$ for each consonant (Eq. 4.3); consonants are ordered along the abscissa based on the NH $\Delta \mathrm{SNR}_{90}$ values in SWN (as in Fig. 4.3). $\overline{\Delta P_e} = \pm 0.4$ is marked for reference.

The consonant recognition error as a function of SNR for both talker tokens ($P_e^M(s)$ and $P_e^F(s)$) along with the average across the two talkers is displayed in 14 sub-plots (one for each consonant), for ears 40L and 34L (Fig. 4.8 (a, c), respectively). Ear 40L reaches $\geq$50% two-talker average error for /b, g, m, n, v/, as noise is introduced; when the error is analyzed at the token level, one finds that the error for /g, m/ is completely due to the female token and that the error for /v/ is completely due to the male token. Ear 34L reaches $\geq$50% two-talker average error for /b, g, k, p, v, z/, as noise is introduced. The largest differences in noise robustness for ear 34L are for tokens of /k, m, s, v/. For this ear, almost all of the average error for /k, m, s/ can be attributed to errors with only the female token. For /v/, the male token is recognized with no error in only the quiet condition, while the female token is robust down to 6 [dB] SNR. Thus, for both ears 40L and 34L, one can observe large differences in the noise robustness of tokens of the same consonant. Although the acoustical differences across these tokens are small enough for them to be recognized as the same consonant by NH listeners, they are significant enough to make a difference in HI perception.

Average Error Difference

To quantify this observation, the difference in error between the two tokens of each consonant is calculated as a function of SNR. These values are used to compute the average error difference, $\overline{\Delta P_e}$ (Eq. 4.3), shown for ears 40L and 34L in Fig. 4.8 (b, d). A negative $\overline{\Delta P_e}$ indicates that the male token is more robust to noise, while a positive value indicates that the female token of a consonant is more robust to noise. $\overline{\Delta P_e} = 0.4$ is marked for reference; the minimum number of experimental presentations for a token at a given SNR is N = 5, a 0.4 error difference corresponds to two trials in error, which is significantly different ($\alpha = 0.05$) from NH performance (Singh and Allen 2012). The consonants with the largest average error differences for ear 40L are /g, m, v/ and /m, k, s/ for ear 34L. The consonants are ordered along the abscissa by the NH ΔSNR_{90} values, as shown in Fig. 4.3 (b). This is done to determine if the token of a consonant that is more robust to noise for a NH listener would also be more robust for a HI listener. Overall, there is some agreement, as a rough increasing trend can be observed in Fig. 4.8 (b, d).

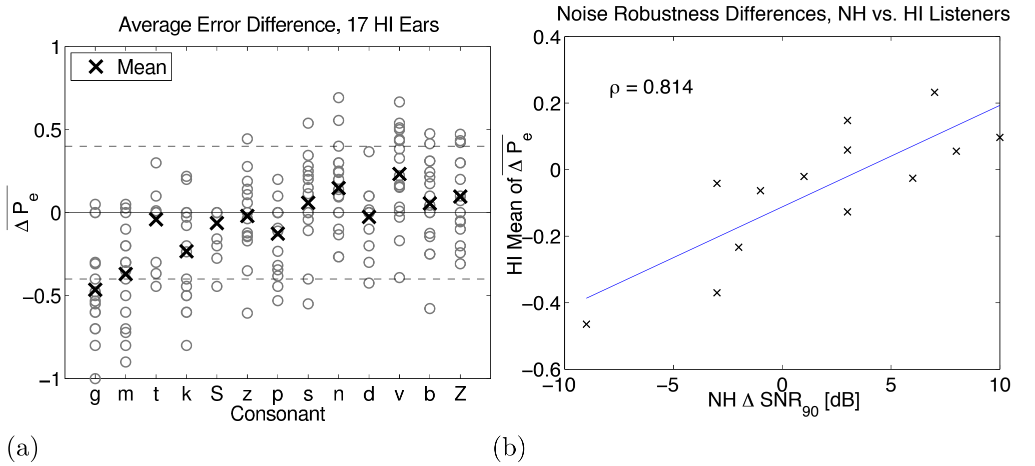


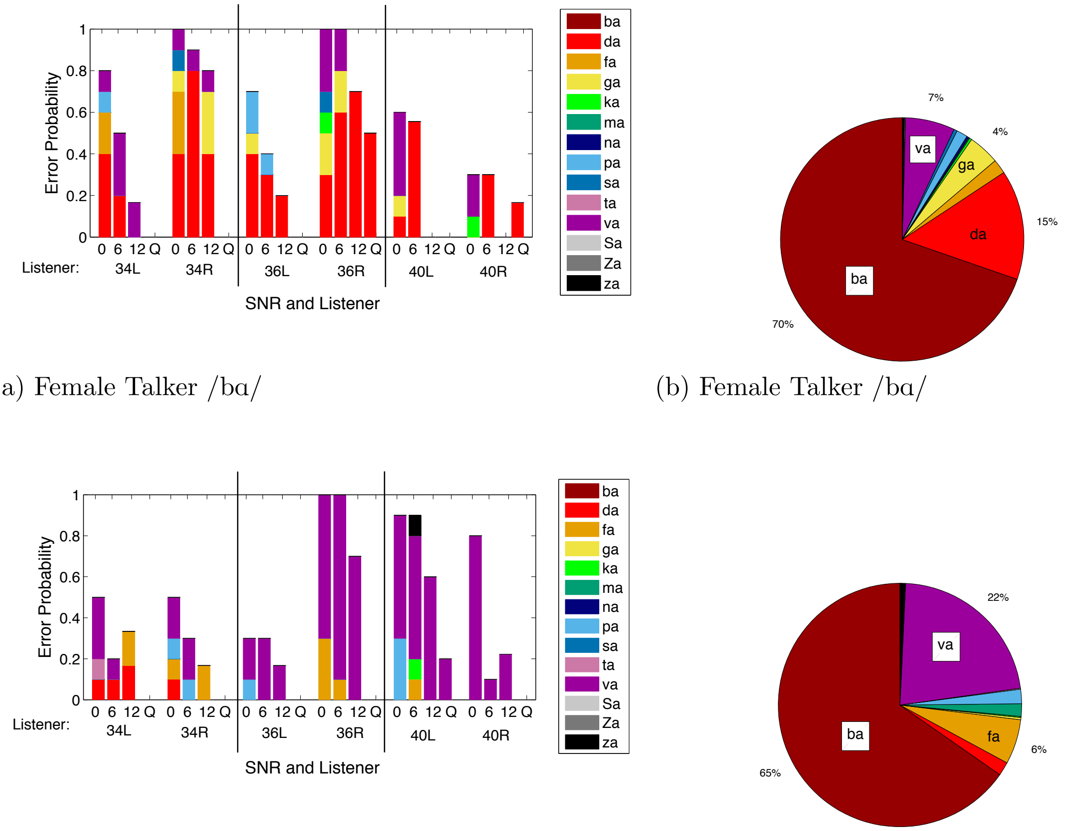
Figure 4.9: (a) $\overline{\Delta P_e}$ for all HI ears; consonants are coded by color and shape. Each point represents the value for a single HI ear, the mean across ears for each consonant is marked with an "x". A negative $\overline{\Delta P_e}$ indicates that the male token has lower error, a positive value indicates that the female token has lower error. Consonants are ordered along the abscissa based on the NH ΔSNR$_{90}$ values (as in Fig. 4.3). $\overline{\Delta P_e} = 0.4$ is marked for reference. (b) Comparison and linear regression of the mean $\overline{\Delta P_e}$ values and the NH ΔSNR$_{90}$ values (see Fig. 4.3), the two values are significantly correlated ($\rho = 0.81$, p-val < 0.001).

The $\overline{\Delta P_e}$ values for all 17 HI ears are shown in Fig. 4.9 (a). Large error differences between tokens is a widespread effect, with 16 out of 17 ears showing at least one average error difference >0.4. The consonants /g, k, m, n, p, v/ have $\overline{\Delta P_e}$ distributions that are significantly different from zero, indicating that one of the two tokens is consistently more robust to noise for HI ears.

Comparison of NH and HI Noise-Robustness Measurements

The SNR_{90} has been found to significantly correlate with the intensity (and thus audibility) of the time-frequency region that contains the NH perceptual cues (Régnier and Allen 2008; Li et al. 2010, 2012). Thus, the NH ΔSNR_{90} is related to the difference in intensity of the NH consonant cue regions. If token robustness for HI listeners is completely dependent on audibility/intensity of the acoustic element(s) containing the consonant cues, then the tokens of a consonant that have the more intense consonant cue regions (i.e., lower NH SNR_{90}s) should also be more robust to noise for HI listeners. For the sake of comparison, the consonants in Fig. 4.9 (a) are plotted in the same order as in Fig. 4.3 (b). A clear increasing trend can be observed in the mean HI $\overline{\Delta P_e}$ values, similar to the trend of the NH ΔSNR_{90} values. A linear regression between the two measures is plotted in Fig. 4.9 (b); the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values are significantly correlated ($\rho = 0.81$, p-value <0.001). If the HI $\overline{\Delta P_e}$ values are computed as the average over all tested SNRs (i.e., $n(S) = 4$ fixed for all consonants, Eq. 4.3), then the correlation coefficient is lower but remains significant ($\rho = 0.77$).

Despite this strong relationship, a notable amount of individual variability can be observed in the data of Fig. 4.9 (a). Tokens that are almost identically noise-robust for a NH listener can show large $\overline{\Delta P_e}$ values for a HI ear. As an example, the two tokens of /z, p, s/ have NH $\Delta\text{SNR}_{90} \leq 3$ [dB], indicating that the two tokens have necessary cue regions that are nearly equal in intensity. Yet, there are individual HI ears for which a $\overline{\Delta P_e} > 50\%$ is observed for /z, p, s/. In such cases, additional signal properties, perhaps the presence of *conflicting cues* (Li et al. 2010, 2012) or variations of the primary cues that the HI ears could be sensitive to, may play a role. To better understand the HI within-consonant differences, we next examine the consonant confusions.



(a) Female Talker /bɑ/

(b) Female Talker /bɑ/

(c) Male Talker /bɑ/

(d) Male Talker /bɑ/

Figure 4.10: Probability of error and confusions for the two tokens of /bɑ/. (a) Error for the female /bɑ/ token, data from six HI ears (34L/R, 36L/R, 40L/R). Confusions as a proportion of the total error are labeled by color. (b) The proportion of responses for the female token, averaged across all HI ears and SNRs; primary confusions are with /d, v, g/. (c) Error for the male /bɑ/ token, data from the same six HI ears (34L/R, 36L/R, 40L/R). Confusions as a proportion of the total error are labeled by color. (d) The proportion of responses for the male token, averaged across all HI ears and SNRs; primary confusions are with /f, v/.

Table 4.3: (a) Confusion matrix for the female /bɑ/ token, data from six HI ears (34L/R, 36L/R, 40L/R), at each SNR [dB]. (b) Confusion matrix for the male /bɑ/ token, data from the same six HI ears (34L/R, 36L/R, 40L/R), at each SNR [dB]. For both confusion matrices, the highest probability confusion in each row is highlighted in **bold**, and probabilities of 0% are removed to reduce clutter. (c) The recognition data for the female token, averaged across all 17 HI ears; primary confusions are with /d, v, g/. (d) The recognition data for the male token, averaged across all 17 HI ears; primary confusions are with /f, v/. The labels sh = ʃ, zh = ʒ, and a = ɑ.

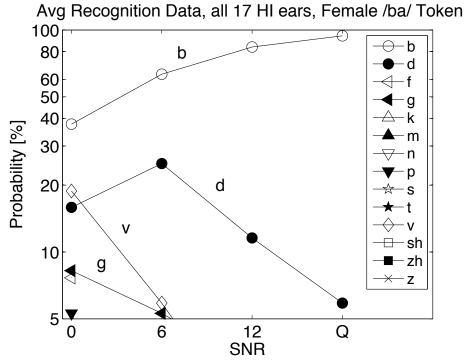
(a) Female bɑ token

Ear	SNR	b	d	f	g	p	v	s, k
34L	Q	100						
	12	83					**17**	
	6	50	20				**30**	
	0	20	**40**	20		10	10	
34R	Q	100						
	12	20	**40**		30		10	
	6	10	**80**				10	
	0		**40**	30	10		10	10
36L	Q	100						
	12	80	**20**					
	6	60	**30**			10		
	0	30	**40**		10	20		
36R	Q	50	**50**					
	12	30	**70**					
	6		**60**		20		20	
	0		**30**		20		**30**	20
40L	Q	100						
	12	100						
	6	44	**56**					
	0	40	10		10		**40**	
40R	Q	83	**17**					
	12	100						
	6	70	**30**					
	0	70					**20**	10

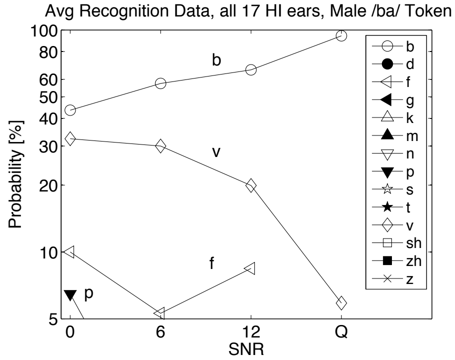
(b) Male bɑ token

Ear	SNR	b	d	f	g	p	v	k, t, z
34L	Q	100						
	12	66	**17**	**17**				
	6	80	**10**				**10**	
	0	50	10				**30**	10
34R	Q	100						
	12	83		**17**				
	6	70				10	**20**	
	0	50	10	10		10	**20**	
36L	Q	100						
	12	83			**17**			
	6	70			**30**			
	0	70			**20**		10	
36R	Q	100						
	12	30					**70**	
	6			10			**90**	
	0			30			**70**	
40L	Q	80					**20**	
	12	40					**60**	
	6	10		10			**60**	20
	0	10				30	**60**	
40R	Q	100						
	12	78					**22**	
	6	90					**10**	
	0	20					**80**	

(c)



(d)



4.2.5 Within-Consonant Differences - Confusions

The common NH listener confusion groups for English consonants were established by Miller and Nicely (1955) (e.g., /b, d, g/, /p, t, k/, /m, n/). When analyzing HI speech perception, some of these same confusion groups are observed. In this section, we investigate the extent of the differences between tokens of a given consonant, in terms of the confusions.

The confusions for a single consonant, /b/, are first analyzed in detail. The probability of error and the confusions (indicated by color) are shown as stacked bars for the two tokens of /bɑ/ in Fig. 4.10 (a, c). Figure 4.10 (a, c) each show the confusions for six HI ears (34L/R, 36L/R, 40L/R), at all 4 tested SNRs (0, 6, 12 [dB] SNR and quiet). The exact confusion matrix values for Fig. 4.10 are shown in Table 4.3 (a, b). For the female /bɑ/ (Fig. 4.10 (a), Table 4.3 (a)), although the HI ears have different degrees of error at different SNRs, one can observe a tendency for the listeners to respond with primarily /d, g, v/ when an error is made. For the male /bɑ/ (Fig. 4.10 (c), Table 4.3 (b)), the primary confusions are instead with /v, f/. This difference in confusion groups for the two /bɑ/ tokens is observed over all 17 of the HI ears. The average responses over all 17 HI ears and SNRs are shown for the female and male /bɑ/ tokens in Fig. 4.10 (b, d), including the proportion of correct responses; these values are shown as a function of SNR in Table 4.3 (c, d). The pie charts show that, for the female token, 27% of all HI confusions (out of 30% total) were from the /d, g, v/ confusion group. For the male token, 28% of all HI confusions (out of 35% total) were from the /f, v/ confusion group.

The average token error across all HI ears and SNRs, and the confusions that make up this error, are shown in Fig. 4.11 and Table 4.4. Here, we can again see the differences in confusion groups for the two /bɑ/ tokens (same data as in Fig. 4.10 (b, d)), but we also see average token differences for the confusion groups of /gɑ, mɑ, sɑ, ʒɑ/. Although some confusions are shared, there are distinct differences in the possible confusions for different tokens of a consonant.

The size of the confusion groups observed in the averages can be small (majority of total responses accounted for by $\leq$4 confusions), indicating, in those cases, that the majority of the responses across all HI ears and noise conditions are drawn from the same confusion group. The HI ears make

Table 4.4: A confusion matrix showing the average response [%] for each token (average taken over the 17 HI ears and 4 SNRs). Each row contains that data for a single token. Confusion probabilities $> 5\%$ are highlighted in **bold**, and probabilities $< 2\%$ are not shown. F, M subscripts denote tokens from female and male talkers.

	b	d	f	g	k	m	n	p	s	t	v	ʃ	ʒ	z
b_F	70	**15**	2	4							**7**			
b_M	65	2	**6**			2		2			**22**			
d_F		93								4				2
d_M		95								4				
f_F			73						**17**		3	3		2
g_F	3	**12**	**5**	62	2					2	**8**			2
g_M		**15**		83										
k_F					80			4		**13**				
k_M					87					**11**				
m_F						79	**9**	2			**7**			
m_M						93	**6**							
n_F						4	86				4			
n_M						**19**	80							
p_F			2		3			82		**12**				
p_M								92		3				
s_F	2		4						84				3	3
s_M									79				**8**	**12**
t_F								2	2	93				
t_M										96				
v_F	3	2	4			4	4				78			2
v_M			4		4	**5**	**5**	**11**		4	63			
ʃ$_F$									4			92		2
ʃ$_M$												96		2
ʒ$_F$				**6**								2	67	**24**
ʒ$_M$		3		**6**							**11**		63	**13**
z_F		4									**6**		**16**	70
z_M									4	2	2		**16**	74

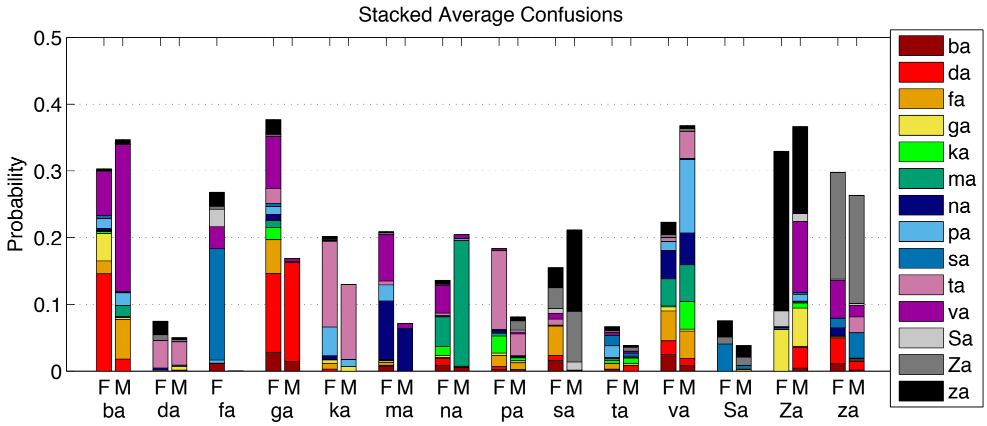


Figure 4.11: The probability of error averaged over all HI ears and SNRs for each token. The confusions that make up the error are color coded, with the size of the stacked bar indicating the probability of each confusion. Token differences in confusion group are primarily seen for /bɑ, ɡɑ, mɑ, sɑ, ʒɑ/.

similar confusions for a given token, despite the many subject differences including degree of hearing loss, age, gender, and background. This consistency across HI ears implies that the acoustic properties of each token (i.e. variable primary and conflicting acoustic cues) determine the possible HI confusions.

When examined on a token (as opposed to a consonant) level, HI ears are much more consistent in their responses; if the consonant confusions from different talkers are averaged together, HI listeners appear more "random" in their speech perception than they actually are.

4.2.6 K-Means Clustering Analysis of the Confusion Matrices

For a given consonant token, HI listeners vary widely in both the degree of error and the SNR threshold at which errors begin to occur. Despite this individual variability, we have observed that different HI ears tend to have similar token-dependent confusions, once an error is made (see Section 4.2.5, (Trevino and Allen 2013b)). If a group of HI listeners generally share a similar confusion group for a particular token, then an auditory training scheme that corrects for this confusion should be effective for that broad population of patients.

In order to explore the extent of the similarities across HI listeners, we employ the spherical k-means clustering algorithm to group the listeners based on individual consonant confusions. The data at all tested SNRs is used together in the k-means clustering analysis, since the different severities of hearing impairment across the many listeners leads to errors at different SNRs, for a given token.

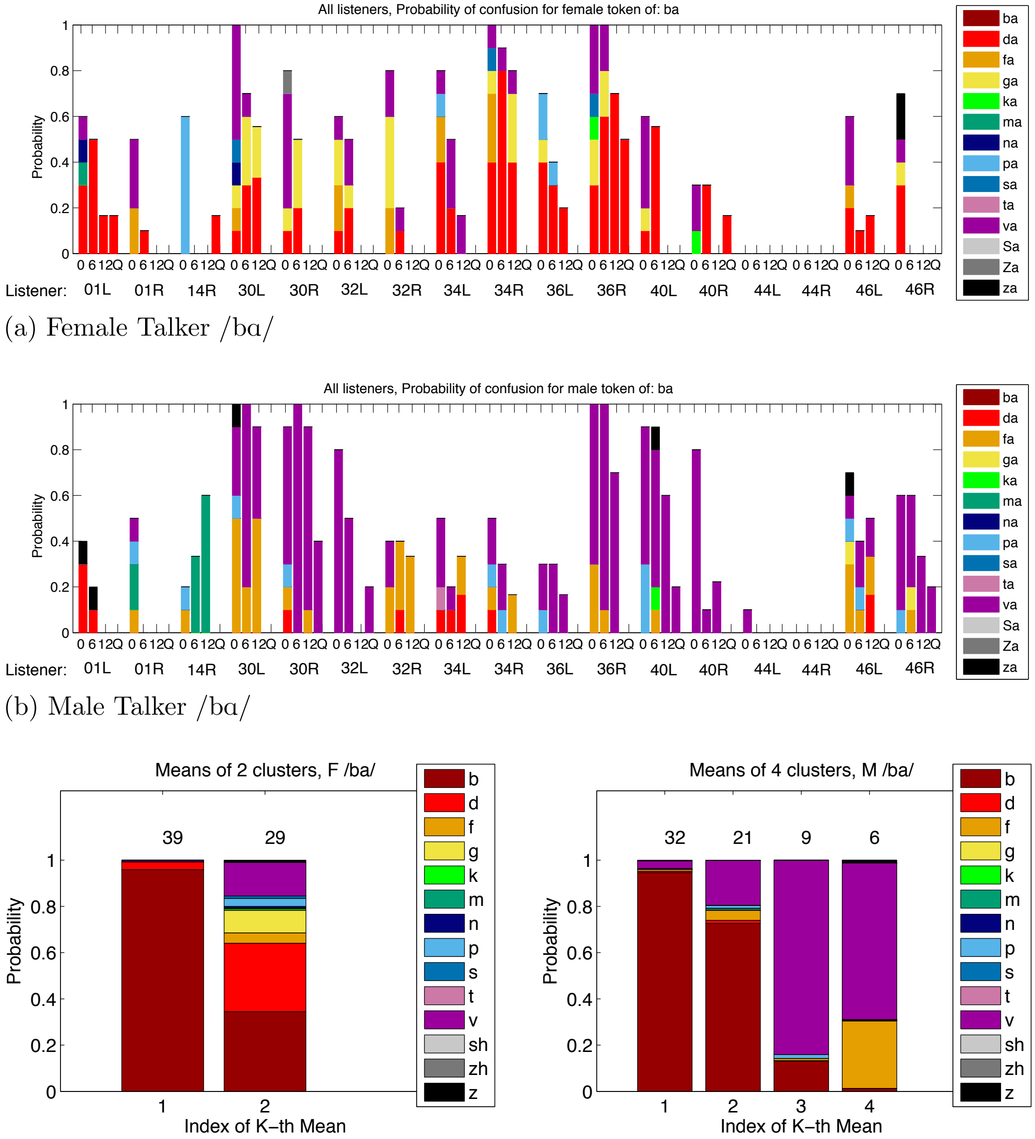
The k-means algorithm clusters HI responses that show similar consonant confusions. The number of clusters, K, for each token is determined by the G-means algorithm, which selects K iteratively based on a statistical test of the cluster distributions (Hamerly and Elkan 2004). As a result of incorporating the statistical test, the number of resulting clusters K is the number of significantly different confusion groups that are present in our data. For example, the case of two resulting clusters, $K = 2$, indicates that all of the listener data is distributed within the cluster of correct-response data points and a second cluster defined by a single confusion group. From the results in Table 4.5, we see that 17 out of the 27 tokens have $K \leq 3$, indicating that all of the HI data for these tokens falls into one of three confusion-based clusters; 22 out of 27 tokens have $K \leq 4$. This small number of clusters for the majority of tokens indicates that, generally, only a few token-dependent confusion groups are present in the HI data.

For each cluster, the primary confusions that define the k$^{\text{th}}$ mean, along with the number N of data points within each cluster, are included in Table 4.5. Results for the cluster of "correct" responses (i.e., the cluster of data with no more than 20% error over 5-10 trials) are also included. From the results in Table 4.5, we see that the confusions that define the clusters can vary across tokens of the same consonant. For example, as shown in Fig. 4.12, /d, ɡ, v/ confusions are present for the female /bɑ/ token, while only /v/ confusions dominate the responses for the male /bɑ/ token. In addition, the large number of data points, N, in the "correct" clusters of all tokens indicates that the mild-to-moderate HI listeners in this study did not have widespread errors. These are observations that have been made in Sections 4.2.3–4.2.5; this analysis shows how these observations can also be found more formally from the results of k-means clustering.

The extent of the similarity across listener responses can be quantified by the angle between the points in the spherical vector space. These angles can be expressed as direction cosines or Hellinger distances, as described in Section 4.1, and can range from 0° to 90°. The angle Θ_{x,μ_k} between a data point $\mathbf{x}$ in the k$^{\text{th}}$ cluster and the k$^{\text{th}}$ cluster mean μ_k provides a measure

Table 4.5: Clustering results for 27 CV tokens. In the "CV" column, talker gender and ID number are indicated by CV subscript and the resulting total number of clusters K is included. Each row shows the data for a single cluster; to focus on clusters with similar listeners, clusters with fewer than six data points are omitted. The main confusions comprising the k^{th} cluster means (> 5%) are listed under "k^{th} Mean (N)", with N being the number of data points within each cluster (out of 68 total). Similarities across HI ears within a cluster are quantified by the average angle, between the members of each cluster and the k$^{\text{th}}$ mean, $\widehat{\Theta}_{x,\mu_k}$.

CV	k$^{\text{th}}$ Mean (N)	$\widehat{\Theta}_{x,\mu_k}$		CV	k$^{\text{th}}$ Mean (N)	$\widehat{\Theta}_{x,\mu_k}$
bɑ$_{F101}$	k_1 : correct (39)	12°		**bɑ**$_{M112}$	k_1 : correct (32)	15°
$K = 2$	k_2 : b, d, g, v (29)	36°		$K = 4$	k_2 : b, v (21)	27°
					k_3 : b, v (9)	19°
dɑ$_{F105}$	k_1 : correct (61)	10°		**dɑ**$_{M118}$	k_1 : correct (61)	10°
$K = 3$				$K = 2$	k_2 : d, g, t (7)	25°
fɑ$_{F109}$	k_1 : correct (39)	14°		-		
$K = 2$	k_2 : f, s, v (29)	34°				
gɑ$_{F109}$	k_1 : correct (35)	8°		**gɑ**$_{M111}$	k_1 : correct (54)	10°
$K = 2$	k_2 : b, d, f, g, v (33)	48°		$K = 4$		
kɑ$_{F103}$	k_1 : correct (50)	11°		**kɑ**$_{M111}$	k_1 : correct (56)	9°
$K = 3$	k_2 : k, p, t (11)	25°		$K = 2$	k_2 : k, t (12)	23°
	k_3 : t (7)	22°				
mɑ$_{F103}$	k_1 : correct (46)	11°		**mɑ**$_{M118}$	k_1 : correct (61)	9°
$K = 3$	k_2 : m, v (12)	28°		$K = 2$	k_2 : m, n, v (7)	16°
	k_3 : m, n (10)	26°				
nɑ$_{F101}$	k_1 : correct (52)	10°		**nɑ**$_{M118}$	k_1 : correct (43)	12°
$K = 4$	k_2 : m, n (9)	25°		$K = 4$	k_2 : m, n (15)	4°
pɑ$_{F103}$	k_1 : correct (59)	13°		**pɑ**$_{M118}$	k_1 : correct (61)	12°
$K = 6$				$K = 2$	k_2 : f, p, t, z (7)	35°
sɑ$_{F103}$	k_1 : correct (55)	11°		**sɑ**$_{M120}$	k_1 : correct (45)	11°
$K = 3$	k_2 : s, ʒ, z (7)	26°		$K = 5$	k_2 : s, z (11)	10°
tɑ$_{F108}$	k_1 : correct (61)	6°		**tɑ**$_{M112}$	k_1 : correct (62)	6°
$K = 2$	k_2 : f, p, s, t (7)	40°		$K = 2$		
vɑ$_{F101}$	k_1 : correct (43)	11°		**vɑ**$_{M118}$	k_1 : correct (29)	14°
$K = 3$	k_2 : f, v (15)	32°		$K = 7$	k_2 : p, v (12)	25°
	k_3 : b, d, m, n, v (10)	38°			k_3 : m, n, v (11)	28°
ʃɑ$_{F103}$	k_1 : correct (60)	7°		**ʃɑ**$_{M118}$	k_1 : correct (65)	6°
$K = 2$	k_2 : s, ʃ, z (8)	24°		$K = 2$		
ʒɑ$_{F105}$	k_1 : correct (42)	11°		**ʒɑ**$_{M107}$	k_1 : correct (36)	13°
$K = 4$	k_2 : z (16)	18°		$K = 3$	k_2 : g, ʒ, z (17)	32°
					k_3 : v, ʒ, z (15)	38°
zɑ$_{F106}$	k_1 : correct (35)	14°		**zɑ**$_{M118}$	k_1 : correct (38)	14°
$K = 7$	k_2 : ʒ, z (11)	9°		$K = 6$	k_2 : ʒ, z (11)	18°
	k_3 : s, ʒ, z (8)	19°			k_3 : v, ʒ, z (9)	20°



(a) Female Talker /bɑ/

(b) Male Talker /bɑ/

(c) Female Talker /bɑ/

(d) Male Talker /bɑ/

Figure 4.12: The stacked confusions for each HI ear and SNR (17 ears x 4 SNRs = 68 stacked bars), for (a) the female /bɑ/ token and (b) the male /bɑ/ token. The resulting means from the k-means clustering of each data set are shown in (c) and (d), respectively. The number over each kth mean indicates the number of data points in each cluster.

of how well each mean represents the overall group of data points. This average of this measure, $\widehat{\Theta}_{x,\mu_k}$, is analogous to the variance within each cluster. Results for $\widehat{\Theta}_{x,\mu_k}$ are shown in Table 4.5. For reference, when each data point $\mathbf{x}$ is the result of 5-10 presentations, as ours are, an angle of 18°-27° lies between a vector of correct responses and a vector with a single incorrect response. Overall, the clusters defined by a larger number of primary confusions have larger $\widehat{\Theta}_{x,\mu_k}$ values. Systematic groupings of HI data in terms of consonant confusions is observed for all the tested tokens. Plots showing the data in each cluster for each token are included in Appendix D.

4.3 Repeatability

A pilot experiment was conducted approximately a year before the data reported in Chapter 4 (Han 2011). The pilot experiment collected consonant recognition data from 46 HI ears, including 16 of the 17 ears in this experiment. The speech materials of the pilot experiment were drawn from the LDC database, six tokens per consonant, with /p, t, k, f, s, ʃ, b, d, g, v, z, ʒ, m, n, θ, ð/+/ɑ/ as the test stimuli. Seventeen of the 28 individual tokens used in this experiment were also tested in the pilot experiment. Consonant recognition was tested at the same SNRs as in this data set (0, 6, 12 [dB] SNR and quiet), but with two presentations at each SNR per token. Presentations were randomized over consonant and talker, but not SNR. The pilot experiment was conducted with the same setup (observers, graphical user interface, location) as the present experiment.

The token data that is common with the pilot experiment can be used to provide a measure of the repeatability. The average error for 16 HI ears across the two experiments is significantly correlated ($\rho = 0.83$, p-val < 0.001), indicating reasonable test-retest reliability of this consonant recognition test.

4.4 Discussion

The long-term goal of this research is to investigate how low-context speech segments (such as consonants) can help to characterize an individual's hearing impairment. The variability of averaged speech measures for listeners with similar audiograms has led researchers in the past to recommend subject-by-subject analysis (Boothroyd 1984). The results of Chapter 4 reinforce this view by showing that the individual differences become even more significant when analyzing the individual token errors. Listeners with similar amounts of hearing loss and average error can differ in both the degree of error and which tokens will be perceived in error. The majority of the listeners (all but 14R) fall within the age range of 65 to 84, which is often treated as a uniform elderly population; the variability within this group implies that even when age is accounted for, one cannot treat listeners with similar audiograms as a uniform group.

We observe that some listeners with slight-to-mild hearing loss have difficulty with only a small subset of stimuli, even at the 0 [dB] SNR high-noise condition. Systematic errors with a particular consonant can lead to difficultly in understanding words that contain these consonants; for natural conversation, the need for correct perception of every phoneme decreases as context is introduced via meaningful words and sentences (Boothroyd and Nittrouer 1988). This provides a plausible explanation for why hearing impairment effects are more difficult to observe in tasks that provide context. The results of Chapter 4 are particularly relevant to clinical interpretations of average speech measures; an average speech measure is insensitive to concentrated, idiosyncratic perceptual problems with a small group of consonants. For example, three of the ears in Fig. 4.7 (a) show ≥50% error for just one stimulus and 0 error for the majority (at least 80%) of the remaining test stimuli. If the error was averaged across all stimuli for these three ears, then their large error with a very specific type of speech stimuli would be attenuated by the number of zero-error sounds; the listeners then appear to have "near normal" consonant perception.

Differences in the noise-robustness of tokens of the same consonant were observed for 16 out of 17 HI ears. These differences can be observed to the extreme that one token of a consonant has no errors at the strongest noise condition of 0 [dB] SNR while the other token of the same consonant reaches 100% error at equal or better SNRs. The average error difference, $\overline{\Delta P_e}$ (Eq. 4.3), can be used to quantify this difference in noise-robustness. Comparing the $\overline{\Delta P_e}$ values for all HI ears (Fig. 4.9 (a)) shows that across the HI ears, one of the two tokens can be consistently more robust to noise than the other. Specifically, the male tokens of /g, m, k, ʃ/ are consistently more robust to noise than the female ones, and the female tokens of /n, v/ are consistently more robust to noise than the male ones. This implies that a physical property of the signal makes one token more noise-robust than the other. To investigate possible signal properties that could lead to the differences in noise-robustness, we have considered a perceptual measure of the consonant cue intensity, the NH SNR_{90}.

For each token, the NH necessary consonant cue region can be isolated in time-frequency space with a combination of time-truncation and high/low-pass filtering psychoacoustic experiments (Phatak and Allen 2007; Li et al. 2010, 2012). The NH SNR_{90} has been found to significantly correlate with the intensity of the NH necessary cue region. Thus, the difference in NH SNR_{90} values can be used to relate the intensity of the NH consonant cue region across tokens. The NH $\Delta\mathrm{SNR}_{90}$ values are compared to the means of the HI $\overline{\Delta P_e}$ values in Fig. 4.9 (b). A significant correlation of $\rho = 0.81$ between the two measures is consistent with the assertion that the more robust NH tokens are also, generally, more robust to noise for HI listeners. This provides some evidence that the HI listeners are using the cues in the same time-frequency region as the NH listeners. If so, selective amplification of the NH consonant cue region (Kapoor and Allen 2012) would make a token more noise-robust for HI listeners. For the individual listener cases where the $\overline{\Delta P_e}$ values differ from the NH $\Delta\mathrm{SNR}_{90}$ prediction, additional signal properties may play a role.

An analysis of the confusion groups reveals additional within-consonant differences. Tokens of the same consonant can have different confusion groups for HI listeners. We observe confusion group differences for the consonants /b, g, m, s, ʒ/ across all of the HI ears in this data set. When examined on a token (as opposed to a consonant) level, one observes that HI ears are much more consistent in their responses; specifically, when the data from two tokens with equivalent error but different confusion groups is averaged together, the entropy of the HI responses increases, causing them to appear more "random". The small sizes of average confusion groups implies that the majority of HI ears share similar confusions; the consistency in token-dependent confusion groups across HI ears provides strong evidence that the HI ears, despite their many differences, may be using the same acoustic cues for perception. Further analysis of the acoustic cues that lead to particular confusions has the potential to provide critical insight into the speech perception strategies being used by HI listeners.

Across all of the HI listeners in this study, there was much more similarity between the left and right ears of a given listener than across ears of different listeners with similar audiograms, even when MCL and age are accounted for. This implies that either peripheral processing deficits that are not captured by the audiogram or degradation of central processing are the source of the inter-subject consonant perception differences. Patterson et al. (1982) concludes that the distortion observed in an impaired ear is due to peripheral damage and not central processing deficits; both can play a role.

It remains clear that speech should be incorporated in the process of hearing aid fitting (Humes et al. 1991); despite this, speech perception measures are not generally used in the fitting, only to evaluate a hearing aid after the fit. As a number of peripheral processing issues can lead to similarly elevated thresholds (Halpin and Rauch 2009), improved objective measures are required to properly characterize each HI ear. The findings of Kujawa and Liberman (2009) have shown how thresholds can be preserved even when significant damage has been done to the peripheral neural pathway. The supra-threshold processing problems (distortion) have been hypothesized to be due to reductions of temporal and spectral resolution. Analysis of which individual consonants cause error for a given HI ear should allow one to provide detailed evidence as to whether spectral and/or temporal processing problems are present.

4.5 Applications to Auditory Training

One of the primary goals of auditory training techniques is to improve the consonant recognition of listeners with sensorineural hearing loss. Training has been shown to be effective treatment in terms of both consonant and word recognition; the work of Boothroyd and Nittrouer (1988); Bronkhorst et al. (1993) generalizes these results by demonstrating how the perception of individual phones and low-context syllables predicts the perception of words and sentences. Although significant improvements can be observed from both analytic and synthetic training (Sweetow and Palmer 2005), the effects are difficult to measure and are most easily observed for listeners with the most pre-training recognition error (Walden et al. 1981). Analysis of the effects of training tend to focus on discrimination ability and overall error; the effects on consonant confusions would provide an additional dimension to the analysis, often without the collection of additional data.

In general, auditory training methodologies do not focus on the listener-specific consonant recognition deficiencies (i.e., individual differences) that are present prior to the training period. Although an identical, overarching approach is desirable when initially assessing the efficacy of a training scheme, it may not be the most beneficial for providing treatment to the patient population.

Our previous works (Trevino and Allen 2013a,b) have shown that patients with mild-to-moderate hearing loss have consonant recognition errors that are usually limited to a small subset of test consonant-vowel tokens. This indicates that, for maximum efficacy and efficiency, a targeted approach is necessary in the implementation of training programs. In addition, we have explored the significant effects of talker variability on HI perception, particularly across tokens of the same consonant (i.e., within-consonant perceptual differences). These within-consonant differences, again, highlight the need for a targeted, patient-specific approach, as well as the importance of considering token variability in the analysis of perceptual data.

The confusion matrix has been the fundamental basis for analyzing consonant recognition data for over 50 years (Miller and Nicely 1955). We have introduced a technique, k-means clustering based on the Hellinger distance, for analyzing similarity of consonant confusions. This analysis is performed on a token-by-token basis, as recommended in the conclusions of our previous works on within-consonant HI perceptual differences (Trevino and Allen 2013a,b). A more precise understanding of how HI listeners are using the acoustic cues that are available to them provides a detailed diagnosis, which could be used to refine the implementation of auditory training programs.

We have found that HI listeners with mild-to-moderate hearing loss make errors with only a small subset ($< 25\%$) of listener-dependent consonant tokens at low noise levels, although the error for these tokens can be as high as chance performance (Trevino and Allen 2013a,b). In addition, we observed significant individual variability across HI ears in terms of the degree of error and which sounds are perceived in error, despite similar hearing thresholds. These findings verify the need for an individualized approach when implementing an auditory training program. Based on our data, an individualized auditory training program would, ideally, first identify the sounds/acoustic cues that a HI listener has difficulty with in quiet and low-levels of noise, in order to focus the training appropriately. In addition, this initial test would provide a precise outcome measure after the training is completed. A test that identifies a HI listener's difficulties in terms of identifying and interpreting acoustic cues would be ideal when prescribing such a training program. A context-free, high-entropy (i.e., large response set), consonant identification task paired with a token-level analysis allows one to identify the specific acoustic cue-processing difficulties of each HI individual.

K-means clustering is a flexible tool for analyzing confusion matrix data. A clustering analysis can be conducted without averaging across tokens, consonants, SNRs or HI ears. The k-means clusters of HI data correspond to different acoustic cue weighting schemes and indicate where auditory correction or training may be useful. Although there are many individual differences across HI listeners, the small number of resulting clusters from the analysis of our data shows that the listeners are processing and interpreting the acoustic cues that are present in speech similarly. These results suggest that, once the sounds that are difficult for a HI listener are diagnosed by a speech test, a common cue correction scheme may be effective for a broad population of listeners.

CHAPTER 5

SUPPLEMENTARY NORMAL-HEARING DATA FOR THE HEARING-IMPAIRED EXP 2 TEST STIMULI

5.1 Introduction

We have observed that HI listeners make systematic token-specific confusions. Here we investigate the source of this consistency in the token-dependent confusions across HI ears. The presence of variable conflicting cues has been studied in the previous works of Li et al. (2010); Li (2010); when present, these cues cause the listener to perceive a non-target consonant when the necessary cue is masked or removed by filtering. Conflicting cues do not affect correct perception by NH listeners if the token is unmodified and presented in quiet.

The data collected from a high/low-pass filtering psychoacoustic experiment identifies existing conflicting cues in each consonant token. With high/low-pass filtering data, we can compare the conflicting cues for NH listeners to the confusions being made by HI listeners.

Twenty-eight consonant-vowel tokens were used as stimuli for the HI consonant recognition test (Experiment 2). These same 28 tokens were also included in the NH psychoacoustic noise-masking experiments of Phatak and Allen (2007); Phatak et al. (2008), but only about half of these tokens were also included as stimuli in the filtering HL07 and time truncation TR07 experiments. The experiment reported on here, which we denote HL11, completes the collection of high/low-pass data for all 28 tokens used in HI Experiment 2. The experimental methods are similar to those described for HL07 (Li et al. 2012), the specifics of which are outlined in Section 5.2.

If conflicting cues are the source of the within-consonant differences in HI confusion groups, then a number of important conclusions can be drawn. First, this would imply that the tested HI listeners are using similar perceptual cues as NH listeners. A conflicting cue causes a confusion for NH listeners when it is amplified or if the necessary cue is attenuated (Kapoor and Allen 2012); if the HI listeners are detecting the conflicting cues, then this would suggest that the HI periphery is causing the necessary cue region to be attenuated, while conserving the conflicting cue. It would also highlight the importance of adjusting hearing aid amplification so as to not amplify conflicting cues over the necessary cue, to increase the probability of the target phone's perception.

5.2 Methods

Normal-hearing listeners for this experiment were recruited who had American English as their first or primary language and were paid for their participation. IRB approval was obtained.

Fourteen isolated consonant-vowels (CV)s: /p, t, k, b, d, g, s, ʃ, f, v, z, ʒ, m, n/+/ɑ/ (no carrier phrase) was chosen from the University of Pennsylvania's Linguistic Data Consortium database (LDC-2005S22, a.k.a. *the Fletcher AI corpus*). The speech sounds were sampled at 16 [kHz]. Two tokens per CV (one male and one female) were tested. Thus, a total of 28 tokens were used (16 CVs × 2 tokens per CV). Sounds were presented diotically (both ears) through Sennheiser HD-280 PRO circumaural headphones, adjusted in level at the listener's MCL for CV tokens in 12 [dB] of WN, i.e., ≈70–75 [dB] SPL. Subjects were allowed to change the sound intensity during the experiment, which was noted in the log files. All experiments were conducted in a single-walled IAC sound-proof booth, situated in a lab with no windows, with the lab outer door shut.

The experiment is composed of 19 filtering conditions, namely one full-band condition (0.25–8 [kHz]), nine high-pass and nine low-pass conditions. The cutoff frequencies were calculated using Greenwood's inverse cochlear map function; the full-band frequency range was divided into 12 bands, each having an equal distance along the human basilar membrane. The common high- and low-pass cutoff frequencies were 3678, 2826, 2164, 1649, 1250, 939, and 697 [Hz]. To this we added the cutoff frequencies 6185, 4775 (high-pass) and 509, 363 [Hz] (low-pass). The filters were implemented as sixth-order elliptic filters having a stop-band attenuation of 60 [dB]. White noise (6-18 [dB] SNR) was added to the modified speech in order to mask out any residual cues that might still be audible (Phatak et al. 2008). The SNR was set for each token based on the NH SNR_{90} in WN to improve the accuracy of the isolated necessary cue region. Most tokens were presented at 12 [dB] SNR; tokens with an SNR_{90} of 12 ± 2 [dB] were presented at 18 [dB] SNR, and tokens with wide frequency range results from HL07 and an SNR_{90} below 0 [dB] were presented at 6 [dB] SNR.

A MATLAB® program was written for the stimulus presentation and data collection. A baseline session, using non-test tokens, was run for each listener at the beginning of his or her participation in the experiment to verify that their error rates were within NH ranges; NH ranges were determined from the results of Singh and Allen (2012). A mandatory practice session, with feedback, was given at the beginning of each experiment. Speech tokens are randomized across talkers, conditions and tokens. Following each stimulus presentation, listeners responded by clicking on the button that is labeled with the CV that they perceived. If the listener heard only noise or a consonant that is not one of the 14 choices, the listener was instructed to click a “Not Sure” button. Frequent (e.g., 20 minute) breaks were encouraged to prevent test fatigue. Subjects were allowed to repeat each token up to three times, after which they were forced to choose a response. The waveform is played via a SoundBlaster 24 bit sound card in a PC Intel computer, running MATLAB® via Ubuntu Linux.

5.2.1 Differences from HL07

A number of small but significant differences in the methods of HL07 (Li et al. 2012) and the methods used in this round of data collection, denoted HL11, are enumerated below.

1. HL07 did not include baseline or unmodified practice sessions (i.e., the practice session played filtered tokens, many of which had removed the primary cue). Due to this, we believe that a small number of listeners had trouble understanding which orthographic symbols corresponded to certain phones, most commonly /ʒ, θ, ð/. For HL11, correct perception of all unmodified consonants by each NH listener was verified at the beginning of the test, during the baseline session. In addition, the practice session presented unmodified tokens, with feedback.

2. In HL07, the example word in the GUI for /ʒ/ was "azure", a word that is commonly mispronounced by native English speakers. The example words "rouge" and "measure" were substituted in for clarification.

3. HL07 computed the SNR based on the standard deviation, this implementation uses the sliding exponential window method outlined in the master's thesis of Cvengros (2011).

4. The automatic technique for removing speech-free segments in speech samples for HL07 (remove_dead_spots.m) damaged some of the tokens by removing low-level consonant regions. The updated implementation had the speech-free regions of the 28 test tokens removed by hand, using Praat.

5. During the HL07 experiment, listeners were instructed to guess a consonant if they heard "any sound that wasn't pure noise", and report "Noise Only" only when nothing but noise is heard. In this implementation, listeners are instructed to report "Not Sure" if they hear any sound that is not one of the possible consonant options. This update prevents false morphs – due to biased guessing – from being introduced into the data.

6. In HL07, each listener heard the experimental tokens in the same randomized order. In the HL11 implementation, each listener has his or her own randomized presentation token order.

7. Due to high average error rates, the consonants /θ, ð/ were not included in HI Experiment 2, but they were included as responses in HL07. The consonants /θ, ð/ are not included as possible responses in HL11.

5.3 Results

5.3.1 Normal-Hearing Frequency Cues

The results of the high/low-pass experiment are shown (labeled HL11) and compared to the existing data from HL07 in Figs. 5.1-5.14. As noted above, there is no HL07 data to report for about half of the tokens. Results from low-pass filtering are shown as solid lines with circle markers; results from high-pass filtering are shown as dashed lines with + markers. “Not sure” responses are indicated by thin black lines. Cue regions are determined using the procedure outlined in Section 3.1.5. For reference, since low-level WN was added to each speech sample, the SNR_{90} values are included in the caption of each figure; values are computed from the WN masking experiment of Phatak et al. (2008).

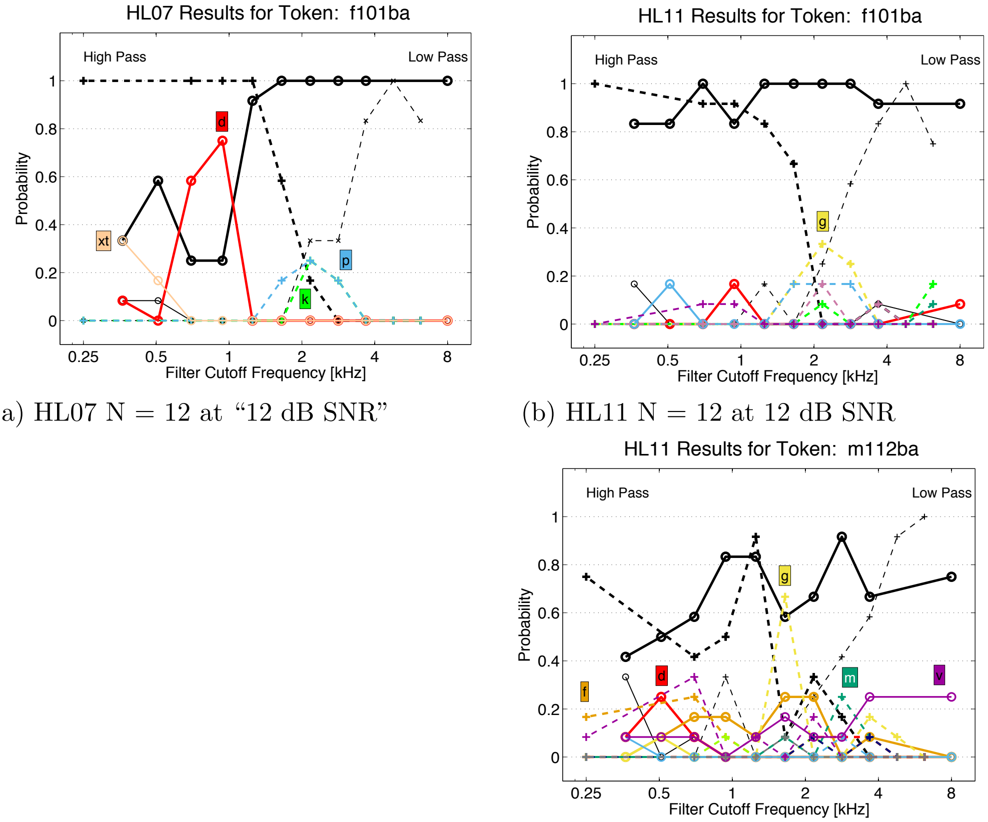
High/low-pass filtering results for /bɑ/

For the /bɑ/ tokens from talkers f101 and m112 (Fig. 5.1), the cue region lies between 0.5–1.5 [kHz]. No conflicting cues are present above 4 [kHz]. For the /bɑ/ from talker f101 (Fig. 5.1 (a, b)), the potential for a low-frequency /dɑ/ conflicting cue can be observed, as well as weak mid-frequency /kɑ, gɑ, pɑ/ conflicting cues. The HL11 results show $>$ 80% performance when this token is low-pass filtered at a cutoff of 0.35 [kHz], indicating that some of the low-frequency primary cue region remained audible after filtering.

For the /bɑ/ from talker m112 (Fig. 5.1 (c)), the results are less clear due to wide-band scores $<$ 80%. Out of the 28 tokens used in HI Experiment 2, the m112 /bɑ/ token was the only one that could not reach 100% scores at any tested levels of WN (except in quiet). Presenting only this token in quiet would have made it easy to pick out by the test subjects, even once the consonant cue region was removed by filtering; therefore, 18 [dB] of noise was added. As a consequence, the scores at the wide-band conditions for the m112 /bɑ/ are below 100%. At low noise levels (18 [dB] SNR), confusions with /fɑ, vɑ/ are observed at the wide-band conditions, and a mid-frequency /gɑ/ conflicting cue can be observed from the high-pass data.

High/low-pass filtering results for /dɑ/

For the /dɑ/ tokens from talkers f105 and m118 (Fig. 5.2), the cue region indicated by the high/low-pass data lies between 1–4 [kHz]. For the /dɑ/ token from talker f105 (Fig. 5.2 (a)), the primary cue region ranges from 1–6 [kHz], the /tɑ/ confusions seen in the high-pass filtering data may be due to the removal of low-frequency voice-onset cues. In addition, a weak mid-frequency /ga/ and low-frequency /fɑ, vɑ/ conflicting cues can be observed.



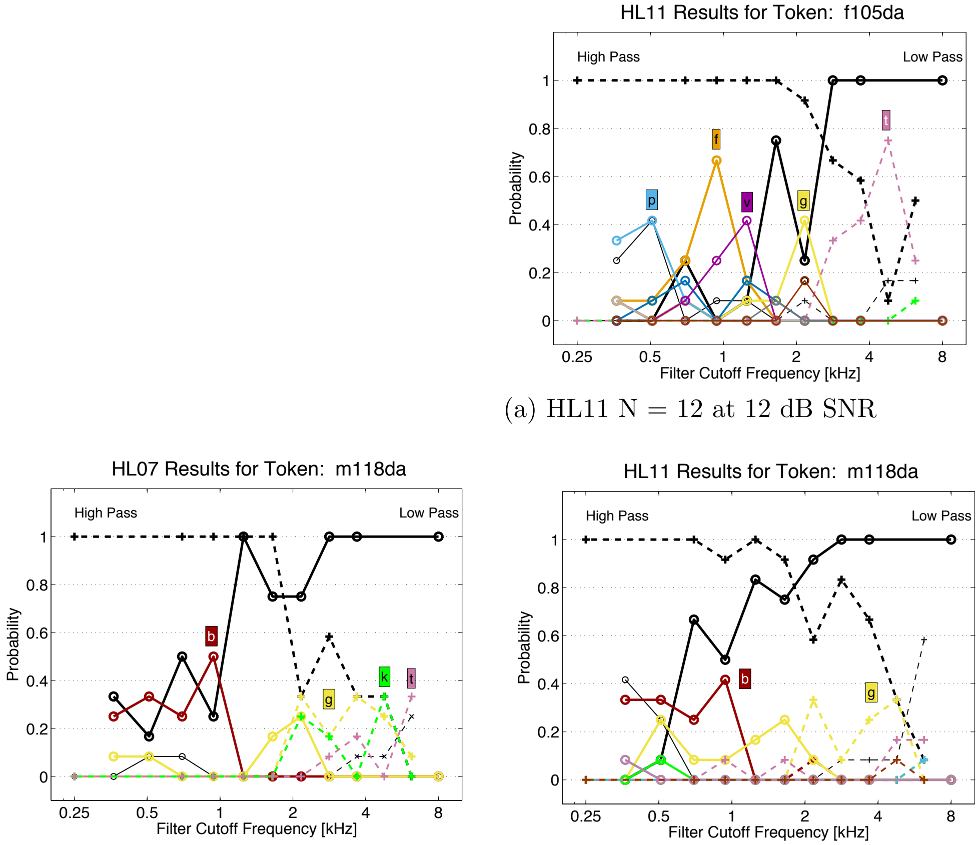
(a) HL07 N = 12 at "12 dB SNR"

(b) HL11 N = 12 at 12 dB SNR

(c) HL11 N = 12 at 18 dB SNR

Figure 5.1: High/low-pass filtering results for the /bɑ/ token from (a, b) female talker 101 and (c) male talker 112. For the female token the $\mathrm{SNR}_{90} \approx -1$ [dB], and for the male token the $\mathrm{SNR}_{90} \approx 13$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008). In the case of the male token, correct perception in the wide-band case only reached 100% in quiet.

The /dɑ/ token from talker m118 (Fig. 5.2 (b, c)) has a lower-frequency primary cue region than that of f105. High-pass filtering of this token leads to weak confusions with /kɑ, gɑ/, the source of which is uncertain. A low-frequency /bɑ/ conflicting cue is the only strong confusion produced from low-pass filtering.



(a) HL11 N = 12 at 12 dB SNR

(b) HL07 N = 12 at "12 dB SNR"

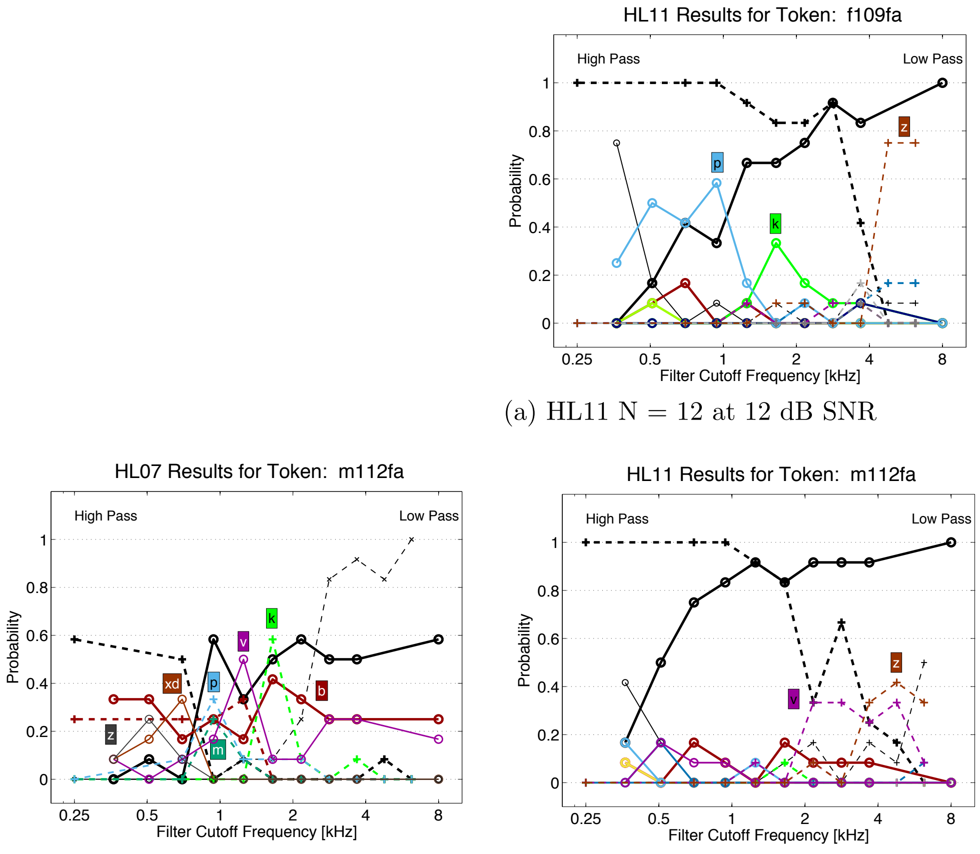
(c) HL11 N = 12 at 12 dB SNR

Figure 5.2: High/low-pass filtering results for the /dɑ/ token from (a) female talker 105 and (b, c) male talker 118. The female token has an $\mathrm{SNR}_{90} \approx -7$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx -2$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

High/low-pass filtering results for /fɑ/

For the /fɑ/ tokens from talkers f109 and m112 (Fig. 5.3), the cue region indicated by the filtering data lies between 1–3 [kHz]. The token from talker f109 (Fig. 5.3 (a)) has no low-frequency fricative conflicting cues, but can be morphed to a /zɑ/ if high-pass filtered above 4 [kHz]. Low-frequency /pɑ, kɑ/ burst confusions indicate that this token has a burst-like release before the onset of voicing.

The /fɑ/ token from talker m112 (Fig. 5.1 (b, c)) has highly differing HL07 and HL11 wide-band results, due to a pre-processing step in HL07 that mistakenly removed the majority of this token's low-level friction region. From the HL11 results, we see that the primary cue region is slightly lower in frequency than that of talker f109. No low-frequency conflicting cues are observed for the m112 token; high-pass filtering above 2 [kHz] causes weak confusions with /vɑ, zɑ/.



(a) HL11 N = 12 at 12 dB SNR

(b) HL07 N = 12 at "12 dB SNR"

(c) HL11 N = 12 at 18 dB SNR

Figure 5.3: High/low-pass filtering results for the /fɑ/ token from (a) female talker 109 and (b, c) male talker 112. The female token has an $\mathrm{SNR}_{90} \approx 0$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx 10$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008). In the case of the male token, the low scores in the HL07 wide-band conditions are due to the removal of the frication region by remove_dead_spots.m.

High/low-pass filtering results for /gɑ/

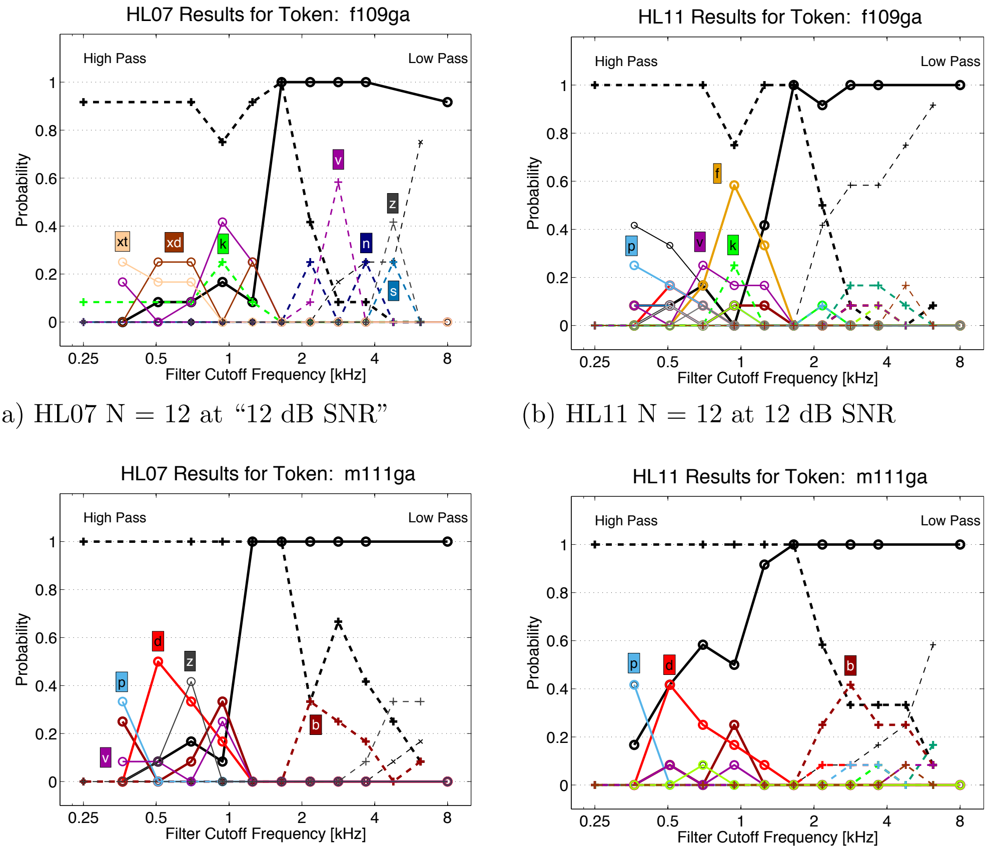
For the /gɑ/ tokens from talkers f109 and m111 (Fig. 5.4), the cue region indicated by the filtering data lies between 1–2 [kHz]. The token from talker f109 (Fig. 5.4 (a, b)) shows a large number of possible confusions across the data from HL07 and HL11. The high-pass filtering (> 2 [kHz]) confusions in HL07 appear to have been due to the forced-choice nature of the task (i.e., select a consonant if any speech-like sound is heard); the HL11 listeners were allowed to select "not sure" for speech-like sounds, which was the primary response in the cases of high-pass filtering above 2 [kHz]. Weak low-frequency /fɑ, vɑ/ conflicting cues are observed.

The /gɑ/ from talker m111 (Fig. 5.4 (c, d)) has confusions with /dɑ, pɑ/ when low-pass filtered, and /bɑ/ confusions when high-pass filtered. This outcome does not match the expected frequency locations of stop consonant burst cues, which typically labels /dɑ/ as a high-frequency cue, and /bɑ/ as a low-frequency cue; this, paired with the low probability of confusion, indicates that strong conflicting cues are not the source of confusions. These confusions occur at the frequencies where the majority but not all of the mid-frequency burst is removed by filtering, leaving a severely degraded mid-frequency cue; the primable voiced-burst responses of the listeners show that the remaining cues identify a burst but are not sufficient for clear identification of a consonant.

High/low-pass filtering results for /kɑ/

For the /kɑ/ tokens from talkers f103 and m111 (Fig. 5.5), the cue region indicated by the filtering data lies between 1–2 [kHz]. The token from talker f103 (Fig. 5.5 (a, b)) has a strong /pɑ/ conflicting cue below 1.2 [kHz], which causes a morph at some of the low-pass filtering conditions. A high-frequency /tɑ/ conflicting cue can also be seen in the data.

The /kɑ/ from talker m111 (Fig. 5.5 (c, d)) also has a low-frequency /pɑ/ conflicting cue. The high-pass filtering data shows a mid-frequency /gɑ/ confusion, possibly caused by the removal of the low-frequency voice-onset cue; in which case, the mid-frequency burst cue for this token would range from 1–2.8 [kHz].



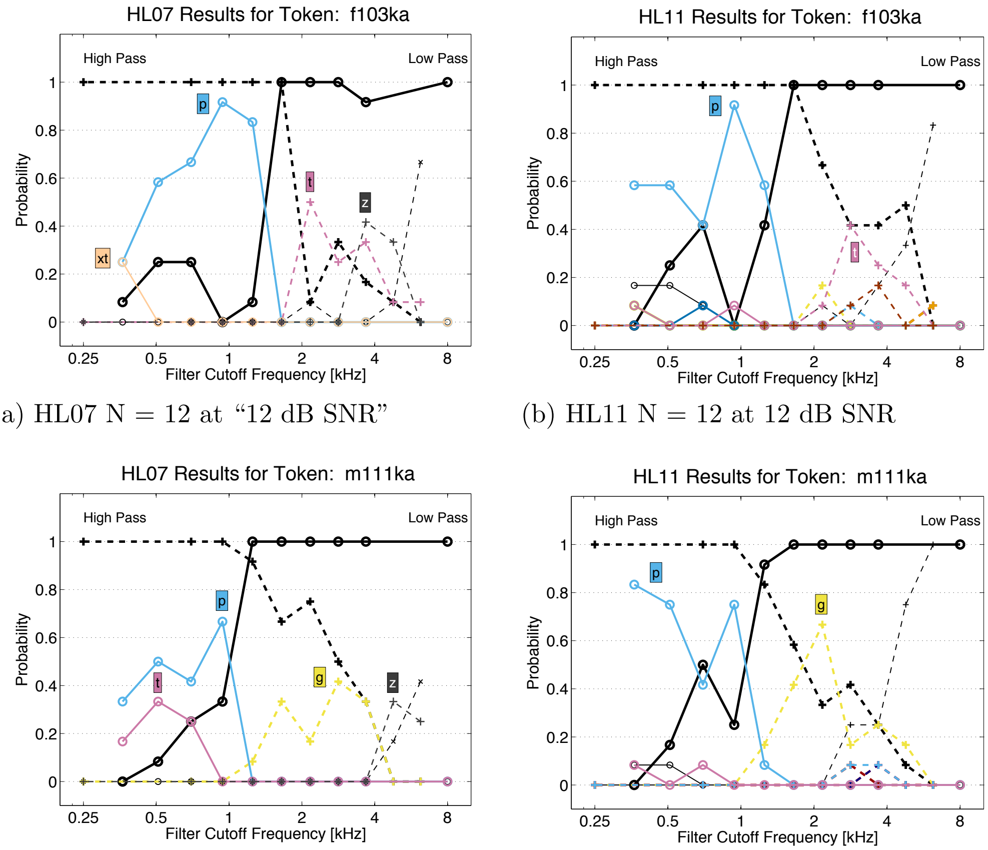
(a) HL07 N = 12 at “12 dB SNR”

(b) HL11 N = 12 at 12 dB SNR

(c) HL07 N = 12 at “12 dB SNR”

(d) HL11 N = 12 at 12 dB SNR

Figure 5.4: High/low-pass filtering results for the /gɑ/ token from (a, b) female talker 109 and (c, d) male talker 111. The female token has an $\mathrm{SNR}_{90} \approx 4$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx -1$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).



(a) HL07 N = 12 at “12 dB SNR”

(b) HL11 N = 12 at 12 dB SNR

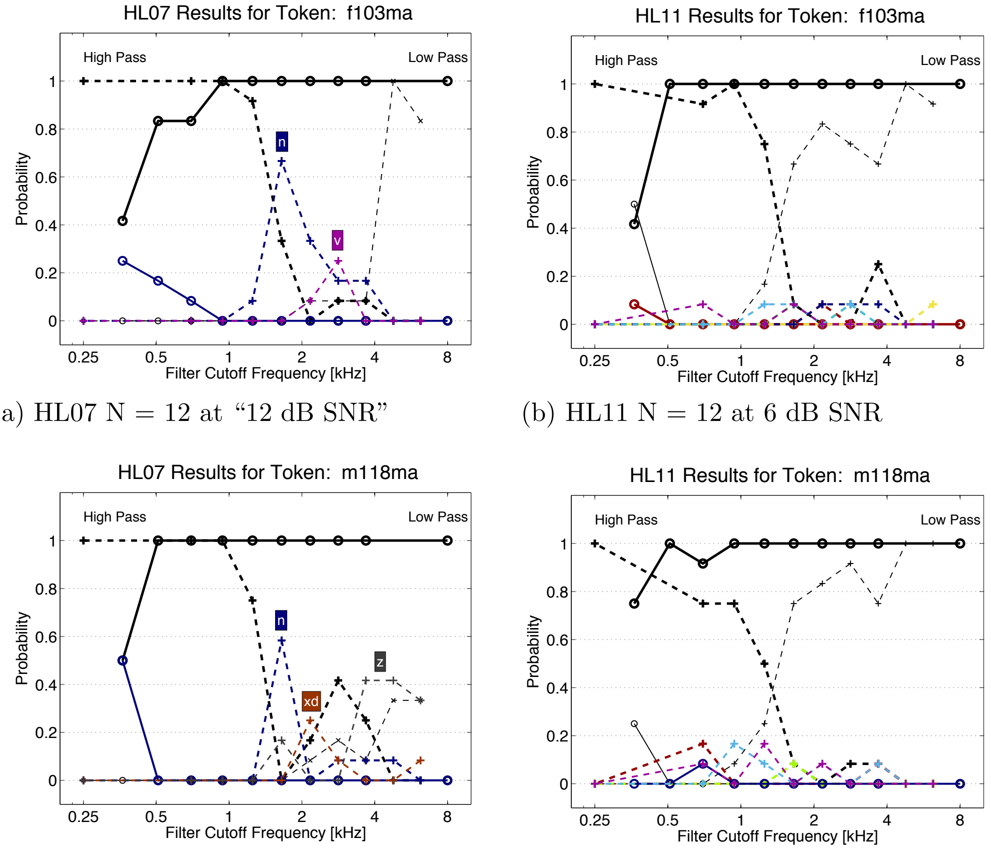
(c) HL07 N = 12 at “12 dB SNR”

(d) HL11 N = 12 at 12 dB SNR

Figure 5.5: High/low-pass filtering results for the /kɑ/ token from (a, b) female talker 103 and (c, d) male talker 111. The female token has an $\mathrm{SNR}_{90} \approx -1$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx -4$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

High/low-pass filtering results for /mɑ/

For the /mɑ/ tokens from talkers f103 and m118 (Fig. 5.6), the cue region indicated by the filtering data lies between 0.5–1.3 [kHz]. Although there appear to be conflicting cues at higher frequencies based on the HL07 data, the HL11 data shows that, when given the option, listeners select "not sure" as the primary response to all high-pass filtering data at and above 1.6 [kHz]. Thus, a /nɑ/ confusion may result if listeners are forced to guess under degraded conditions, but it would not be due to the presence of a clear /nɑ/ conflicting cue.



(a) HL07 N = 12 at "12 dB SNR"

(b) HL11 N = 12 at 6 dB SNR

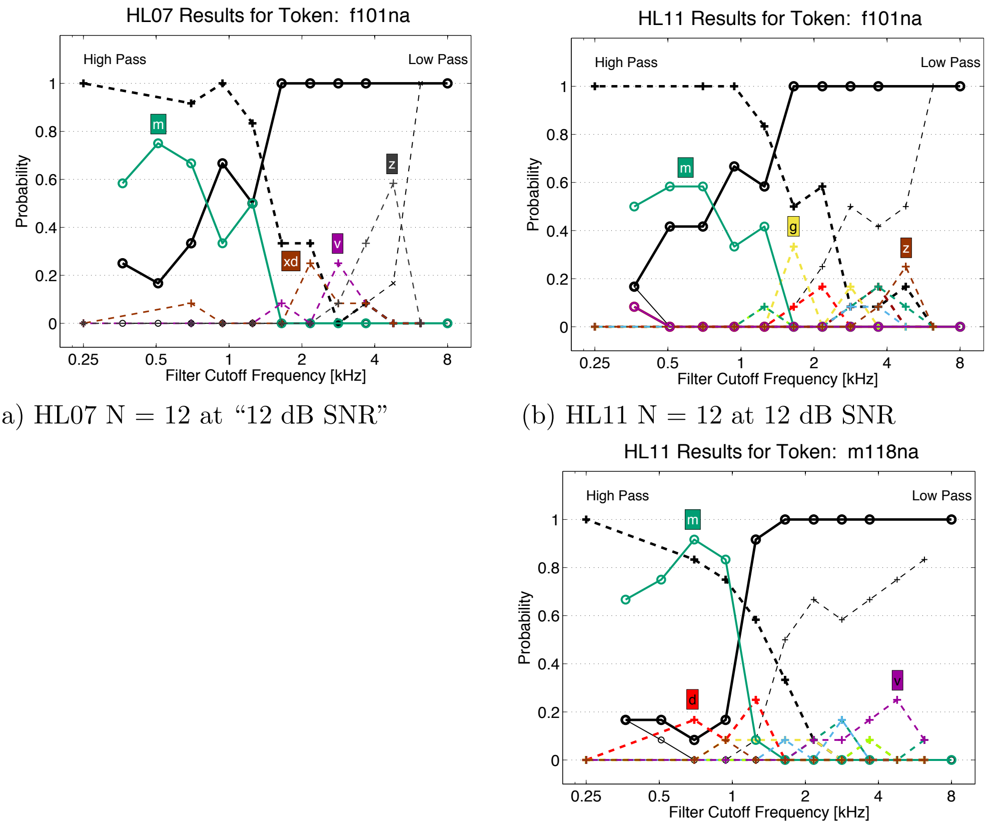
(c) HL07 N = 12 at "12 dB SNR"

(d) HL11 N = 12 at 6 dB SNR

Figure 5.6: High/low-pass filtering results for the /mɑ/ token from (a, b) female talker 103 and (c, d) male talker 118. The female token has an $\text{SNR}_{90} \approx -9$ [dB], and the male token has an $\text{SNR}_{90} \approx -11$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

High/low-pass filtering results for /nɑ/

For the /nɑ/ tokens from talkers f101 and m118 (Fig. 5.7), the cue region indicated by the filtering data lies between 0.7–1.3 [kHz]. For both tokens, a strong /mɑ/ conflicting cue lies between 0.25–1 [kHz], which produces a complete morph in the low-pass filtering data of the token from talker m118. For the /nɑ/ token from talker f101 (Fig. 5.7 (a, b)), the high-pass filtering data shows low-probability /gɑ, zɑ/ confusions; the primary listener response for high pass filtering above 2 [kHz] was "not sure", indicating a lack of clear consonant cues in this frequency range. The filtering data for the /nɑ/ token from talker m118 (Fig. 5.7 (c)) shows a strong low-frequency /mɑ/ conflicting cue; the high-pass results are primarily "not sure" responses above 1.3 [kHz] with low-probability (≤25%) /dɑ, vɑ/ confusions.



(a) HL07 N = 12 at "12 dB SNR"

(b) HL11 N = 12 at 12 dB SNR

(c) HL11 N = 12 at 12 dB SNR

Figure 5.7: High/low-pass filtering results for the /nɑ/ token from (a, b) female talker 101 and (c) male talker 118. The female token has an $\mathrm{SNR}_{90} \approx -6$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx -1$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

High/low-pass filtering results for /pɑ/

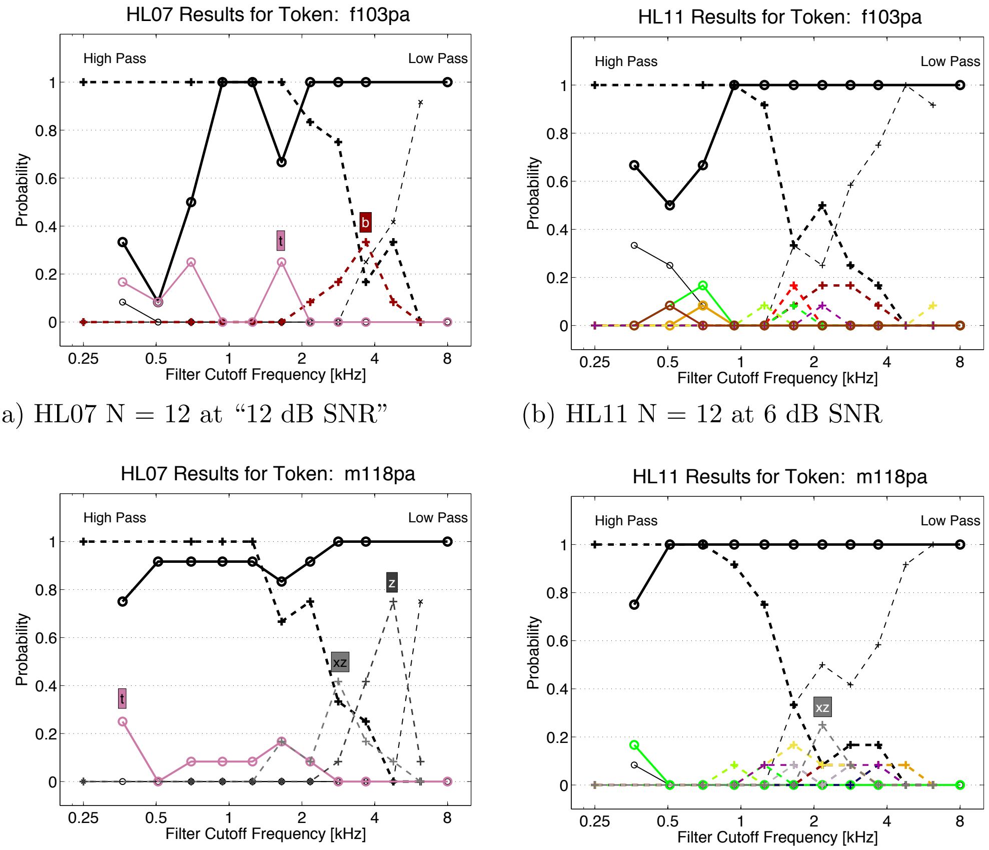
For the /pɑ/ tokens from talkers f101 and m118 (Fig. 5.8), the cue region indicated by the filtering data lies between 0.35–1.3 [kHz]. The /pɑ/ token from talker f103 (Fig. 5.8 (a, b)) has no conflicting cues that can be observed from the filtering data; the primary listener response when the cue region is removed by high/low-pass filtering is "not sure" or "noise only".

The /pɑ/ token from talker m118 (Fig. 5.8 (c, d)) has a slightly lower frequency range than that of the token from talker f101. The m118 token may have low-intensity /ʒa, za/ conflicting cues between 2–8 [kHz], since these confusions can be seen in the HL07 high-pass filtering data at 12 [dB] SNR, but not in the HL11 data at 6 [dB] SNR. Alternatively, these HL07 high-pass filtering confusions may be due to bias from the forced-choice nature of the task, the HL11 responses at these frequencies are primarily "not sure".

High/low-pass filtering results for /sɑ/

For the /sɑ/ tokens from talkers f103 and m120 (Fig. 5.9), the cue region indicated by the filtering data lies between 3.5–8 [kHz]. The /sɑ/ token from talker f103 (Fig. 5.9 (a)) shows confusions with /zɑ/ when high-pass filtered at cutoff frequencies from 1–4.8 [kHz]; these confusions do not seem to be due to a conflicting cue, and instead may be related to the loss of the lower-frequency vowel region. Low-pass filtering of the female token at cutoff frequencies from 0.5–3.7 [kHz] causes /fɑ/ confusions, and, at some frequencies, complete morphs, indicating the presence of a conflicting cue within this frequency range. A lower-probability /pɑ/ confusion in the low-pass filtering data at 0.5 [kHz] is due to the remaining voicing onset after the higher-frequency frication is removed by filtering.

The /sɑ/ token from talker m120 (Fig. 5.9 (b)) has a slightly lower frequency range for the primary cue than the token from talker f103. The low-pass filtering data shows /fɑ, pɑ/ confusions, similar to the results for the token from talker f103. In addition, the high-pass filtering data only shows confusions at the 6 [kHz] cutoff, when the majority of the high-frequency frication is removed, leading to confusions with /zɑ/.



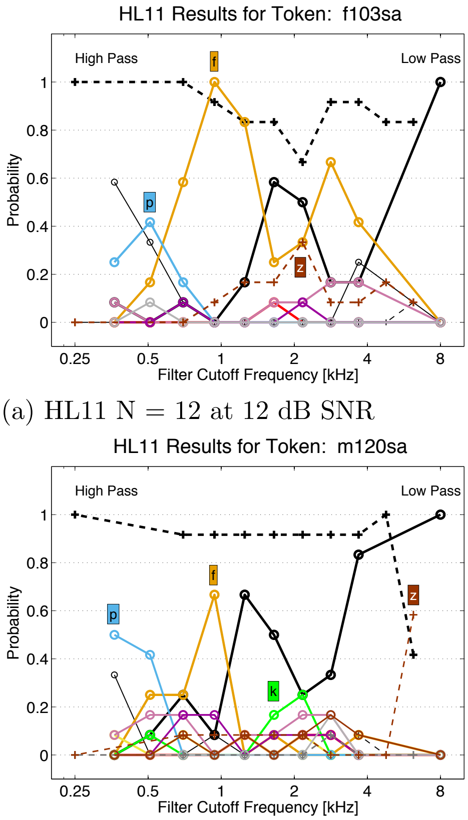
(a) HL07 N = 12 at “12 dB SNR”

(b) HL11 N = 12 at 6 dB SNR

(c) HL07 N = 12 at “12 dB SNR”

(d) HL11 N = 12 at 6 dB SNR

Figure 5.8: High/low-pass filtering results for the /pɑ/ token from (a, b) female talker 103 and (c, d) male talker 118. The female token has an $\text{SNR}_{90} \approx -1$ [dB], and the male token has an $\text{SNR}_{90} \approx -1$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).



(a) HL11 N = 12 at 12 dB SNR

(b) HL11 N = 12 at 12 dB SNR

Figure 5.9: High/low-pass filtering results for the /sɑ/ token from (a) female talker 103 and (b) male talker 120. The female token has an $\mathrm{SNR}_{90} \approx 0$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx 0$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

High/low-pass filtering results for /ʃɑ/

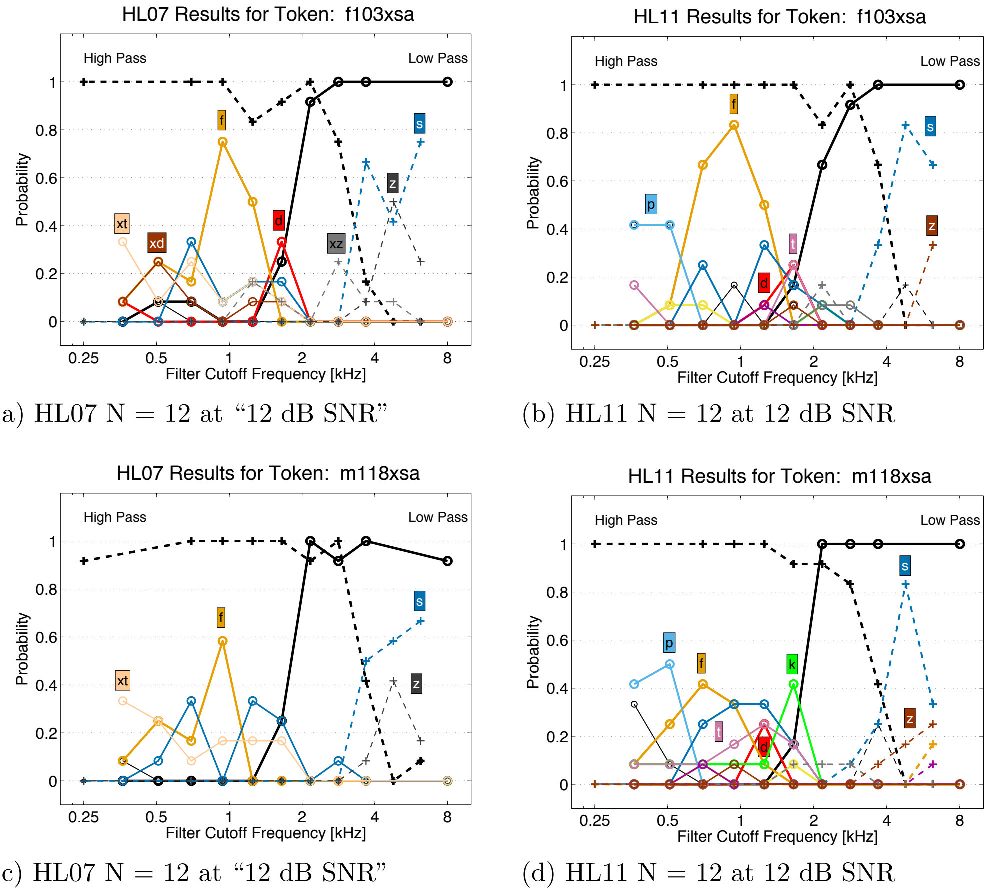
For the /ʃa/ tokens from talkers f103 and m118 (Fig. 5.10), the cue region indicated by the filtering data lies between 2–3.5 [kHz]. The token from talker f103 (Fig. 5.10 (a, b)) has /fɑ, pɑ/ confusions when the primary cue is removed by low-pass filtering, similar to the results for both /sɑ/ tokens. The confusion with /fɑ/, despite a lack of frication energy <1.4 [kHz], indicates that the only remaining energy, the voicing onset, can provide some cue for /fɑ/. Confusions with plosives are due to the remaining release at the onset of voicing. The high-pass filtering data indicates that there is a /sɑ, zɑ/ conflicting cue above 4 [kHz].

The /ʃa/ token from talker m118 (Fig. 5.10 (c, d)), again, has /fɑ, pɑ/ confusions when low-pass filtered below 1 [kHz]. A mid-frequency /kɑ/ confusion at the 2 [kHz] low-pass filtering condition is due to a small portion of remaining frication which resembles a burst. A high-frequency /sɑ, zɑ/ conflicting cue can be observed from the high-pass filtering data, similar to that of the token from talker f103.

High/low-pass filtering results for /tɑ/

For the /tɑ/ tokens from talkers f108 and m112 (Fig. 5.11), the cue region indicated by the filtering data lies between 4–6 [kHz]. Both tokens have low-frequency /pɑ/ and mid-frequency /kɑ/ confusions in their low-pass filtering data. The /tɑ/ token from talker f108 (Fig. 5.11 (a, b)) has a stronger /pɑ/ conflicting cue, with this confusion dominating the majority of the low-pass data.

The /tɑ/ token from talker m112 (Fig. 5.11 (c, d)) has a /kɑ/ conflicting cue that is intense enough to cause a morph at the 1.6–2.1 [kHz] cutoff low-pass filtering conditions. The high-pass filtering data for the m112 token shows a /dɑ/ confusion at the cutoffs above 2.1 [kHz], this confusion is not due to a conflicting cue, but, instead, to a loss of the voice-onset timing cue which discriminates a /t/ from a /d/ burst.



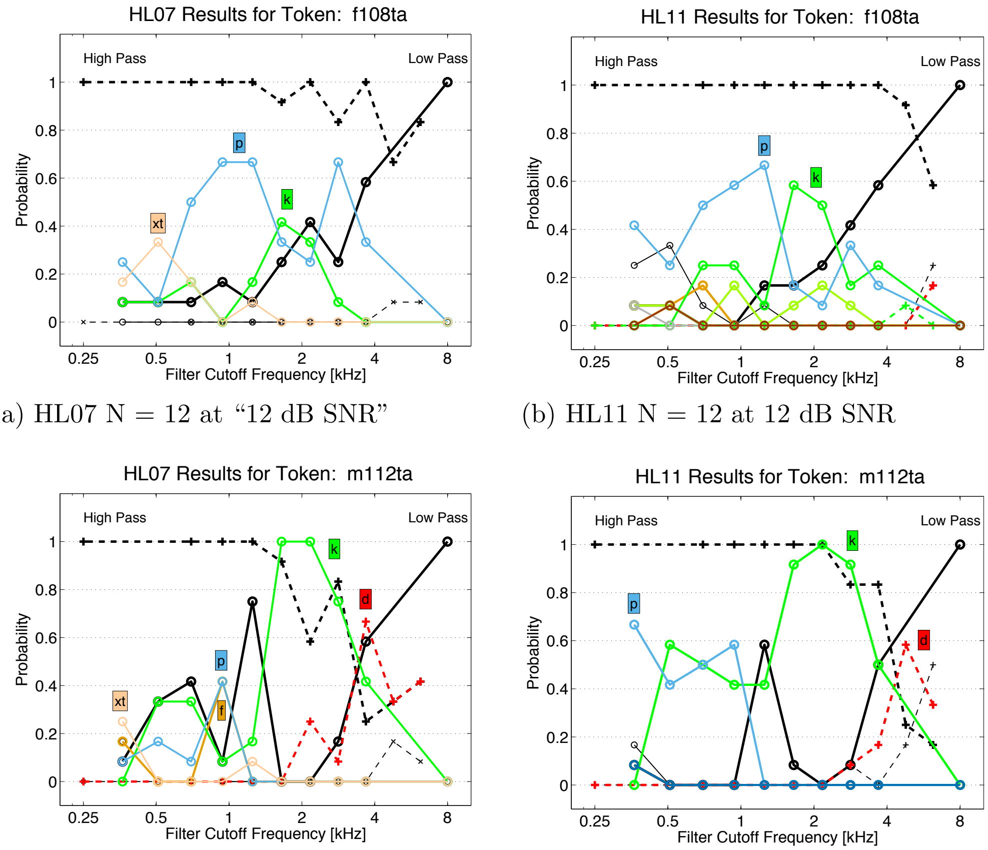
(a) HL07 N = 12 at “12 dB SNR”

(b) HL11 N = 12 at 12 dB SNR

(c) HL07 N = 12 at “12 dB SNR”

(d) HL11 N = 12 at 12 dB SNR

Figure 5.10: High/low-pass filtering results for the /ʃɑ/ token from (a, b) female talker 103 and (c, d) male talker 118. The female token has an $\mathrm{SNR}_{90} \approx -1$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx -6$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).



(a) HL07 N = 12 at “12 dB SNR”

(b) HL11 N = 12 at 12 dB SNR

(c) HL07 N = 12 at “12 dB SNR”

(d) HL11 N = 12 at 12 dB SNR

Figure 5.11: High/low-pass filtering results for the /tɑ/ token from (a, b) female talker 108 and (c, d) male talker 112. The female token has an $\text{SNR}_{90} \approx -1$ [dB], and the male token has an $\text{SNR}_{90} \approx 0$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

High/low-pass filtering results for /vɑ/

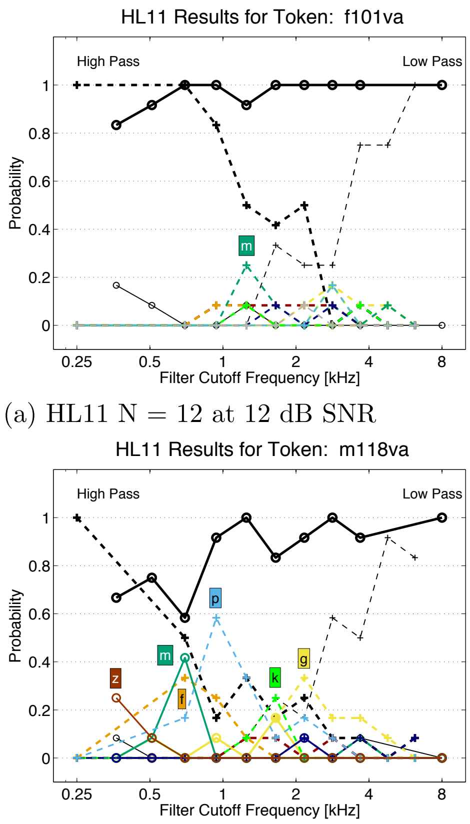
For the /vɑ/ tokens from talkers f101 and m118 (Fig. 5.12), the cue region indicated by the filtering data lies between 0.35–1 [kHz]. The token from talker f101 (Fig. 5.12 (a)) shows a weak (25%) confusion with /mɑ/ at the high-pass filtering condition where the necessary frication is mostly but not completely removed. No conflicting cues are observable from the filtering data.

The data for the token from talker m118 (Fig. 5.12 (b)) shows a wider array of confusions. The high-pass filtering data for the token from talker m118 shows some confusions with /fɑ/ at cutoffs between 0.5–1 [kHz], indicating that energy below these frequencies contains a necessary voicing cue. Low-pass filtering within this 0.5–1 [kHz] range produces /mɑ/ confusions, indicating that the sustained voicing cues are preserved, but cues for discriminating place of articulation have been degraded. High-pass filtering with a cutoff frequency between 1–3 [kHz] produces confusions with plosives /gɑ, kɑ, pɑ/, which could be due to conflicting cues or the release following the frication within this frequency range.

High/low-pass filtering results for /ʒɑ/

For the /ʒa/ tokens from talkers f105 and m107 (Fig. 5.13), the cue region indicated by the filtering data lies between 1.2–2.8 [kHz]. The token from talker f105 (Fig. 5.13 (a)) has a high-frequency /zɑ/ conflicting cue, observable in the high-pass filtering data. The /gɑ/ confusion at the 2 [kHz] low-pass filtering condition is due to the small remaining portion of the frication near the onset of voicing, which resembles a burst. The low-pass filtering data mainly shows /vɑ, zɑ/ confusions below 2 [kHz], due to remaining frication cues.

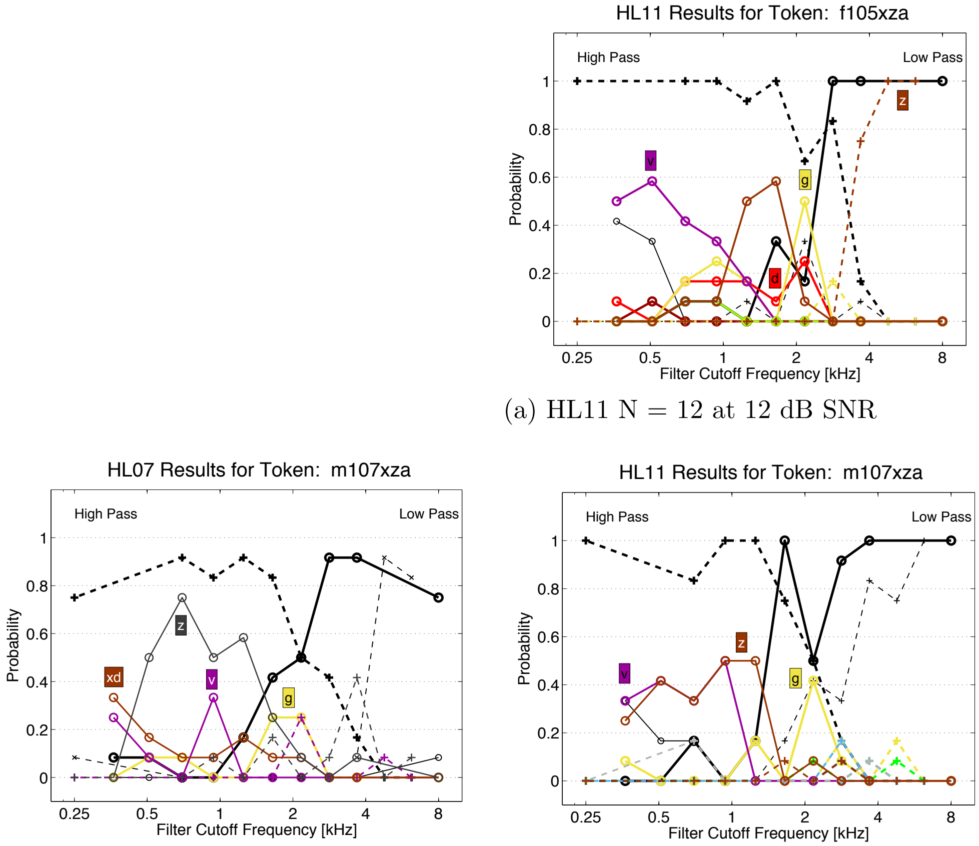
The /ʒa/ token from talker m107 (Fig. 5.13 (b, c)) does not show confusions at the high-pass filtering conditions. At the low-pass filtering conditions, similar to the results for the token from talker f105, /vɑ, zɑ, gɑ/ confusions can be observed.



(a) HL11 N = 12 at 12 dB SNR

(b) HL11 N = 12 at 12 dB SNR

Figure 5.12: High/low-pass filtering results for the /vɑ/ token from female talker 101 and male talker 118. The female token has an $\mathrm{SNR}_{90} \approx -3$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx 8$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

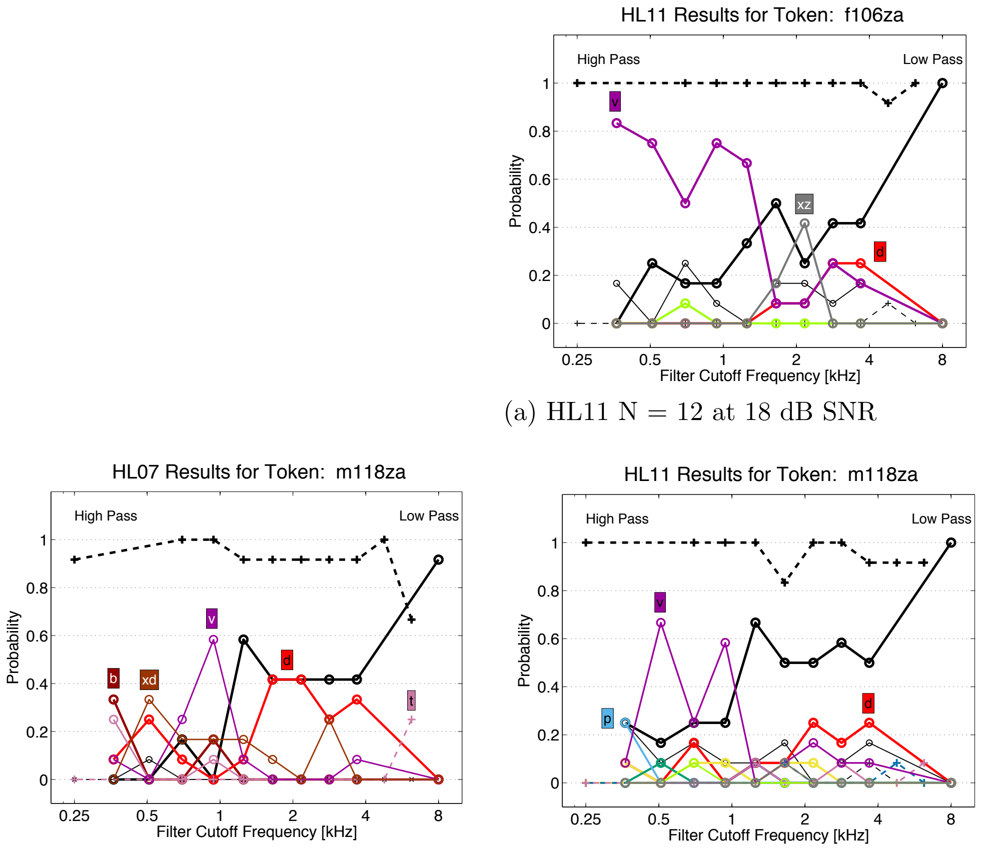


(a) HL11 N = 12 at 12 dB SNR

(b) HL07 N = 12 at “12 dB SNR”

(c) HL11 N = 12 at 12 dB SNR

Figure 5.13: High/low-pass filtering results for the /ʒɑ/ token from female talker 105 and male talker 107. The female token has an $\mathrm{SNR}_{90} \approx -7$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx 5$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).



(a) HL11 N = 12 at 18 dB SNR

(b) HL07 N = 12 at “12 dB SNR”

(c) HL11 N = 12 at 12 dB SNR

Figure 5.14: High/low-pass filtering results for the /zɑ/ token from female talker 106 and male talker 118. The female token has an $\mathrm{SNR}_{90} \approx 12$ [dB], and the male token has an $\mathrm{SNR}_{90} \approx -1$ [dB]; these values are computed from the WN masking data of Phatak et al. (2008).

High/low-pass filtering results for /zɑ/

For the /zɑ/ tokens from talkers f106 and m118 (Fig. 5.14), the cue region indicated by the filtering data lies between 6–8 [kHz]. The /zɑ/ token from talker f106 (Fig. 5.14 (a)) shows some confusions with /dɑ/ when the majority of the intense high-frequency frication is removed by low-pass filtering, leaving an intense high-frequency voicing onset that resembles a voiced burst. Lower-frequency confusions with /ʒɑ, vɑ/ are due to remaining but low-intensity frication noise regions just before the onset of voicing.

The /zɑ/ token from talker m118 (Fig. 5.14 (b, c)), similar to the female token, shows /dɑ/ confusions when the majority of the high-frequency frication region is removed by low-pass filtering, leaving an intense, high-frequency, burst-like onset at voicing. Below 1 [kHz], the low-pass filtering data shows confusions with /vɑ/, due to the remaining low-intensity frication at these frequencies.

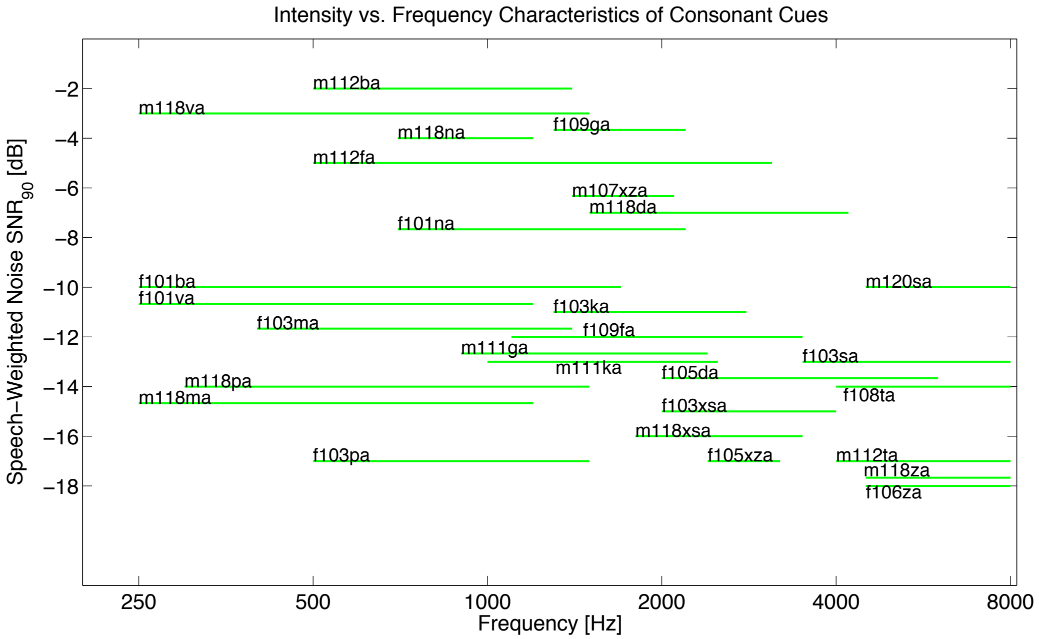


Figure 5.15: Visual summary of noise-masking and high/low-pass filtering results for NH listeners. Each line represents the results for a single token, labeled by the LDC filename. The ordinate is SNR_{90} of each token, in SWN, computed from the noise-masking data of Phatak and Allen (2007). The frequency range of each primary cue, based on the results of HL11, is marked by the location of the green line along the abscissa.

5.4 Summary of Cue Frequency Range and Noise Masking Results

A visual summary of the high/low-pass filtering primary cue results of HL11 and the SWN masking results of Phatak and Allen (2007) is shown in Fig. 5.15. The results of Régnier and Allen (2008); Li et al. (2010, 2012) show that the SNR_{90} provides an objective psychophysical measure that is correlated to the relative intensity of the primary cue regions. Thus, this figure provides an estimate of the relative cue intensities, for all of the HI Experiment 2 test tokens. The SNR_{90}s for SWN are used in this summary to match the conditions of HI Experiment 2.

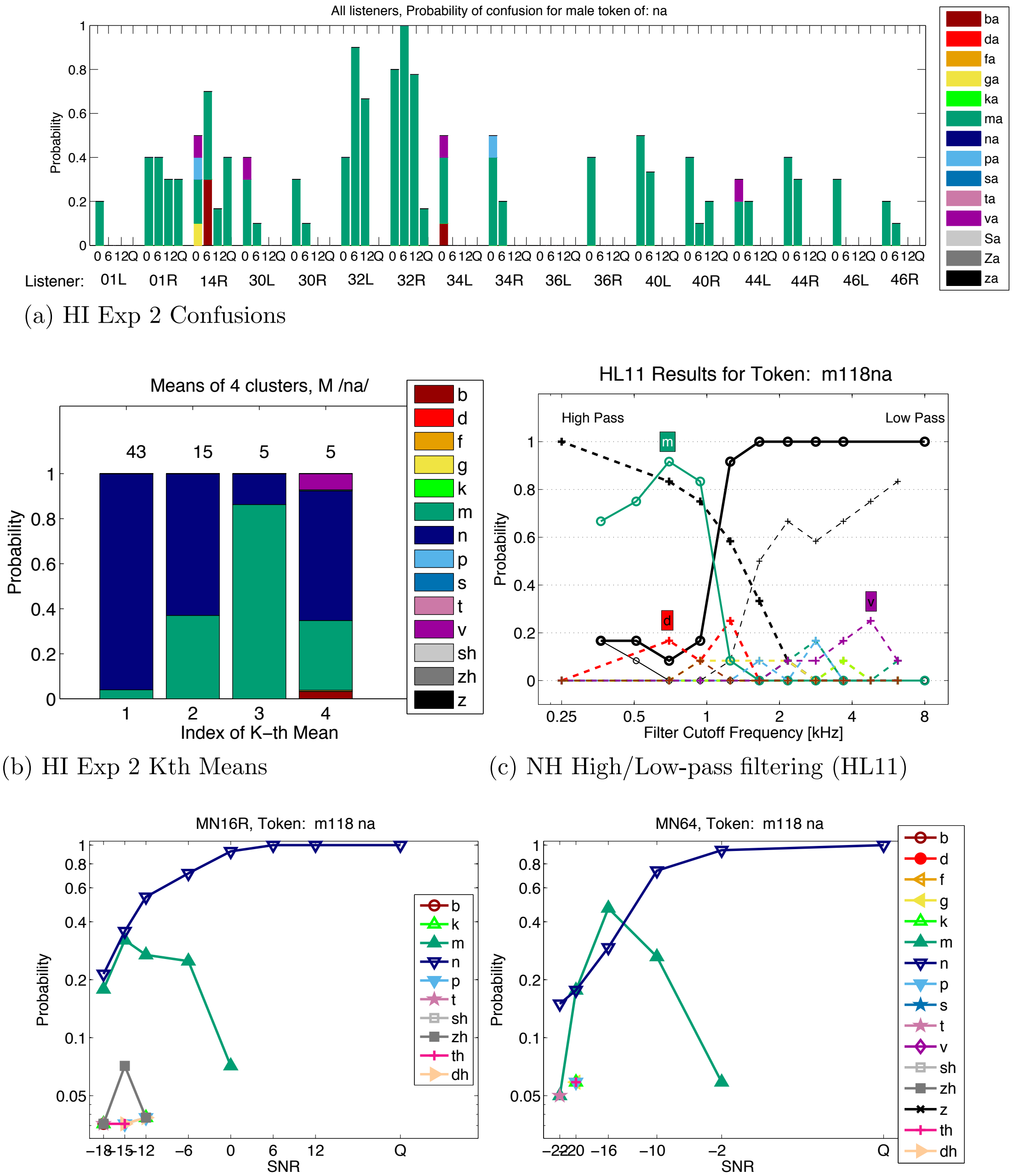
This intensity and frequency characterizations of the tokens could be used both in the design of hearing tests and in diagnosis of hearing impairment. The SNR_{90} variability across tokens of the same consonant allows one to control for the difficulty of a hearing test in noise; the lower the SNR_{90} of a token, the more intense the primary acoustic cue (i.e., easier to hear). Characterizing the variability in the frequency locations of the primary consonant cues can be used to create a fine-tuned speech test which investigates perceptual issues as a function of frequency and intensity.

5.5 Discussion: Insights on the Hearing-Impaired Consonant Confusions

The raw consonant confusion data for all tokens and HI ears is compiled in Appendix D, along with the k-means cluster means, the NH noise-masking results, and the NH high/low-pass filtering results. As an example, this data is shown in Fig. 5.16 for the /nɑ/ token from talker m118. By comparing the NH and HI data, we can determine which confusions resemble NH responses when the primary cue is masked by noise (data from MN16R and MN64) and which confusions may be due to conflicting cues (data from HL11).

5.5.1 Noise-Masking

For the majority of tokens in Experiment 2, the HI confusions in SWN can also be observed in the NH noise-masking data, at lower SNRs. The HI



(a) HI Exp 2 Confusions

(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure 5.16: Overview of data collected for the /nɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars, with each color representing a confusion. (b) The k-means cluster means; in this case $K = 4$, with 43 cases in cluster 1, 15 in cluster 2, and 5 each in clusters 3 and 4. (c) The NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.

Table 5.1: For each token, the HI confusions which are also observed in NH noise-masking data.

			NH Match
CV	Talker	HI Conf	Noise Type
bɑ	m112	f	WN
		v	WN, SWN
fɑ	f109	b	WN
		v	WN, SWN
gɑ	f109	d, v, z	WN
gɑ	m111	d	WN
fɑ	f109	b, v	WN
kɑ	f103	p, t	WN
kɑ	m111	p, t	WN
mɑ	f103	n, v	WN
mɑ	m118	n	WN
nɑ	f101	d, v	WN
nɑ	m118	m	WN, SWN
pɑ	f103	f, k, t	WN
pɑ	m118	f, k, t	WN
sɑ	f103	f	WN
		z	SWN
sɑ	m120	z	WN, SWN
ʃɑ	f103	z	WN
		ʒ	WN, SWN
ʃɑ	m118	ʒ	WN, SWN
tɑ	f108	p	WN
tɑ	m112	d, k	WN, SWN
vɑ	f101	b, f	WN, SWN
		n	SWN
vɑ	m118	t	WN
		m, p	WN, SWN
ʒɑ	f105	g	WN
		z	WN, SWN
ʒɑ	f105	v, z	WN
		d, g	WN, SWN
zɑ	f106	d, ʒ	WN
zɑ	m118	d, s, v	WN
		t	WN, SWN

confusions that are most common for each token are compared to the NH responses at all SNRs, in WN and SWN.

For example, in Fig. 5.16 (a, b), we see that the most common HI confusion for the /nɑ/ token from talker m118 is /m/. We then examine the NH noise-masking confusion patterns in Fig. 5.16 (d, e), to see if either of these data sets shows confusions with /m/. The fact that both the WN and SWN results show confusions with /m/ indicates that, when the /n/ cue for this token is degraded by noise, NH listeners make the same confusions as HI listeners. Based on this data, we can hypothesize that the /m/ confusions could be due to either a conflicting cue that has not been fully masked by the noise or an /m, n/ ambiguity caused by partial masking of the primary cue.

Table 5.1 lists the HI confusions which are also observed for NH listeners in WN and/or SWN. The row for the m118 /nɑ/ example shows that the HI /m/ confusion is also observed for NH listeners in WN and SWN.

Overall, we see that the HI confusions are most like the NH confusions in WN. This may be due to the fact that high-frequency sloping loss is the predominant audiometric configuration for the HI listeners who participated in Experiment 2 (16 out of 17 ears). A high-frequency sloping hearing loss more severely attenuates high-frequency cues; WN, similarly, also masks higher frequencies more severely than SWN.

5.5.2 High/Low-Pass Filtering

The results of the high/low-pass filtering data shows where conflicting cues are present. Conflicting cues cause confusions for NH listeners when the primary cue region is masked or removed by filtering. The filtering conditions at which NH listeners make the same confusions as HI listeners are listed in Table 5.2. Confusions due to conflicting cues that are observed in the NH high/low-pass filtering data, but are not observed in the NH noise-masking data, may indicate that those conflicting cues are masked by noise before the primary cue (i.e., at SNRs $>$ SNR$_{90}$).

For the example /nɑ/ token from talker m118, we examine the NH high/low-pass filtering data in Fig. 5.16 (c) to see if the HI /m/ confusion is present at any filtering condition. The results show that NH listeners make

Table 5.2: For each token, the HI confusions which are also observed in NH high-pass (HP) or low-pass (LP) filtering data.

		HI	NH Match
CV	Talker	Conf	Filter Cutoff
bɑ	f101	d	LP 0.7–1 kHz
		g	HP 2.2–2.8 kHz
fɑ	f109	s, z	HP 4.7–6.2 kHz
gɑ	f109	f, v	LP 0.7–1.2 kHz
kɑ	f103	p	LP 0.35–1.2 kHz
		t	HP 2.8–4.7 kHz
kɑ	m111	p	LP 0.35–1 kHz
nɑ	f101	m	LP 0.35–1.2 kHz
nɑ	m118	m	LP 0.35–1 kHz
sɑ	f103	f	LP 0.35–1.2 kHz
ʃɑ	f103	f	LP 0.7–1.2 kHz
	f103	s	HP 3.7–6.2 kHz
	f103	z	HP 6.2 kHz
ʃɑ	m118	s, z	HP 3.7–6.2 kHz
tɑ	f108	p	LP 0.35–3.7 kHz
		k	LP 0.7–3.7 kHz
tɑ	m112	p	LP 0.35–1 kHz
		k	LP 0.5–3.7 kHz
vɑ	f101	m	HP 1.2 kHz
vɑ	m118	m	LP 0.7 kHz
		p	HP 0.7–1.2 kHz
		k	HP 1.6 kHz
ʒɑ	f105	g	LP 0.7–1, 2.2 kHz
		z	HP 3.7–6.2 kHz
ʒɑ	m107	g	LP 1.2, 2.2 kHz
		v	LP 0.35–1 kHz
zɑ	f106	v	LP 0.35–3.7 kHz
		ʒ	LP 1.6–2.2 kHz
zɑ	m118	v	LP 0.5–1 kHz

/m/ confusions for this token at the low-pass (LP) filtering conditions with cutoff frequencies of 0.35–1 [kHz].

NH listeners show confusions and, at times, morphs when conflicting cues are selectively amplified while the primary cue is attenuated (Li 2010). In the HI ear, attenuation of cues at particular frequencies could have a similar effect, allowing the conflicting cues to be more salient than the primary. One way to test this hypothesis in future experiments would be to remove conflicting cues from the experimental tokens and record the effect on HI confusions.

5.5.3 Other Sources of Consonant Confusions

In Table 5.3, we list other possible sources of the HI confusions, primarily truncation-like modifications of the consonant region by a HI ear due to loss of audibility or loss of the nonlinear onset transient. Truncation of voiceless fricatives can lead to voicing confusions and, when truncated to the duration of a burst, confusions with stop consonants (Li et al. 2012).

Table 5.3: HI confusions from Experiment 2 that can be caused in NH listeners with either truncation of a section of the consonant region or the loss of the voicing onset cue.

CV	Talker	Conf	Source
dɑ	f105	t	Loss of voicing cue
dɑ	m118	t	Loss of voicing cue
sɑ	f103	z	Truncation
sɑ	m120	z	Truncation
ʃɑ	f103	ʒ	Truncation
ʃɑ	m118	ʒ	Truncation
tɑ	m112	d	Loss of voicing cue
vɑ	f101	f	Truncation
ʒɑ	f105	g	Truncation
ʒɑ	m107	g	Truncation
zɑ	f106	d	Truncation
zɑ	m118	d, t	Truncation

Additional possible sources these consonant confusions, that have not yet been explored, include loss of forward masking, loss of temporal fine structure, reduced frequency selectivity, and changes in the central processing of cues at the auditory cortex.

5.5.4 Conclusions

HL11 completes the high/low-pass filtering data collection for all tokens used in HI Experiment 2, providing a characterization of the NH acoustic cues, as a function of frequency. To fully complete the 3DDS analysis for all test tokens, a truncation experiment would be needed. A prediction of truncation results that could match HI confusions, based on previous 3DDS findings, is provided in Table 5.3.

This comparison of NH and HI data is a first step toward understanding the consonant perception scheme being used by HI listeners. The information in Fig. 5.15 could be used to investigate if a HI patient shows loss of primary cues at a particular frequency or below an SNR_{90} threshold. When the intensity-frequency characterization of the primary cues cannot explain the individual token errors of a HI patient, additional signal properties must play a role. One such additional signal property is the presence of token-dependent conflicting cues.

The NH high/low-pass filtering data and NH masking noise data show which conflicting cues may be present in each individual speech token. This allows us to hypothesize which HI confusions may be due to clear conflicting cues and which may be due to the ambiguity of a degraded primary cue. An assumption of this approach is that the HI and NH listeners decode acoustic cues similarly, which is supported by our observations for the majority of HI ears in this dissertation (Trevino and Allen 2013a,b). The token-specific confusion data, along with the observation that the mild-to-moderate listeners have problems with only a subset of tokens, could be exploited to develop a fine-tuned, individualized diagnosis of each HI ear.

APPENDIX A

COMPUTING THE AUDIOGRAM FIT PARAMETERS

Three parameters are computed for the piecewise linear fit to the audiogram that is reported in Chapter 5. The resulting fits are shown in Figs. A.1–A.3; an overlay plot of the fits for 17 HI ears is shown in Fig. A.4. The fit is formulated as

$$h = \begin{cases} h_0 & \text{if } f \leq f_0 \\ h_0 + s_0(log_2(f/f_0)) & \text{if } f > f_0 \end{cases} \tag{A.1}$$

where h is the hearing loss in [dB] and f is frequency in [kHz]. The parameter f_0 estimates the frequency at which the sloping loss begins; h_0 estimates the low-frequency ($f \leq f_0$) flat loss in [dB]; s_0 estimates the slope of the high-frequency loss in [dB/octave]. The parameters are fit to minimize the mean-squared-error (MSE).

To find the equations for h_0 and s_0, the parameter derivative of mean-squared error is set to zero and solved for the relevant parameter. Frequency is transformed to a decibel scale $log_2(f)$. For h_0 the calculation is derived as follows

$$MSE = (\frac{1}{n}\sum_{i=1}^{n}(h_i - h_0 - s_0 log_2(f_i))^2, s_0 = 0$$

$$\frac{dMSE}{dh_0} = \frac{-2}{n}\sum_{i=1}^{n}(h_i - h_0) = 0$$

$$h_0 = \frac{1}{n}\sum_{i=1}^{n} h_i$$

with $n = 10$, the number of frequencies measured for the audiogram. For s_0 the calculation is derived as follows

$$MSE = (\frac{1}{n}\sum_{i=1}^{n}(h_i - h_0 - s_0(log_2(f_i) - log_2(f_0))^2$$

$$\frac{dMSE}{ds_0} = \frac{-2}{n}\sum_{i=1}^{n}(h_i - h_0 - s_0(log_2(f_i) - log_2(f_0)))(log_2(f_i) - log_2(f_0)) = 0$$

$$s_0 = \frac{\sum_{i=1}^{n}(h_i log_2(f_i/f_0)) - h_0 \sum_{i=1}^{n} log_2(f_i/f_0)}{\sum_{i=1}^{n}(log_2(f_i/f_0))^2}$$

Finally, the third parameter f_0 was determined by iterating over the 8 possible frequencies at which the breakpoint could be located and setting the final value based on the minimum MSE.

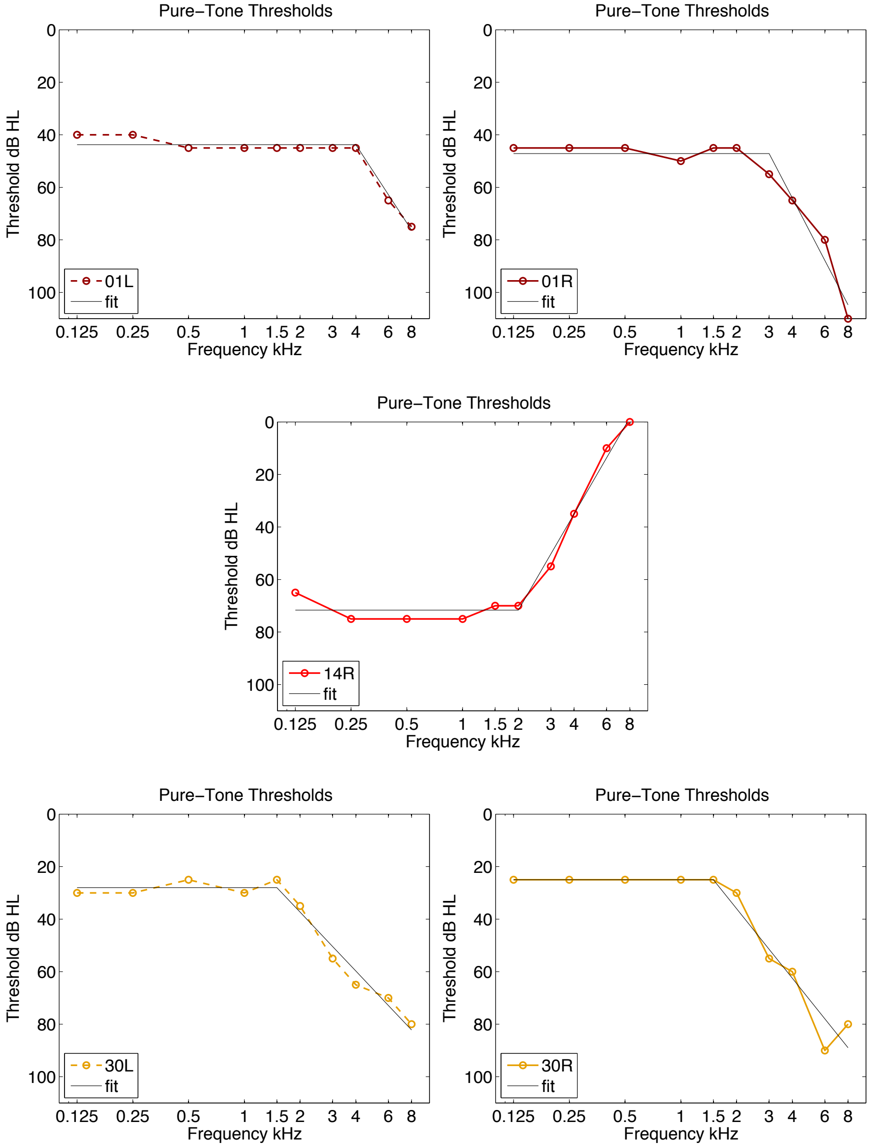


Figure A.1: Pure-tone thresholds and the three-parameter linear fit for HI listeners 01, 14 and 30.

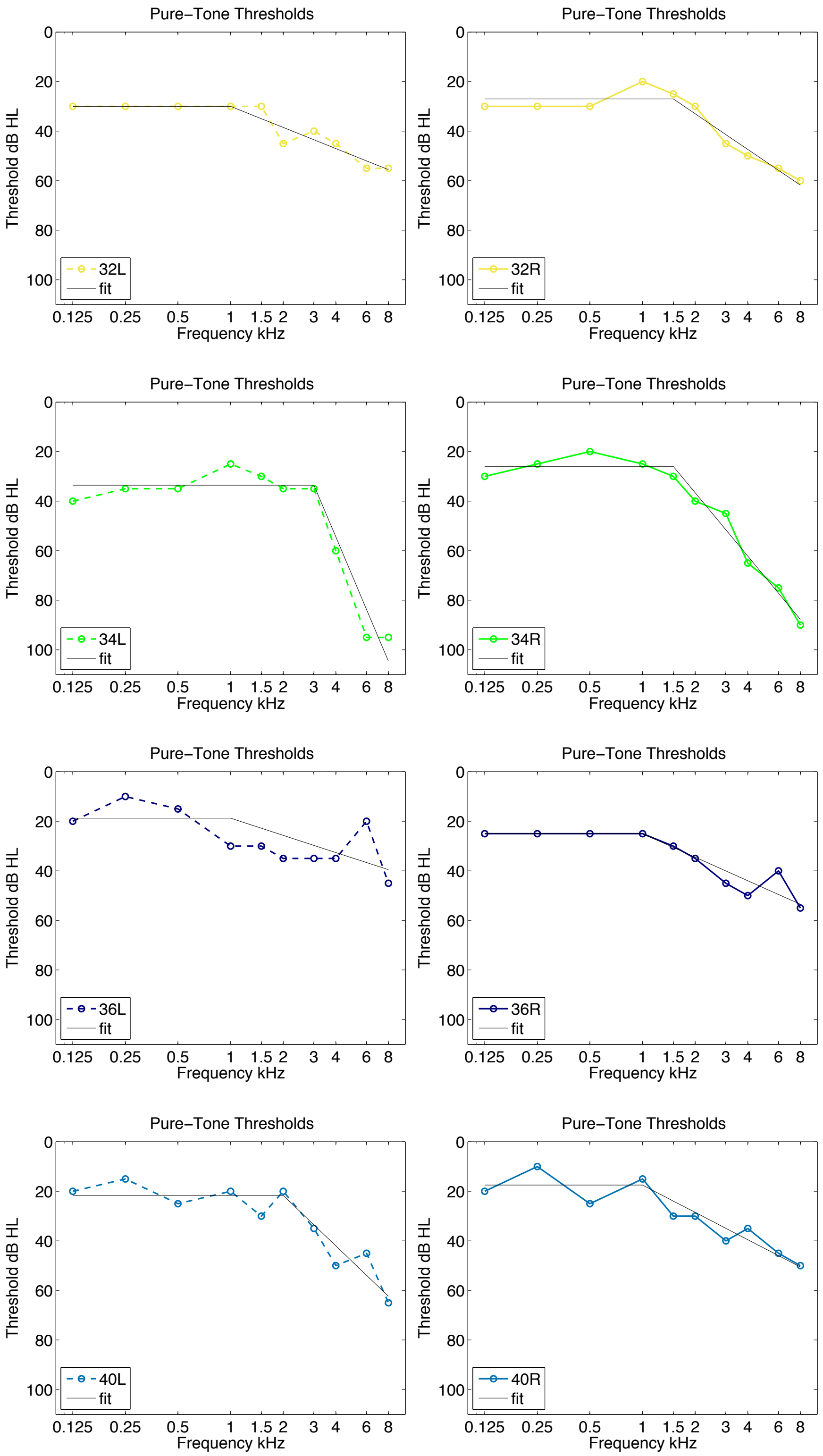


Figure A.2: Pure-tone thresholds and the three-parameter linear fit for HI listeners 32, 34, 36, and 40.

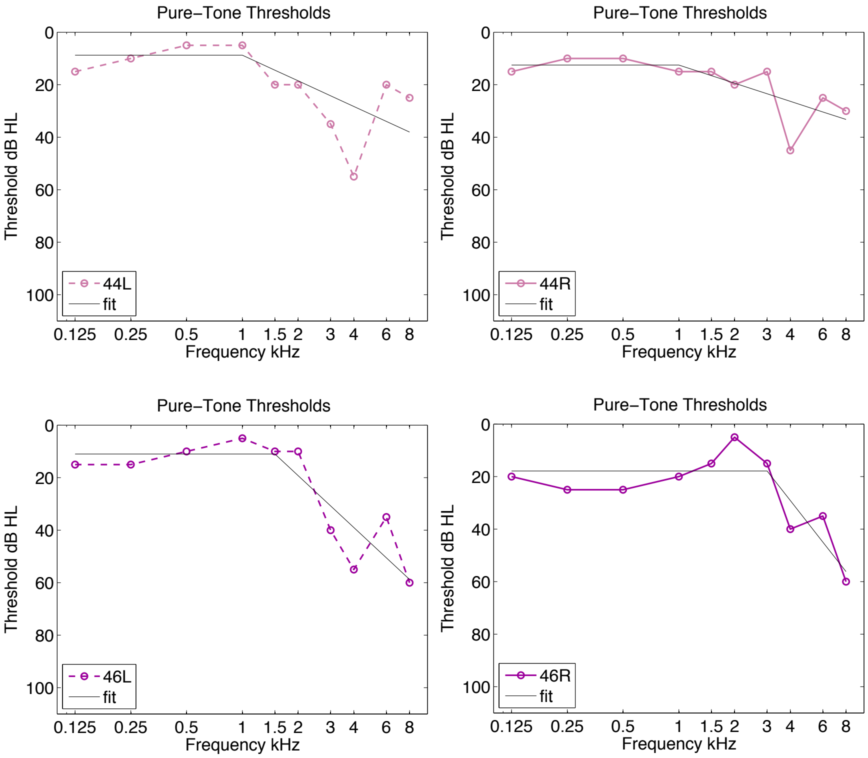


Figure A.3: Pure-tone thresholds and the three-parameter linear fit for HI listeners 44 and 46.

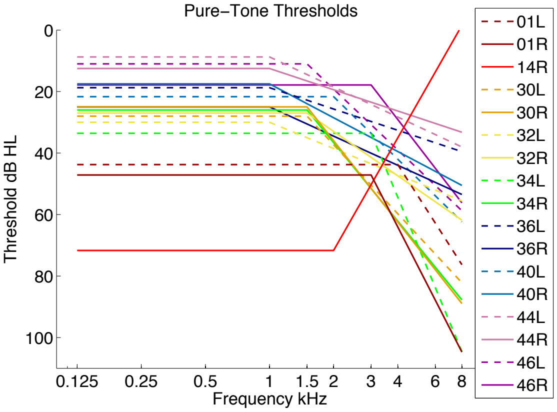
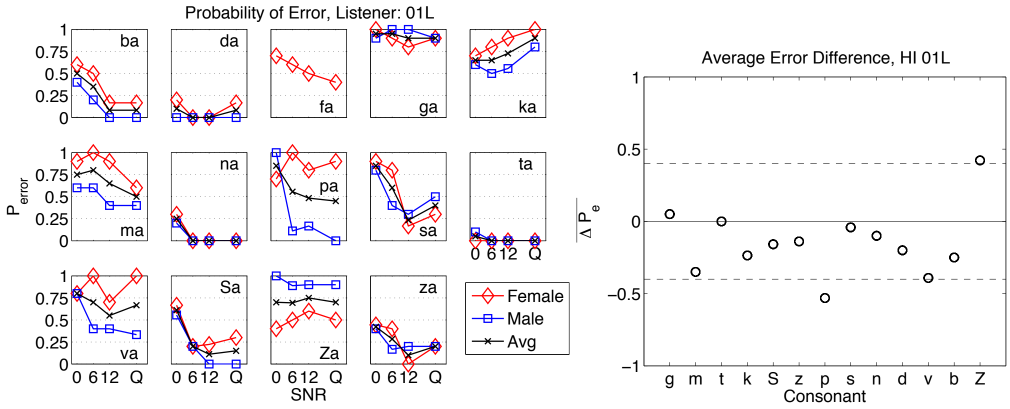


Figure A.4: The three-parameter linear fits for the 17 HI ears.

APPENDIX B

AVERAGE ERROR DIFFERENCES: HI EXP 2

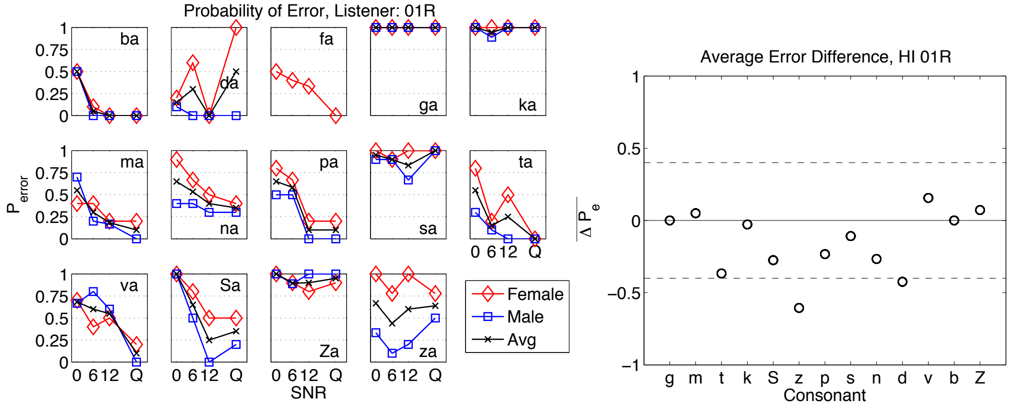
The figures in this appendix (Figs. B.1–B.17) show the raw individual token errors and the computed $\overline{\Delta P_e}$ for each consonant, with a figure for each HI ear. Subplot (a) of each figure shows the consonant recognition error as a function of SNR for both talker tokens ($P_e^M(s)$ and $P_e^F(s)$) along with the average across the two talkers is displayed in 14 sub-plots (one for each consonant). Subplot (b) the $\overline{\Delta P_e}$ for each consonant; consonants are ordered along the abscissa based on the NH ΔSNR$_{90}$ values in SWN (as in Fig. 4.3).



(a) Probability of Error

(b) Average Error Difference

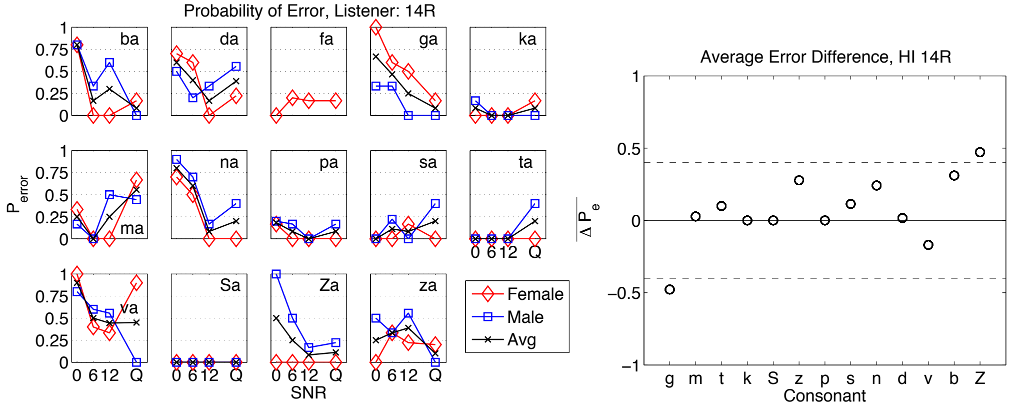
Figure B.1: Data for HI Listener 01L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR$_{90}$ values is not significant for ear 01L; the female token shows higher errors, in general.



(a) Probability of Error

(b) Average Error Difference

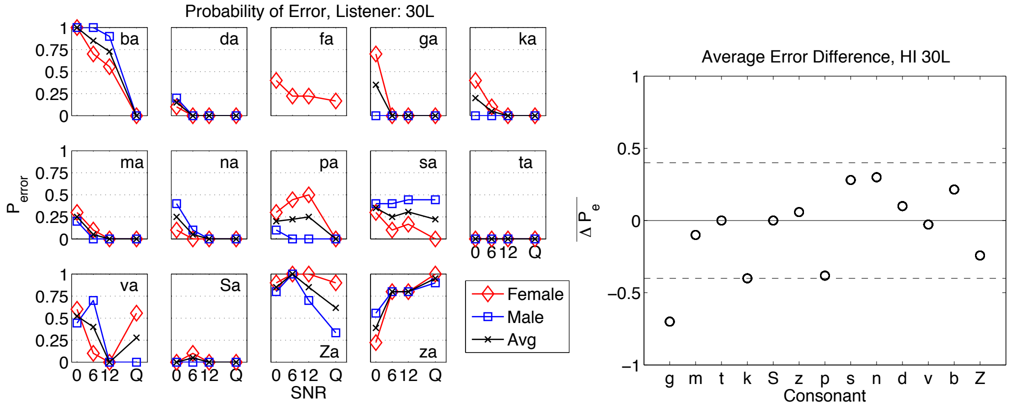
Figure B.2: Data for HI Listener 01R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is not significant for ear 01R; the female token shows higher errors, in general.



(a) Probability of Error

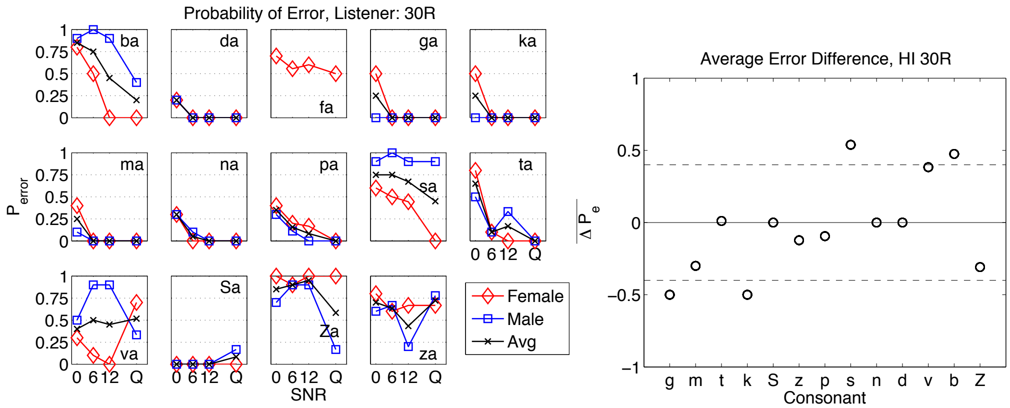
(b) Average Error Difference

Figure B.3: Data for HI Listener 14R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 14R ($\rho = 0.65$, p-val $= 0.017$).



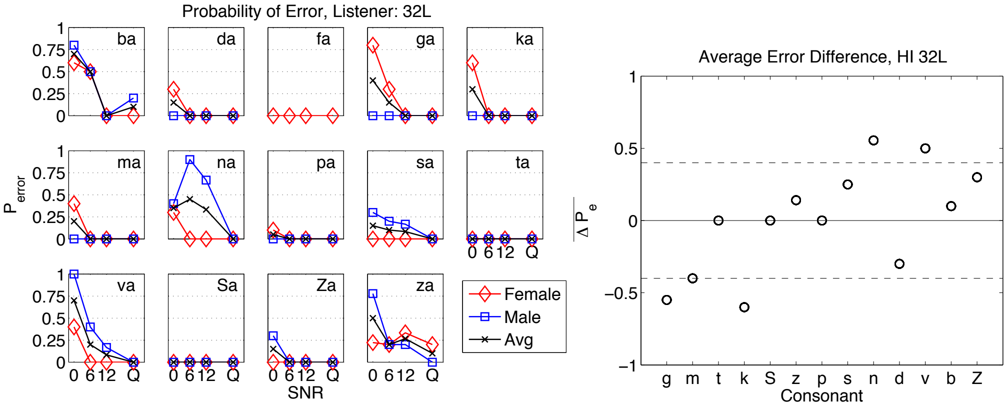
(a) Probability of Error (b) Average Error Difference

Figure B.4: Data for HI Listener 30L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH $\Delta \mathrm{SNR}_{90}$ values is not significant for ear 30L.



(a) Probability of Error (b) Average Error Difference

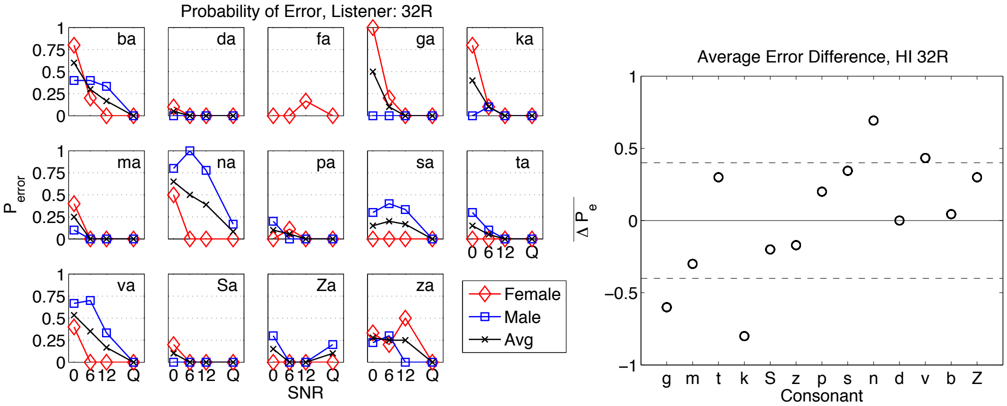
Figure B.5: Data for HI Listener 30R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH $\Delta \mathrm{SNR}_{90}$ values is not significant for ear 30R.



(a) Probability of Error

(b) Average Error Difference

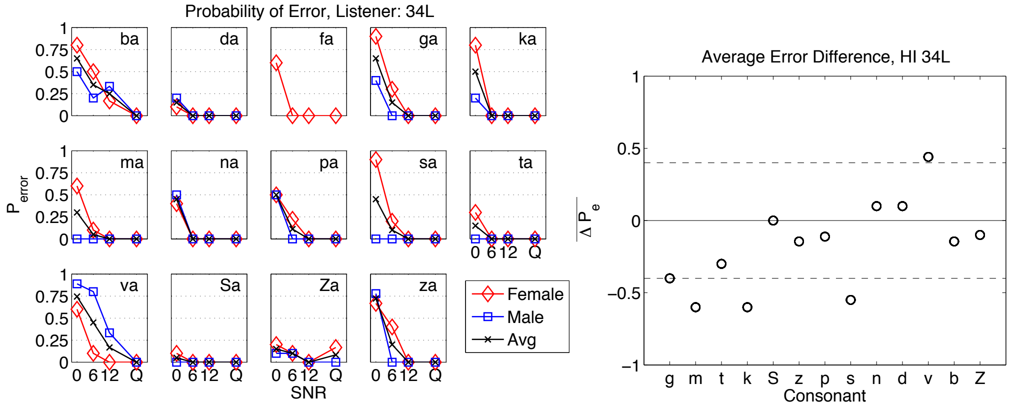
Figure B.6: Data for HI Listener 32L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 32L ($\rho = 0.65$, p-val $= 0.016$).



(a) Probability of Error

(b) Average Error Difference

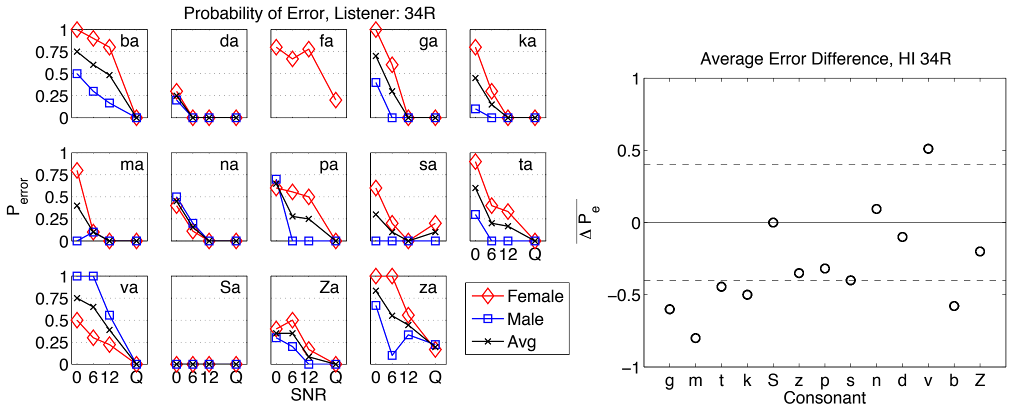
Figure B.7: Data for HI Listener 32R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 32R ($\rho = 0.62$, p-val $= 0.025$).



(a) Probability of Error

(b) Average Error Difference

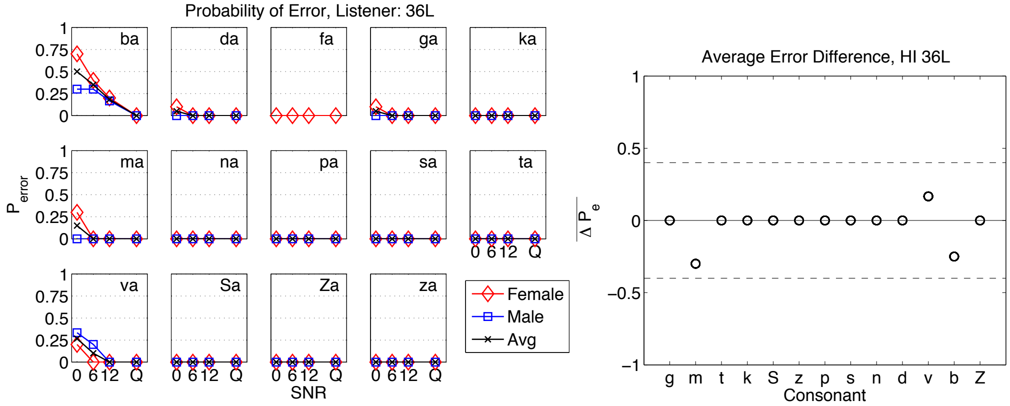
Figure B.8: Data for HI Listener 34L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 34L ($\rho = 0.57$, p-val $= 0.04$).



(a) Probability of Error

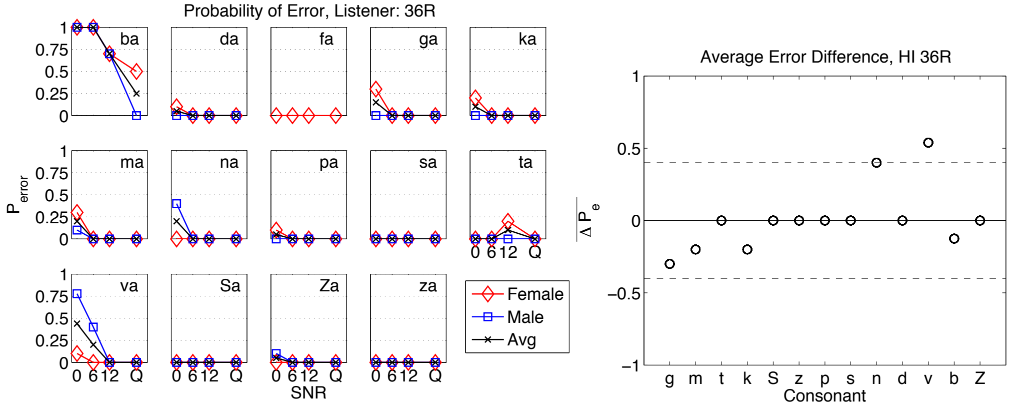
(b) Average Error Difference

Figure B.9: Data for HI Listener 34R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is not significant for ear 34R.



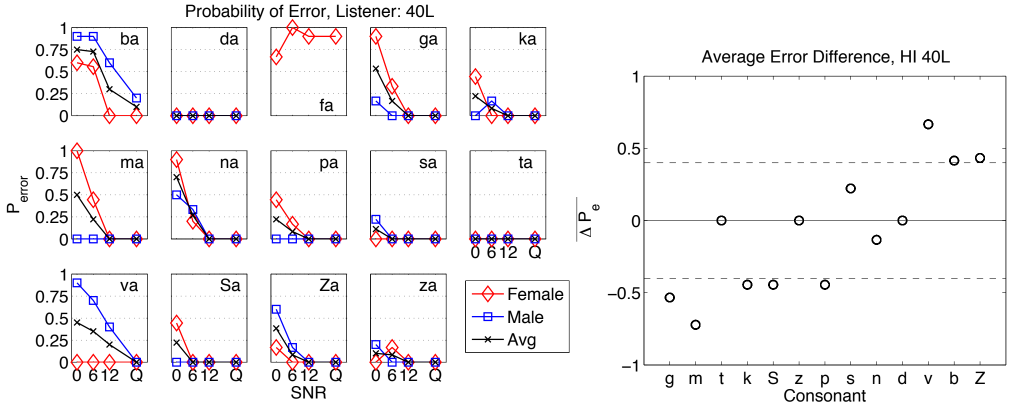
(a) Probability of Error (b) Average Error Difference

Figure B.10: Data for HI Listener 36L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. No comparison is made to the NH ΔSNR_{90} values due to the lack of significantly nonzero $\overline{\Delta P_e}$ values.



(a) Probability of Error (b) Average Error Difference

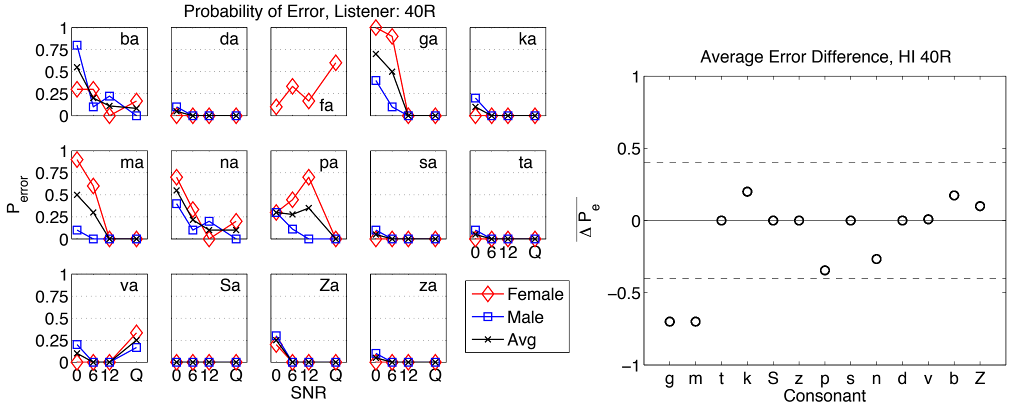
Figure B.11: Data for HI Listener 36R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The only $\overline{\Delta P_e}$ values with absolute values > 0.4 are for /n, v/, both of which agree with the NH ΔSNR_{90} values that show the male token as having a weaker cue.



(a) Probability of Error

(b) Average Error Difference

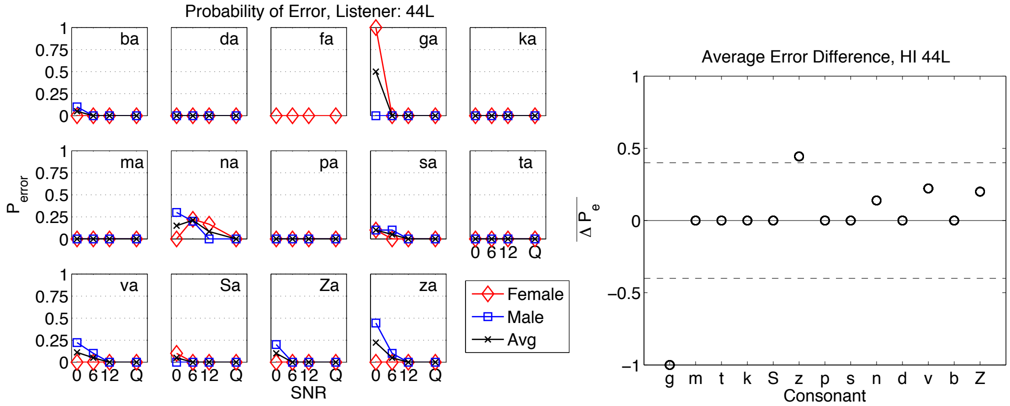
Figure B.12: Data for HI Listener 40L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH $\Delta\mathrm{SNR}_{90}$ values is significant for ear 40L ($\rho = 0.78$, p-val $= 0.001$).



(a) Probability of Error

(b) Average Error Difference

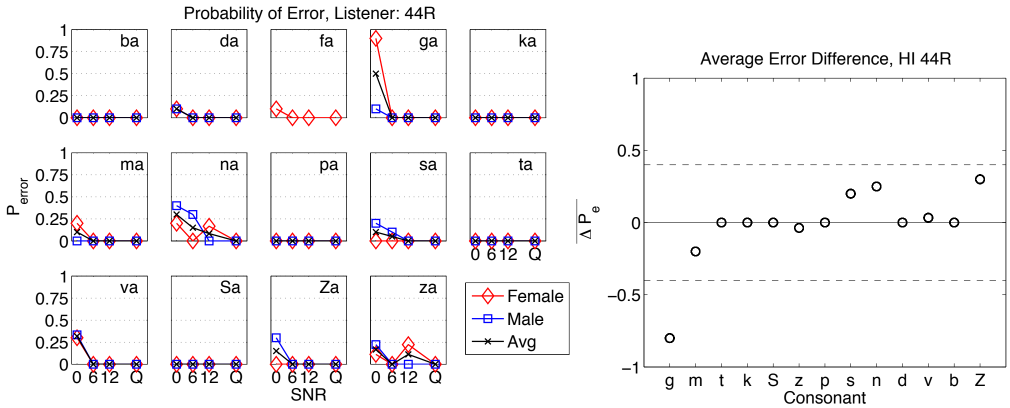
Figure B.13: Data for HI Listener 40R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH $\Delta\mathrm{SNR}_{90}$ values is significant for ear 40R ($\rho = 0.59$, p-val $= 0.032$).



(a) Probability of Error

(b) Average Error Difference

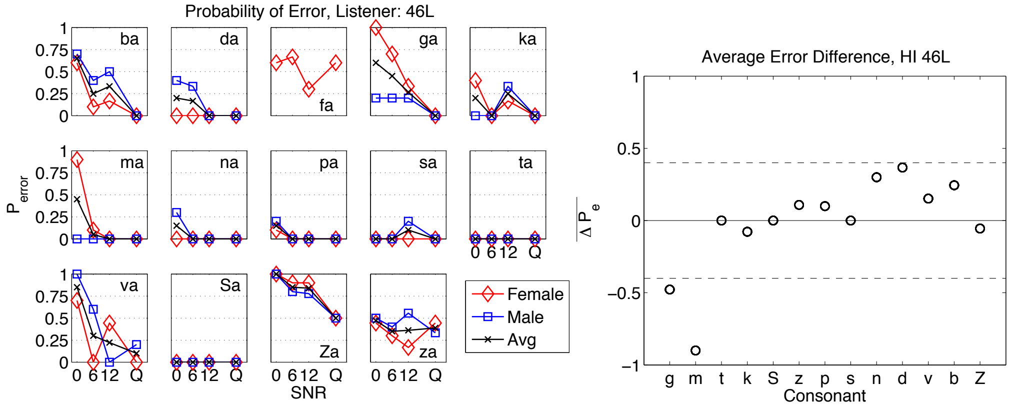
Figure B.14: Data for HI Listener 44L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 44L ($\rho = 0.64$, p-val $= 0.019$).



(a) Probability of Error

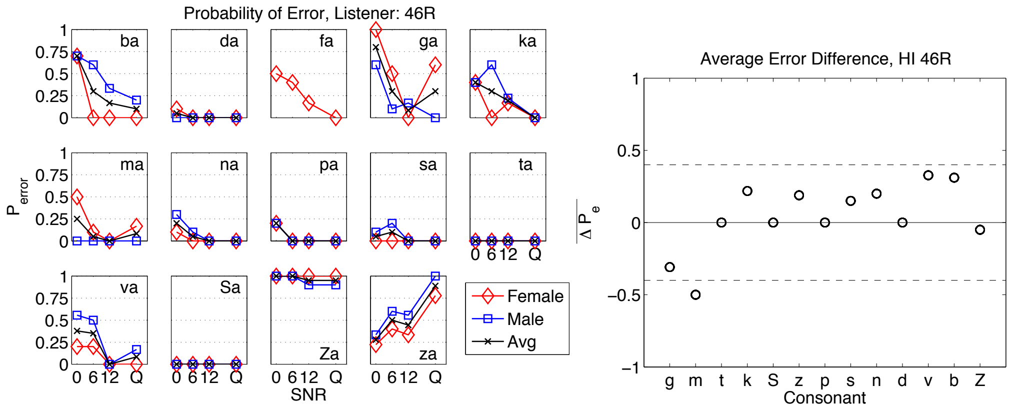
(b) Average Error Difference

Figure B.15: Data for HI Listener 44R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 44R ($\rho = 0.74$, p-val $= 0.003$).



(a) Probability of Error (b) Average Error Difference

Figure B.16: Data for HI Listener 46L. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 46R ($\rho = 0.63$, p-val $= 0.02$).



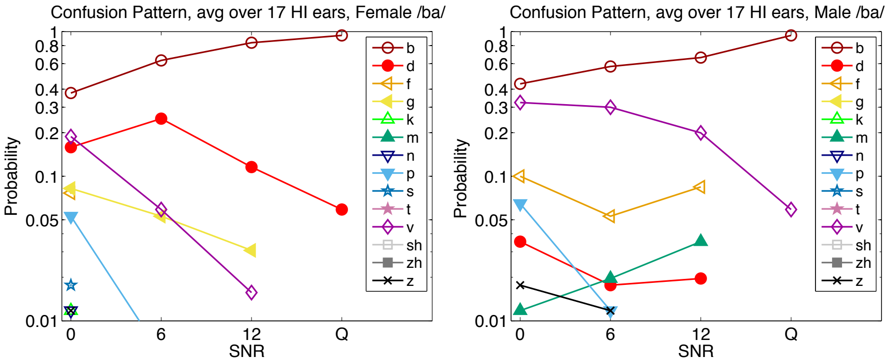
(a) Probability of Error (b) Average Error Difference

Figure B.17: Data for HI Listener 46R. (a) Consonant recognition error as a function of SNR; plots display the error for the female token (red diamond), male token (blue square), and the average across the two talkers (black x). (b) $\overline{\Delta P_e}$ for each consonant. The correlation between the HI $\overline{\Delta P_e}$ and NH ΔSNR_{90} values is significant for ear 46R ($\rho = 0.55$, p-val $= 0.049$).

APPENDIX C

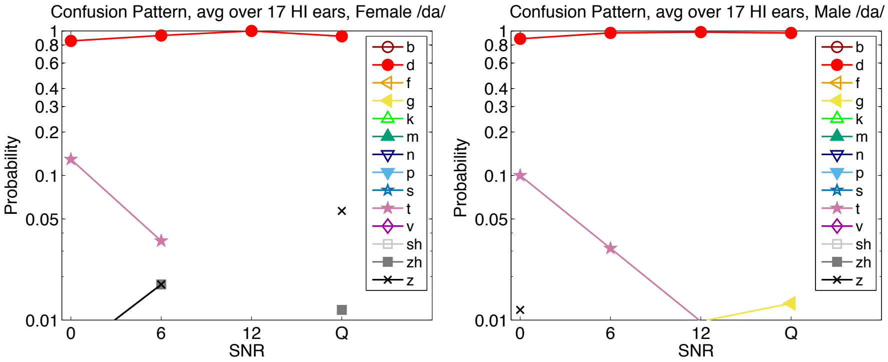
HEARING-IMPAIRED CONFUSION PATTERNS: HI EXP 2

The confusion patterns for each token that was used in HI Experiment 2 (Han 2011) are shown in Figs. C.1–C.14. These confusion patterns were computed by averaging the data for 17 HI ears, and, therefore, have significant underlying variance.



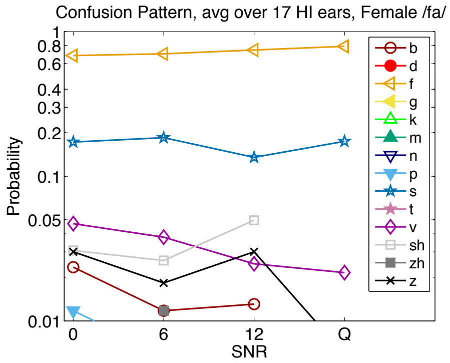
(a) Female Talker (b) Male Talker

Figure C.1: Confusion patterns for /ba/.



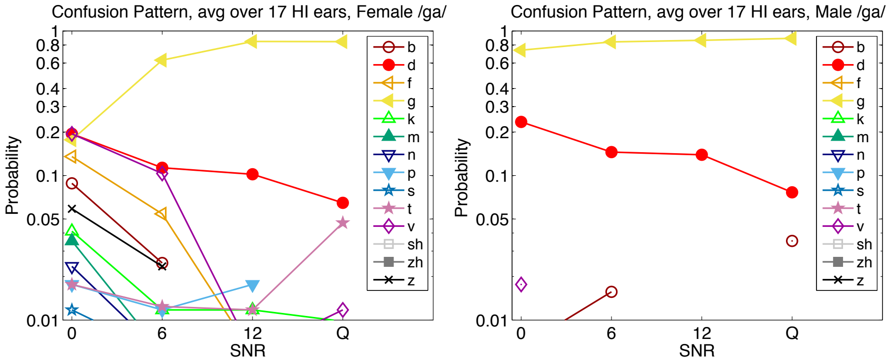
(a) Female Talker (b) Male Talker

Figure C.2: Confusion patterns for /da/.



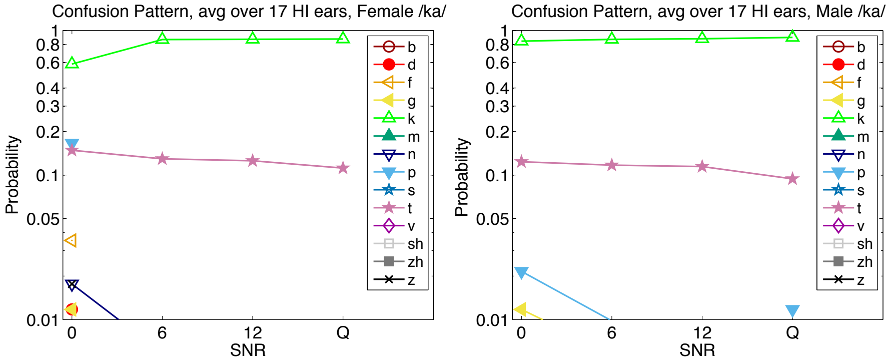
(a) Female Talker

Figure C.3: Confusion patterns for /fa/.



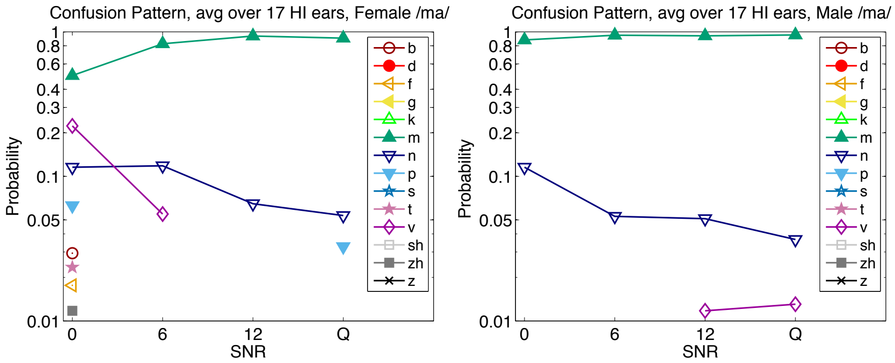
(a) Female Talker (b) Male Talker

Figure C.4: Confusion patterns for /ga/.



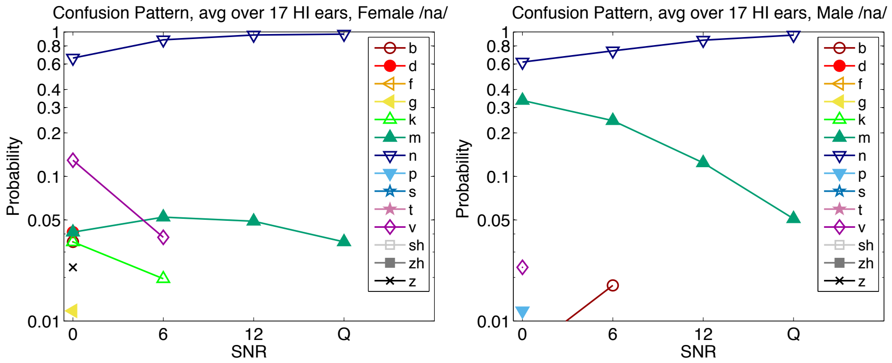
(a) Female Talker (b) Male Talker

Figure C.5: Confusion patterns for /ka/.



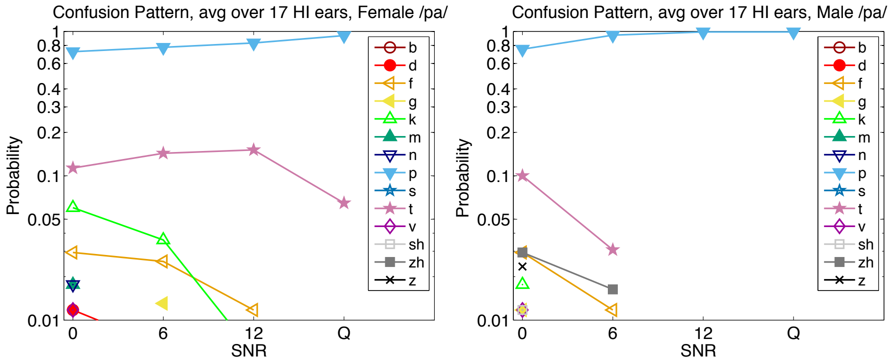
(a) Female Talker (b) Male Talker

Figure C.6: Confusion patterns for /ma/.



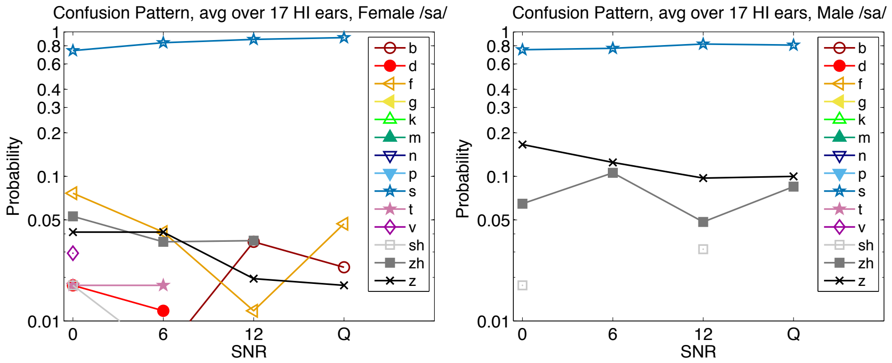
(a) Female Talker (b) Male Talker

Figure C.7: Confusion patterns for /na/.



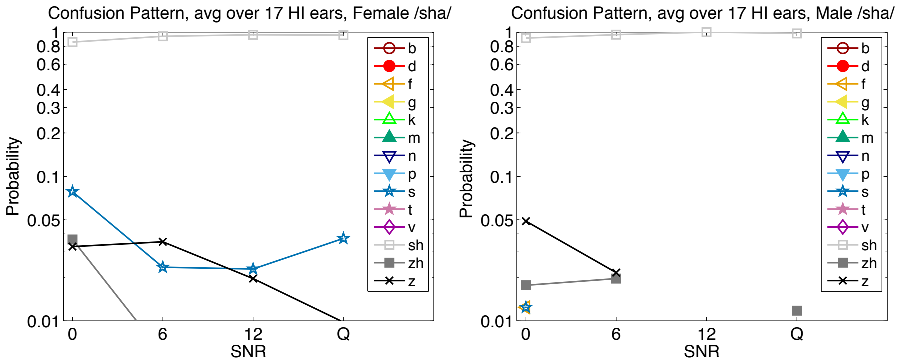
(a) Female Talker (b) Male Talker

Figure C.8: Confusion patterns for /pa/.



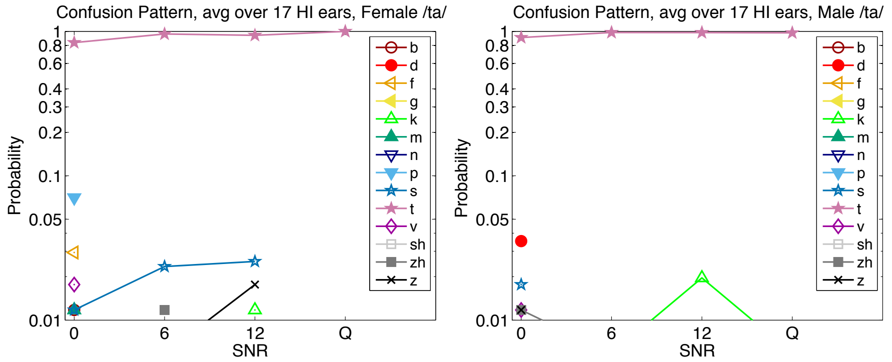
(a) Female Talker (b) Male Talker

Figure C.9: Confusion patterns for /sa/.



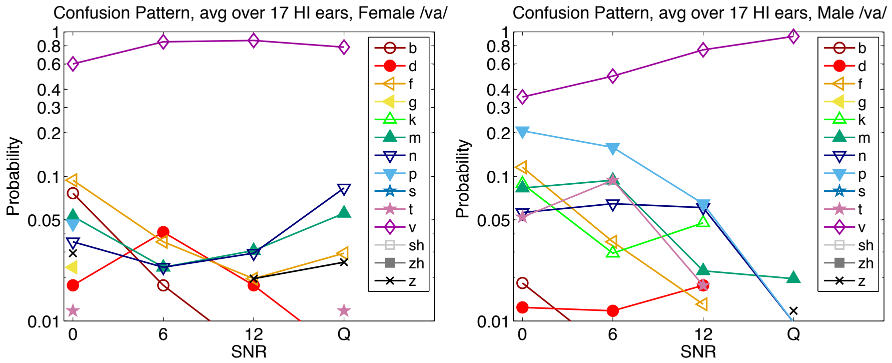
(a) Female Talker (b) Male Talker

Figure C.10: Confusion patterns for /ʃa/.



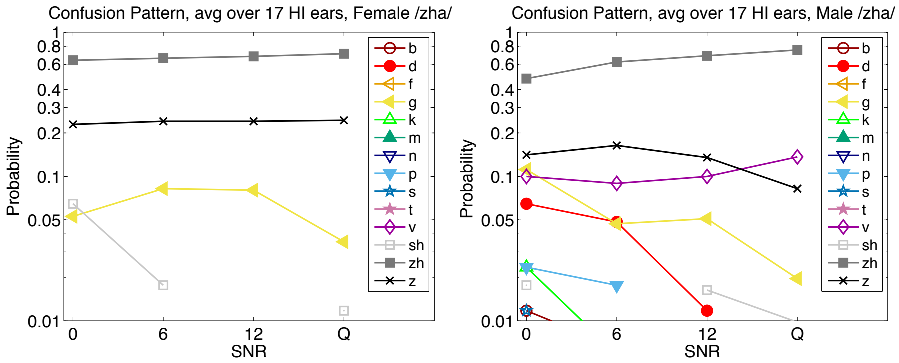
(a) Female Talker (b) Male Talker

Figure C.11: Confusion patterns for /ta/.



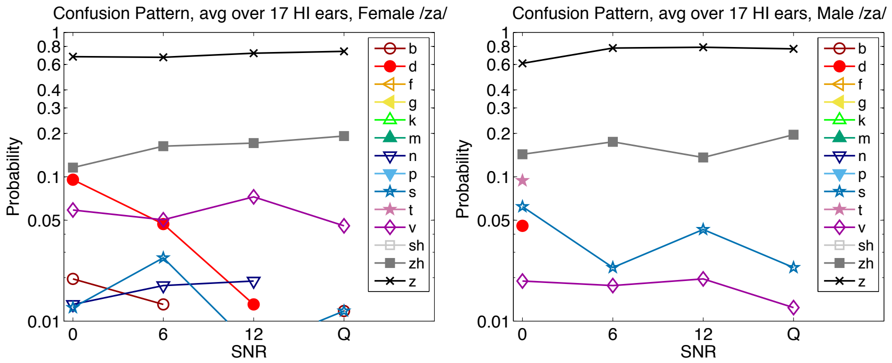
(a) Female Talker (b) Male Talker

Figure C.12: Confusion patterns for /va/.



(a) Female Talker (b) Male Talker

Figure C.13: Confusion patterns for /ʒa/.



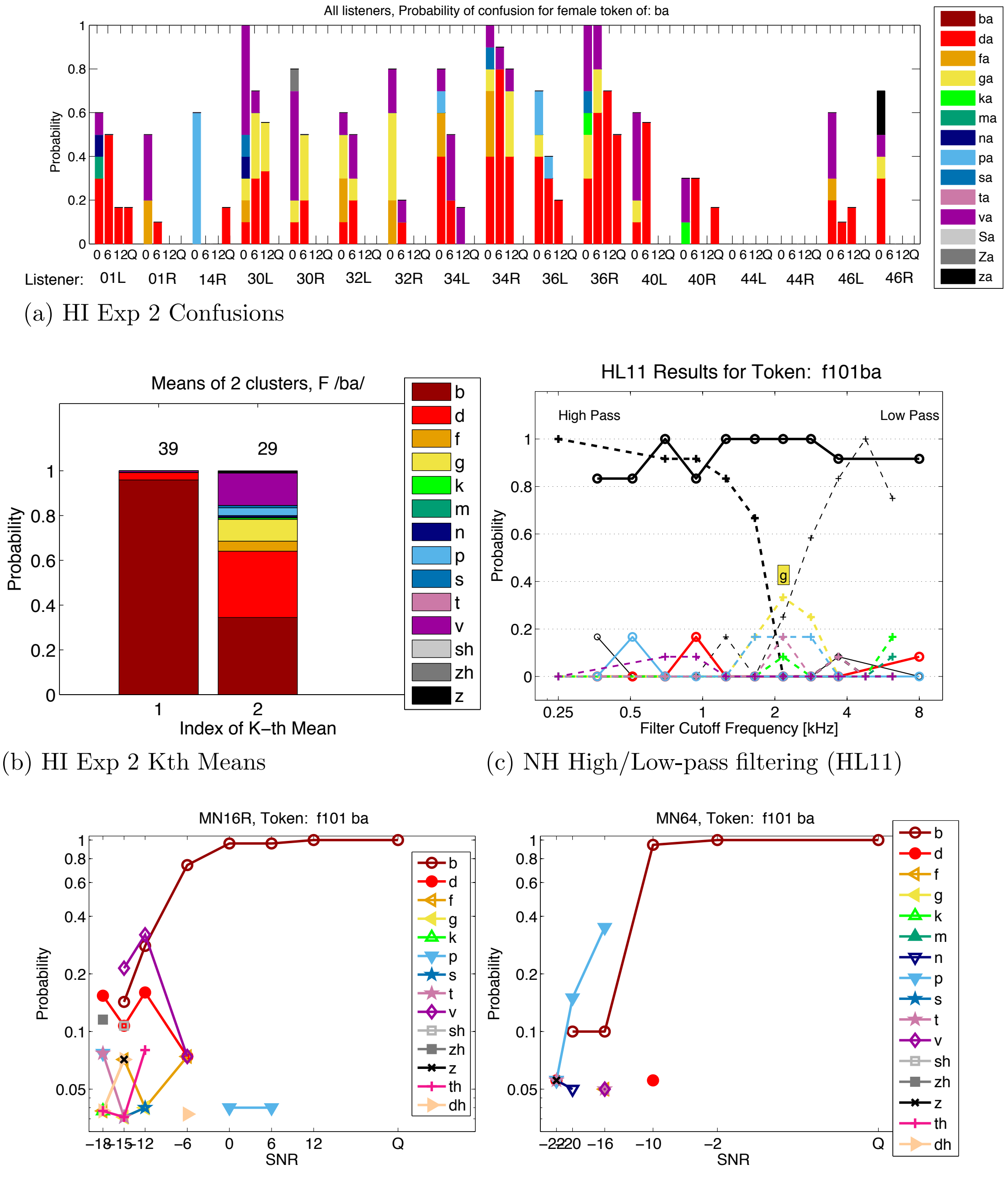
(a) Female Talker (b) Male Talker

Figure C.14: Confusion patterns for /za/.

APPENDIX D

HI CLUSTERED CONFUSIONS VS. NH NOISE-MASKING AND FILTERING DATA

The raw consonant confusion data for all tokens and HI ears is compiled in this appendix, along with the k-means cluster means, the NH noise-masking results, and the NH high/low-pass filtering results. The data is shown for each individual token in Figs. D.1–D.54.



(a) HI Exp 2 Confusions

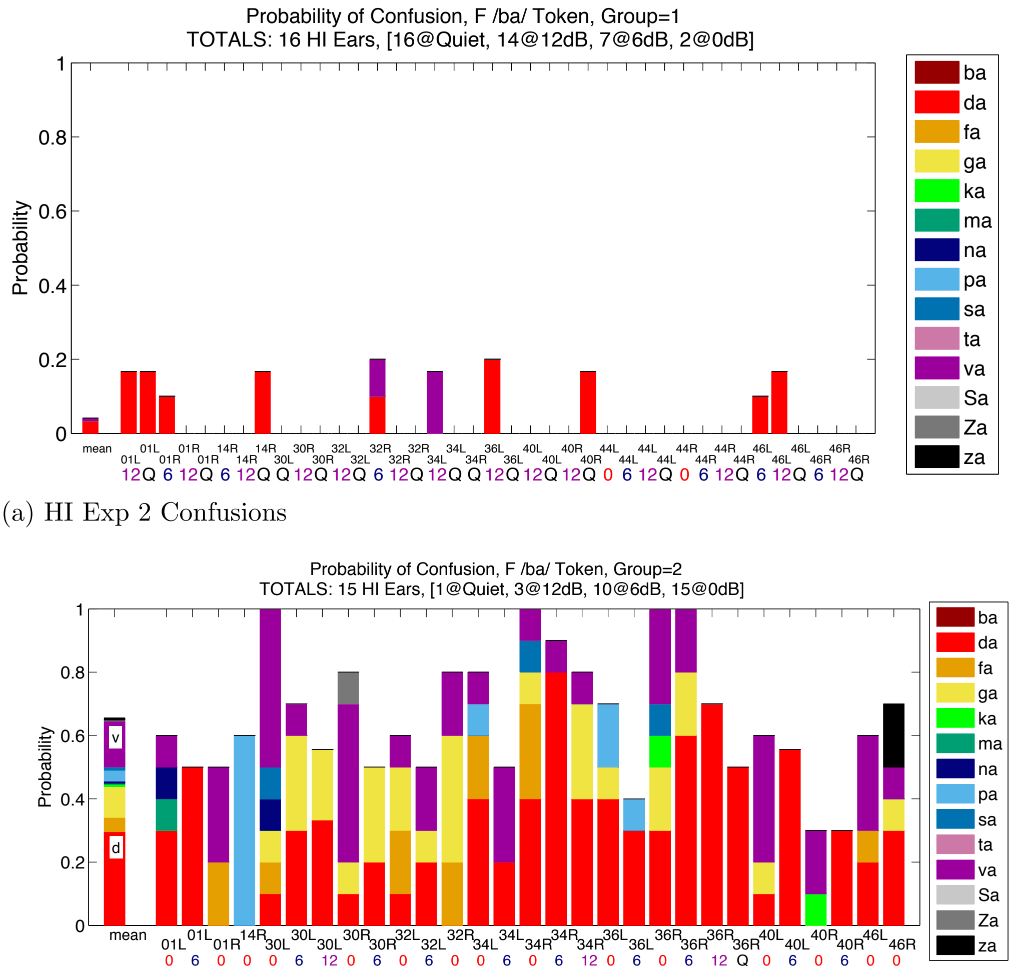
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

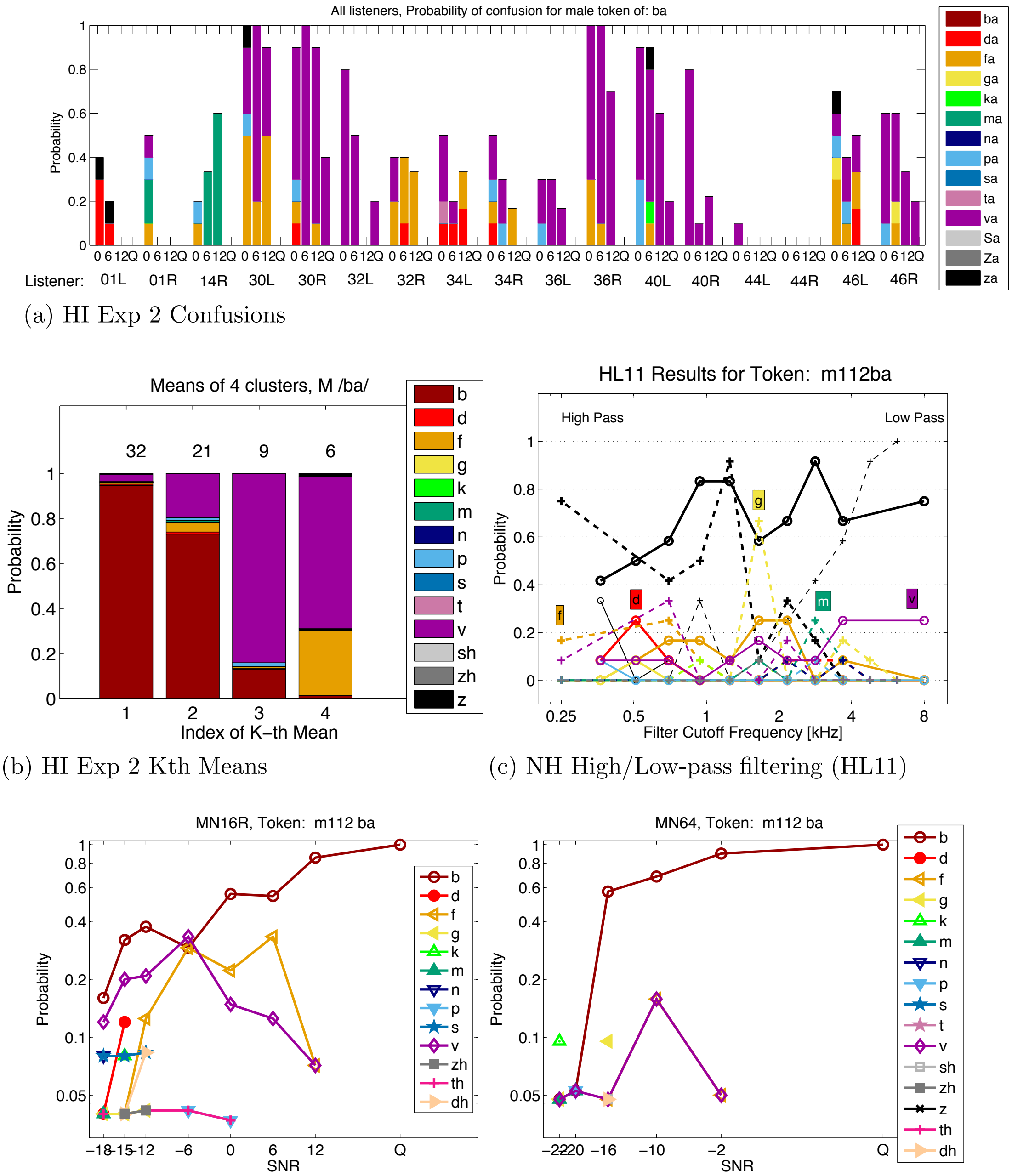
Figure D.1: Overview of data collected for the /bɑ/ token from talker f101. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.2: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f101 bɑ.



(a) HI Exp 2 Confusions

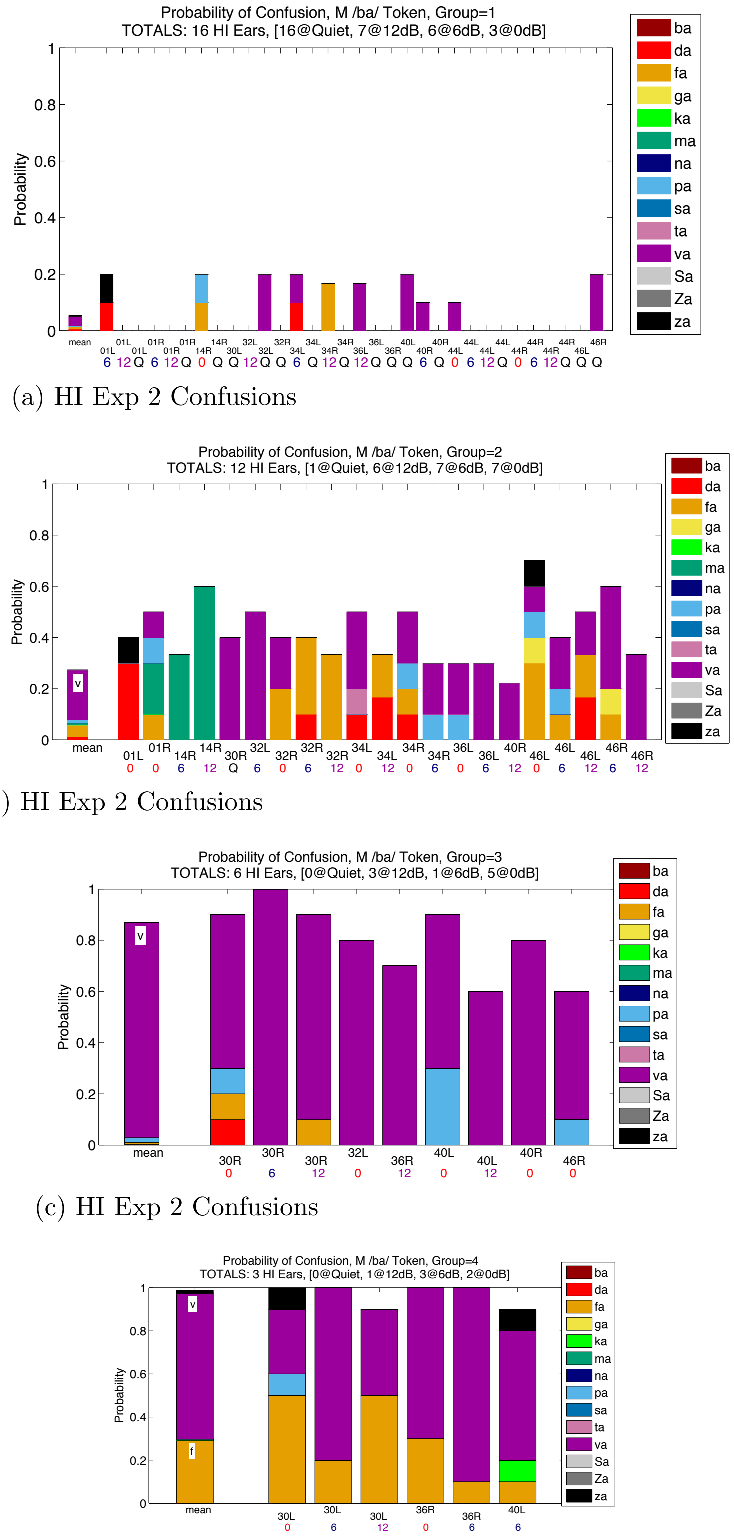
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.3: Overview of data collected for the /bɑ/ token from talker m112. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



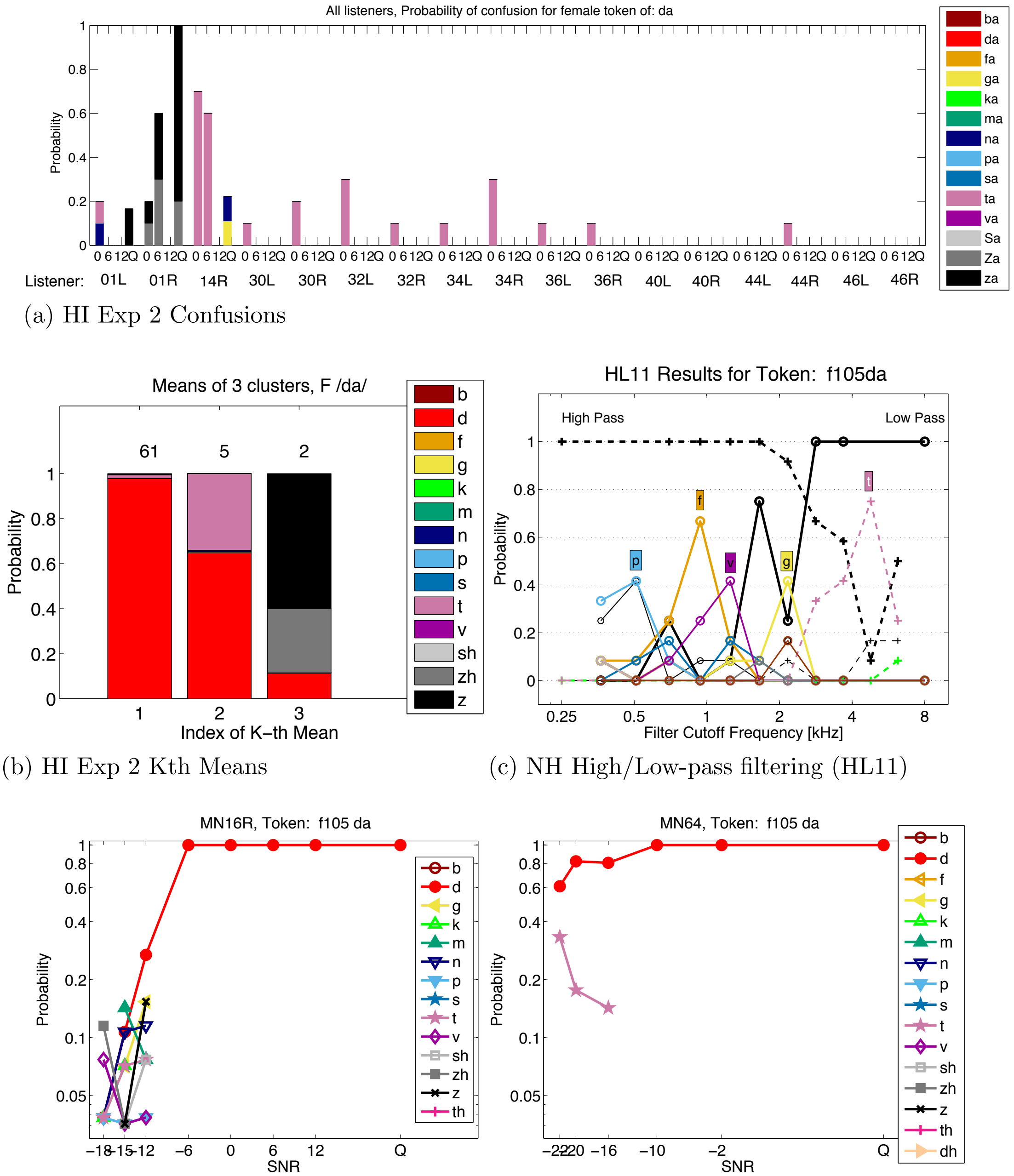
(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

Figure D.4: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m112 bɑ.



(a) HI Exp 2 Confusions

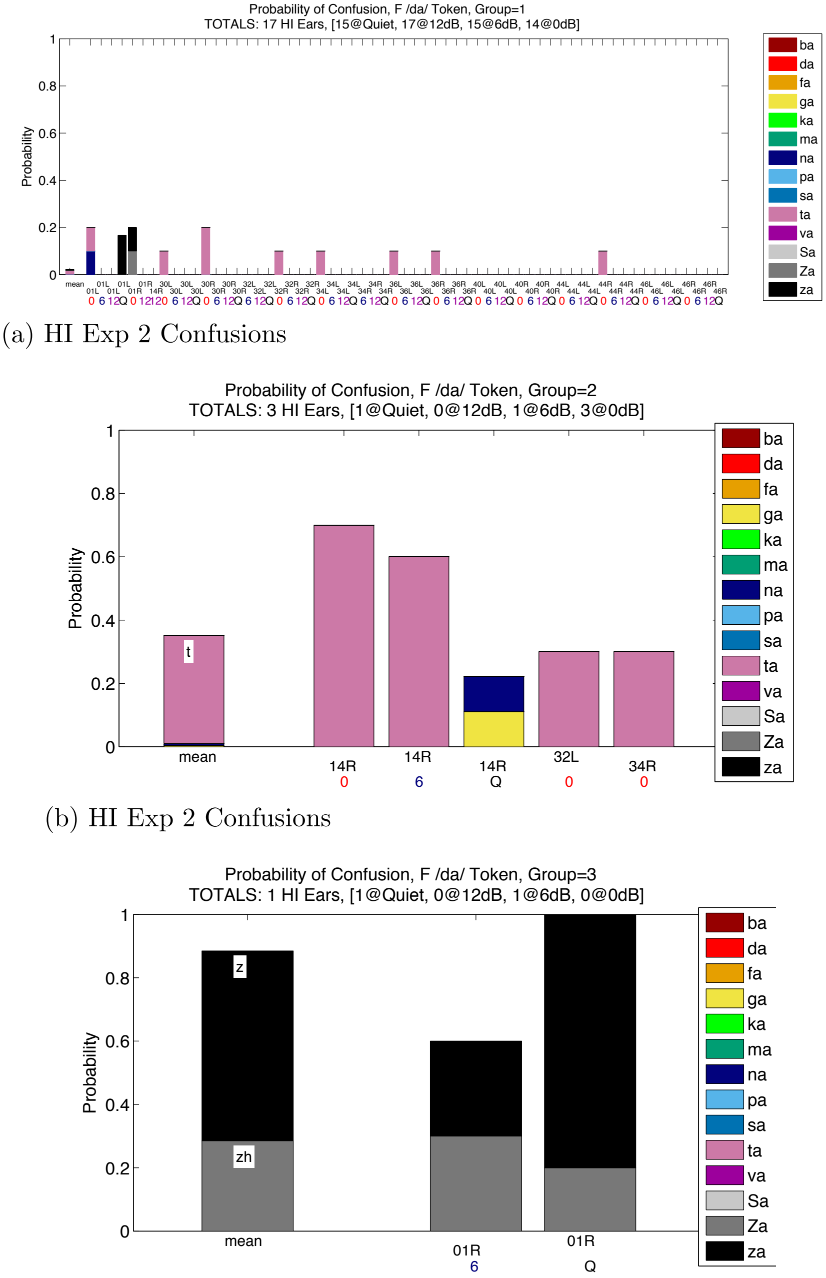
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.5: Overview of data collected for the /dɑ/ token from talker f105. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.

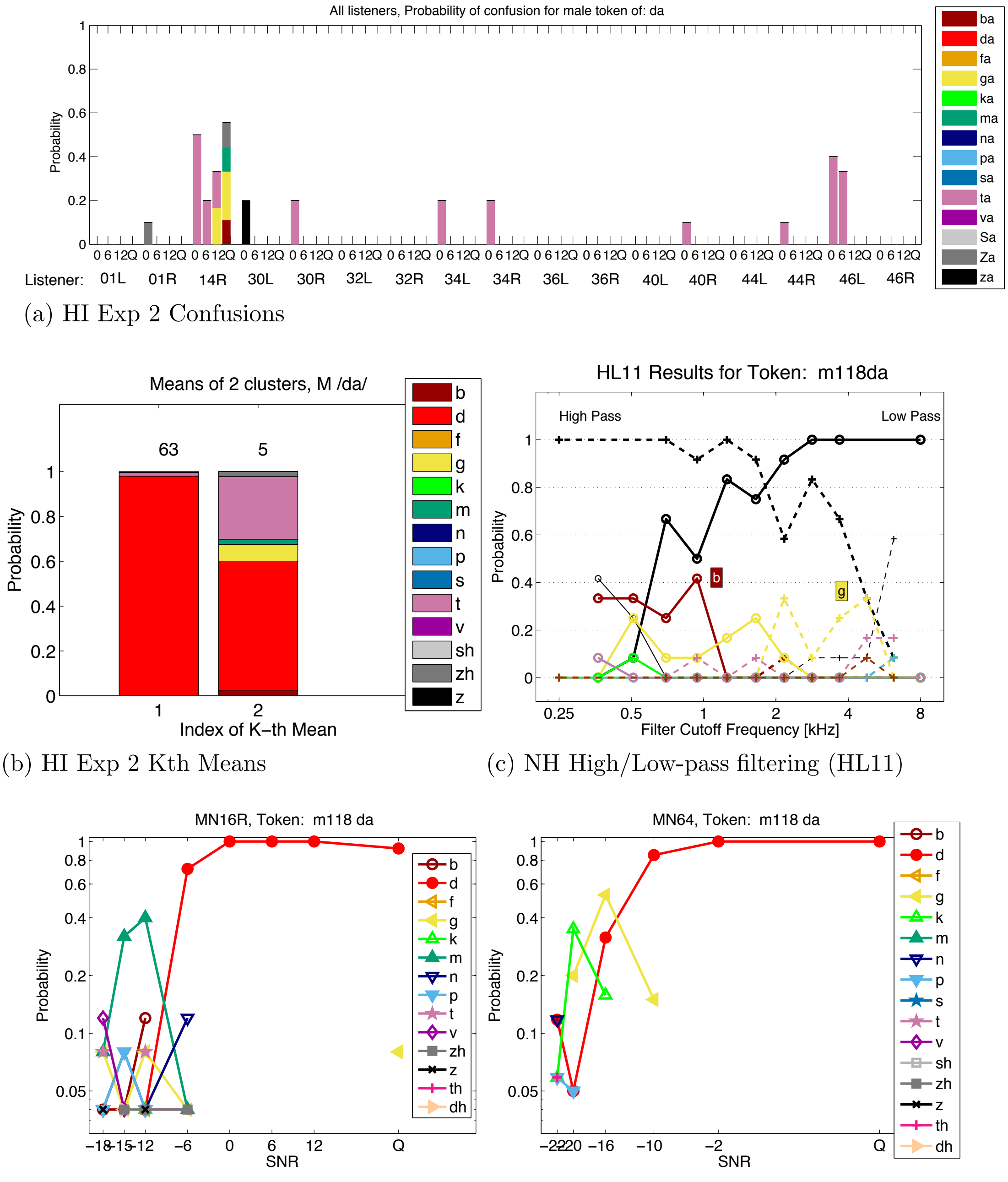


(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

Figure D.6: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f105 dɑ.



(a) HI Exp 2 Confusions

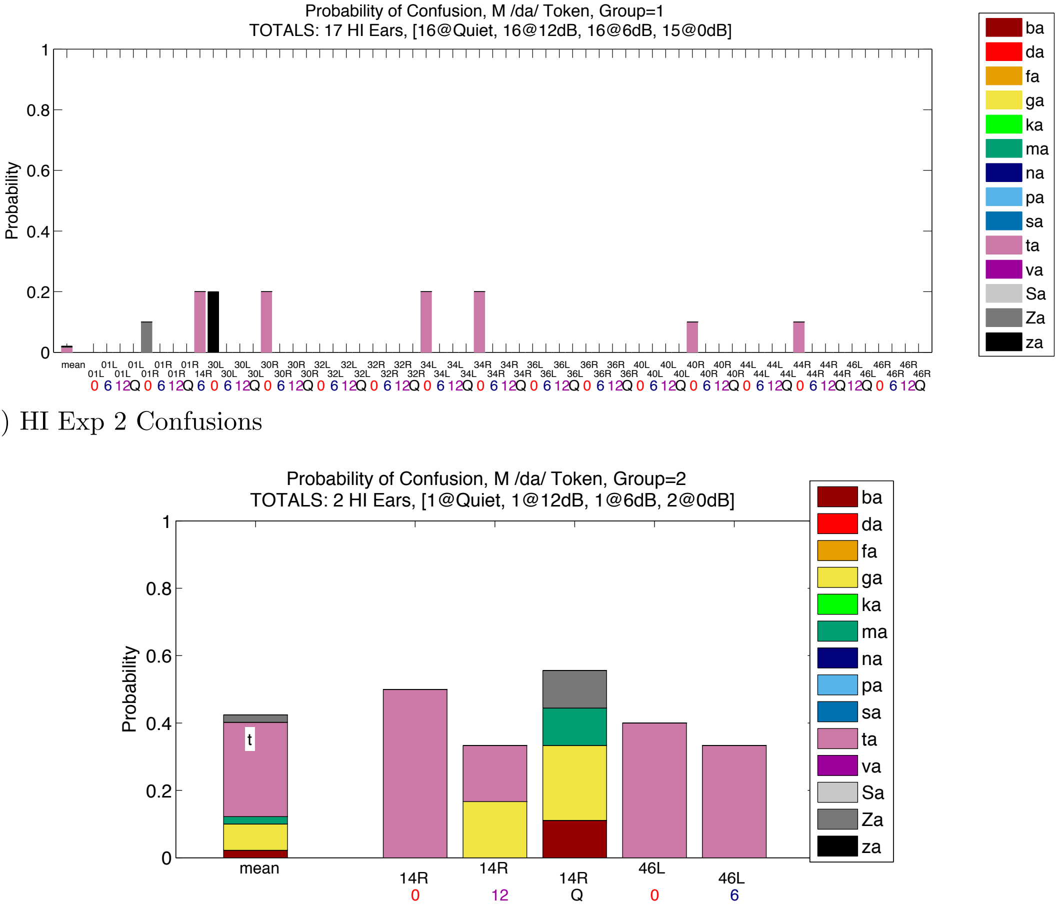
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

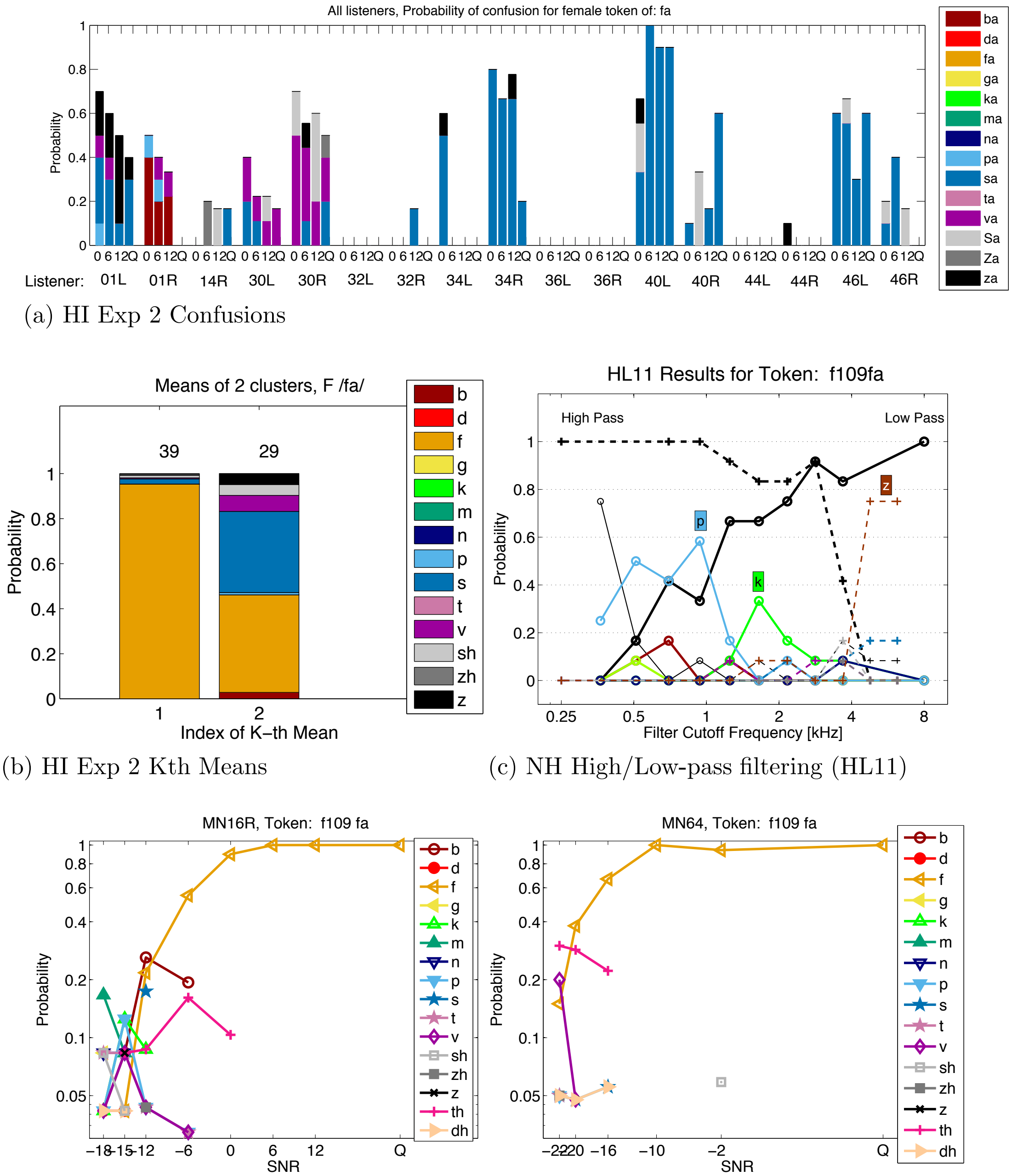
Figure D.7: Overview of data collected for the /dɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.8: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m118 dɑ.



(a) HI Exp 2 Confusions

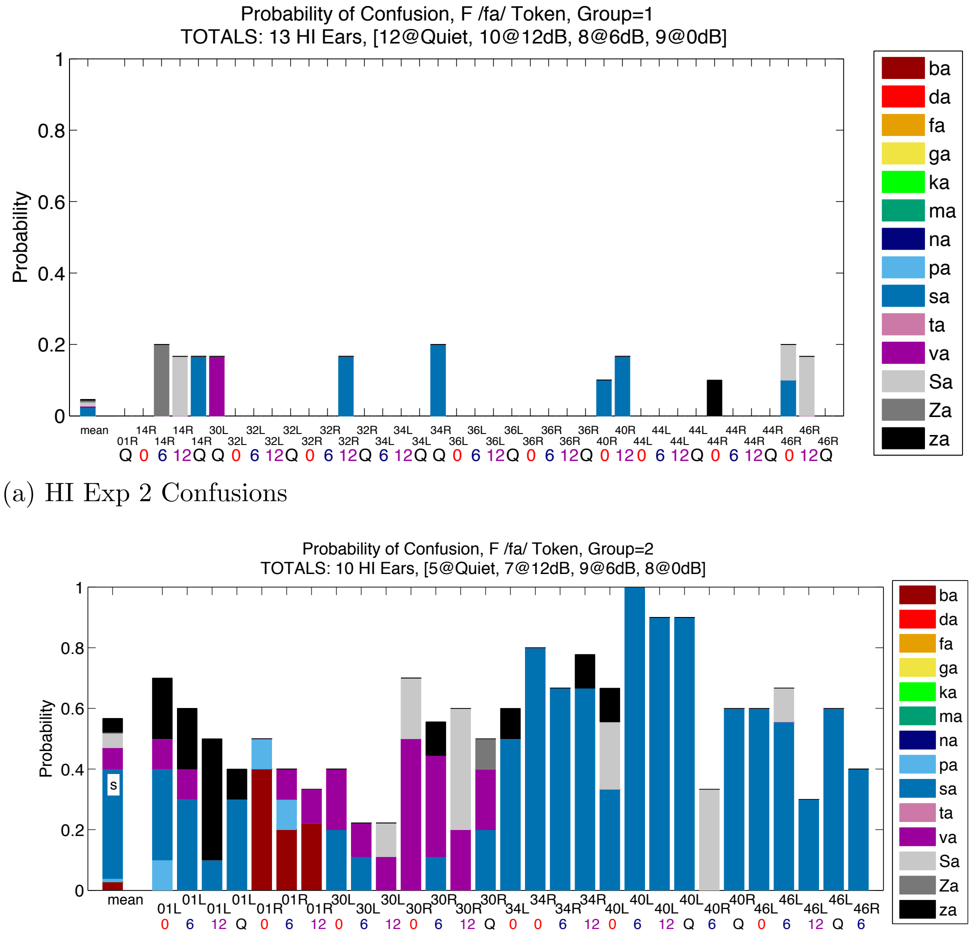
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

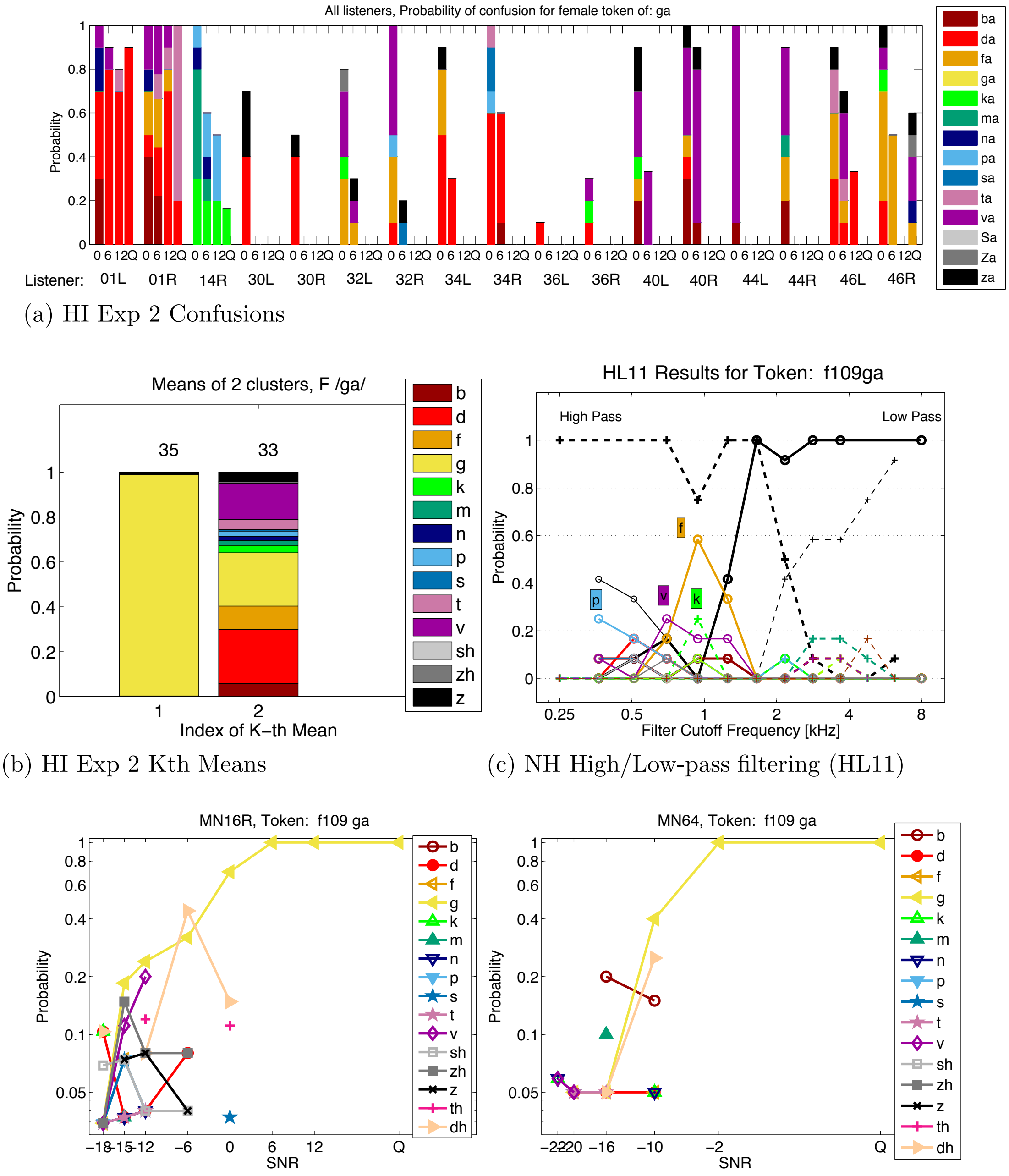
Figure D.9: Overview of data collected for the /fa/ token from talker f109. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.10: Each subplot shows the k$^{\mathrm{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f109 fa.



(a) HI Exp 2 Confusions

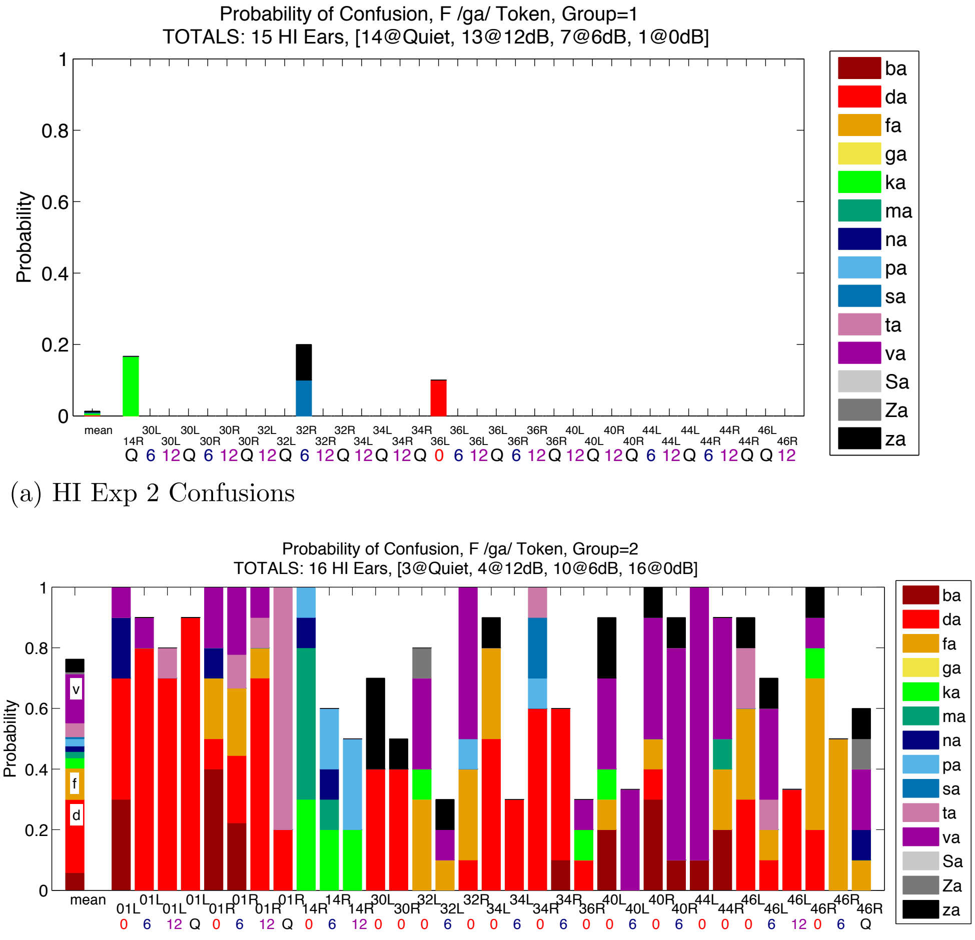
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

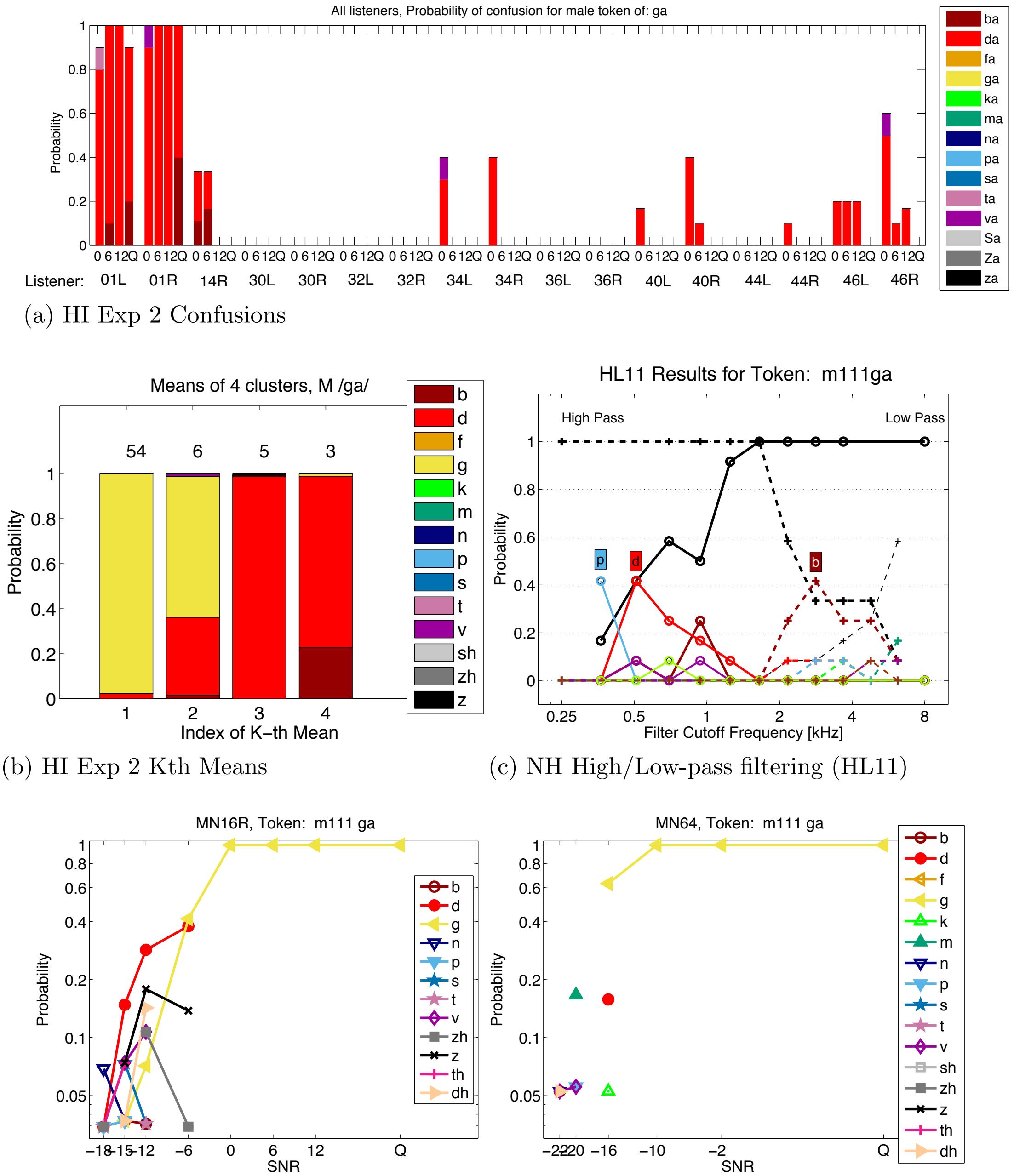
Figure D.11: Overview of data collected for the /gɑ/ token from talker f109. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.12: Each subplot shows the $\mathrm{k^{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f109 ga.



(a) HI Exp 2 Confusions

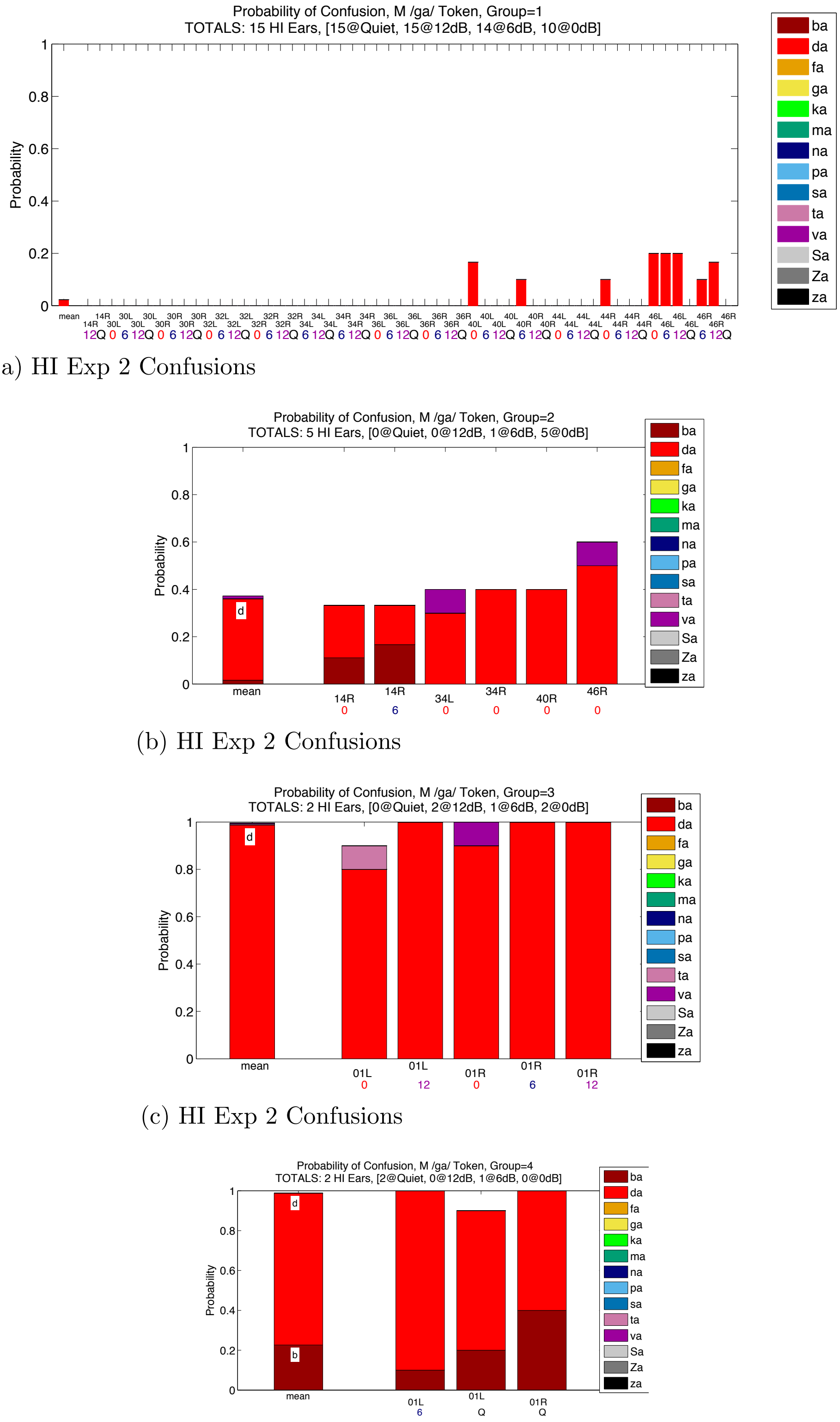
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.13: Overview of data collected for the /ga/ token from talker m111. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



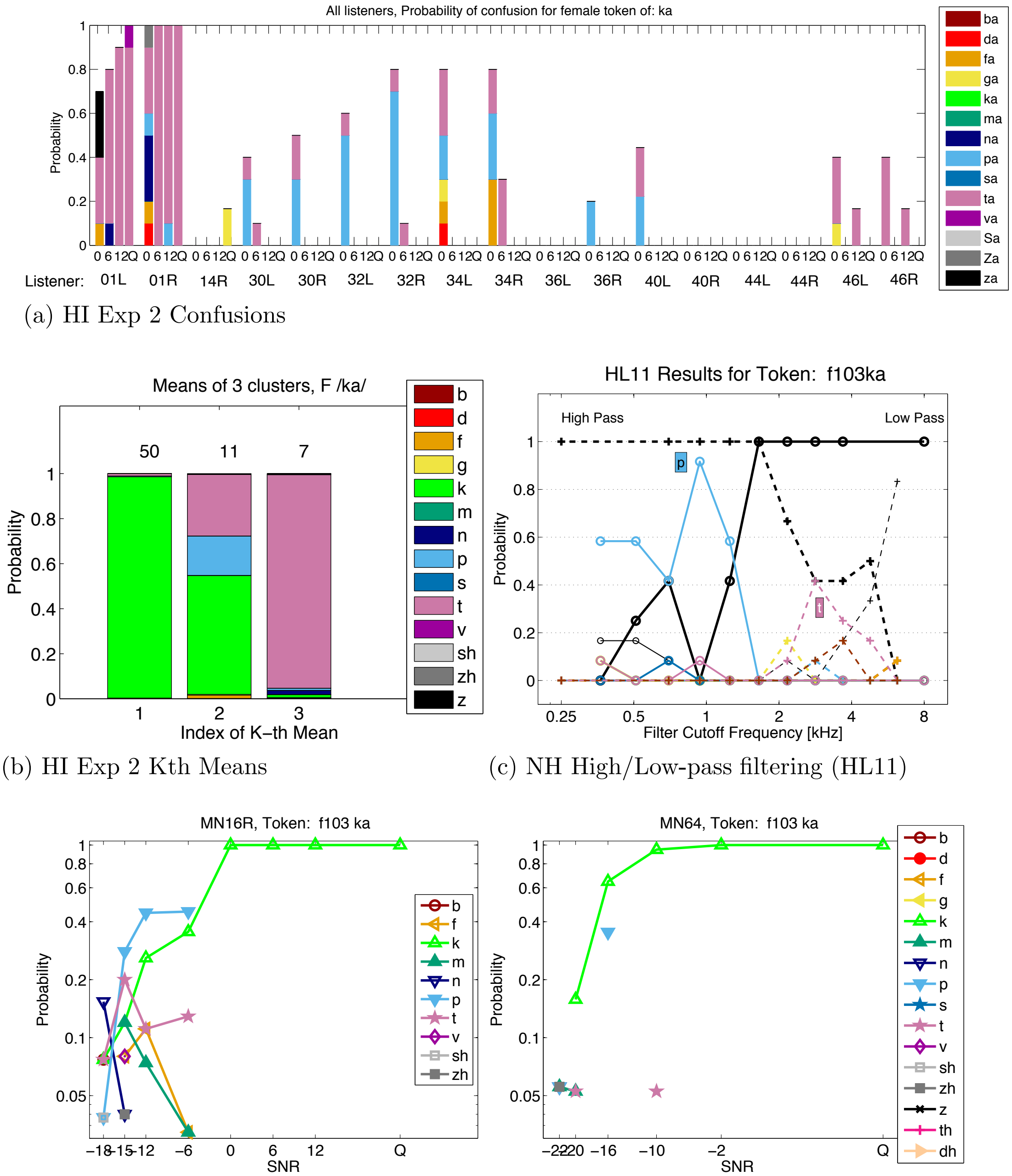
(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

Figure D.14: Each subplot shows the kth cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m111 ga.



(a) HI Exp 2 Confusions

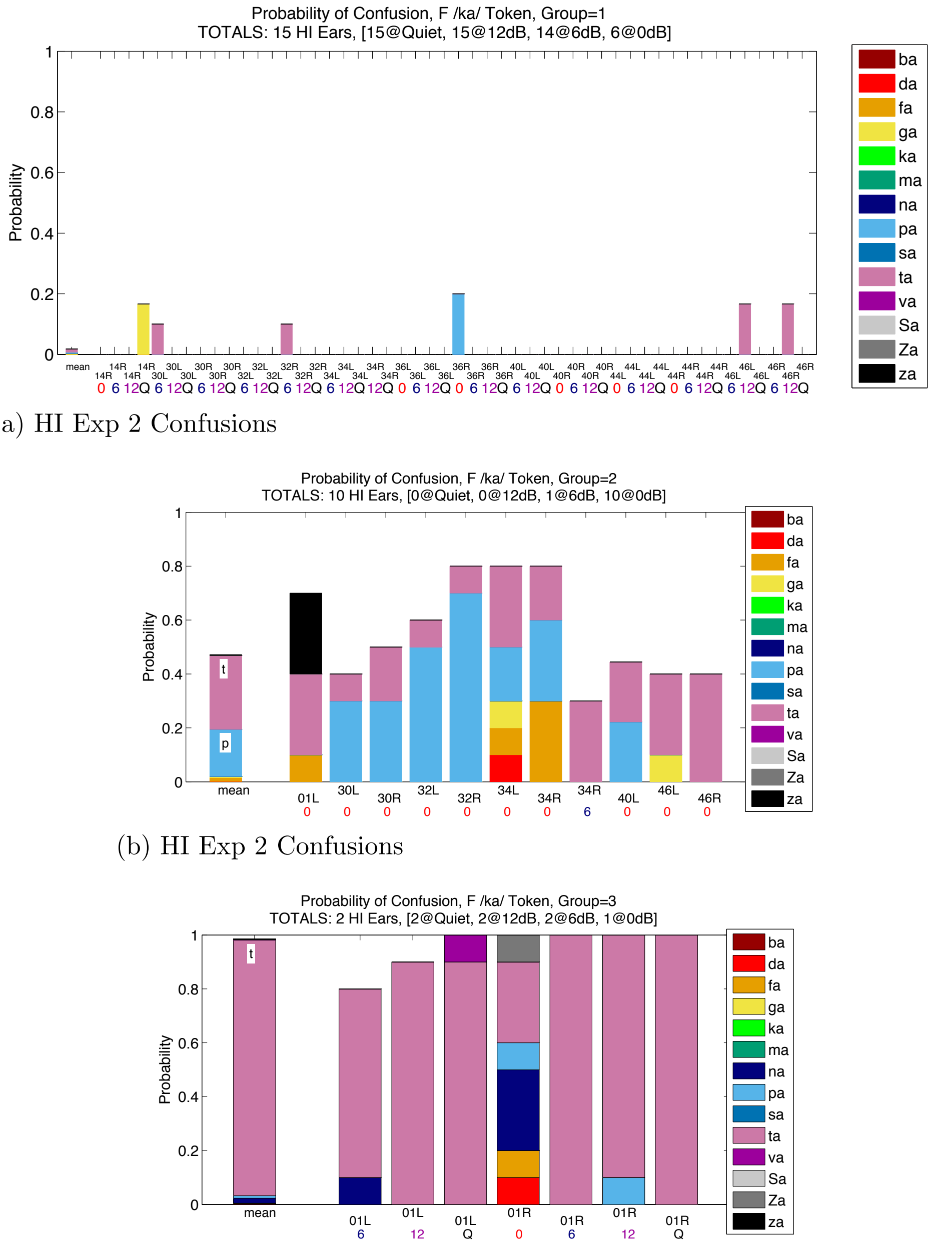
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.15: Overview of data collected for the /kɑ/ token from talker f103. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.

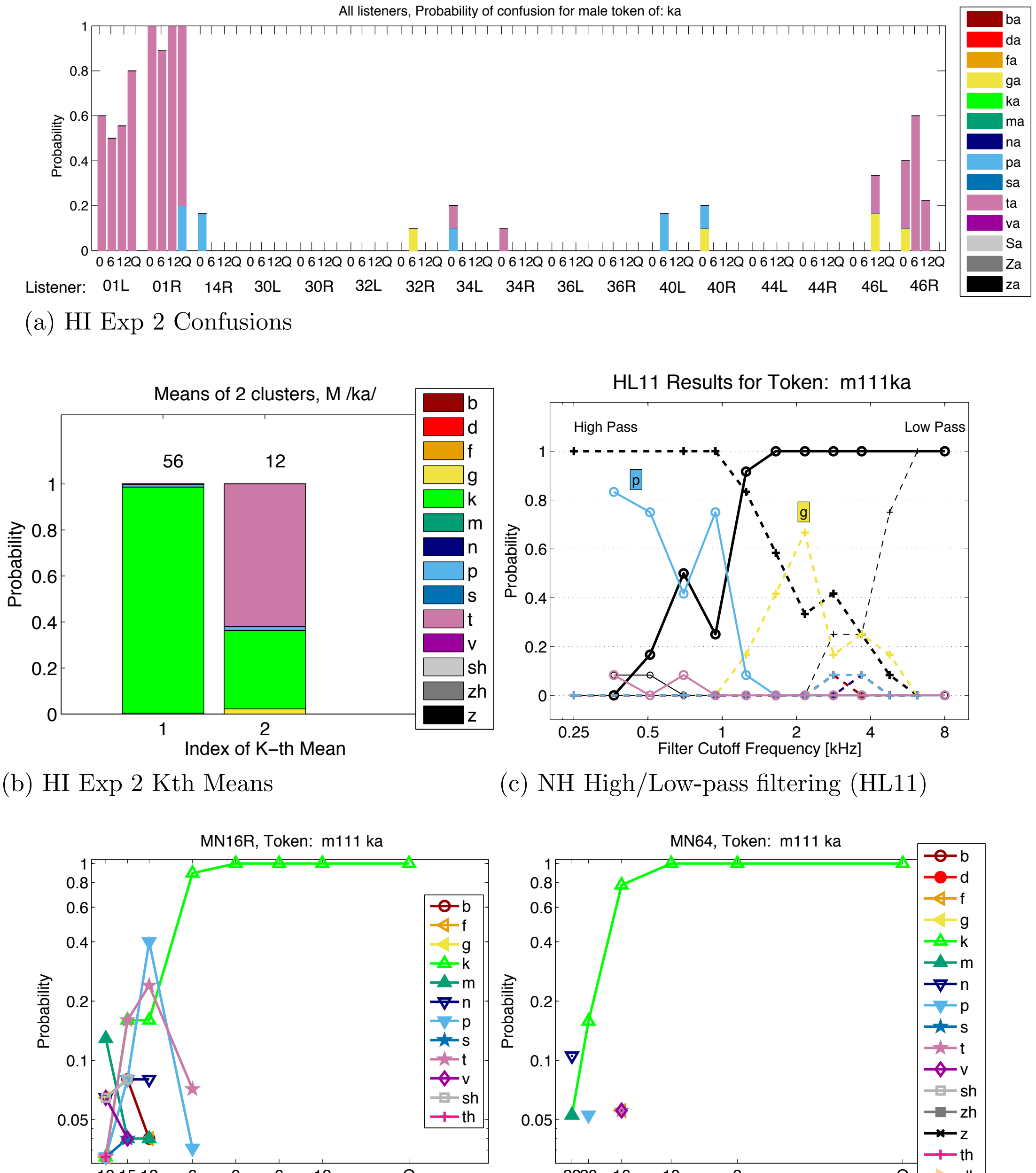


(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

Figure D.16: Each subplot shows the kth cluster mean as well as the collection of data points in that cluster (k indicated by Group = k in each title). Data for f103 kɑ.



(a) HI Exp 2 Confusions

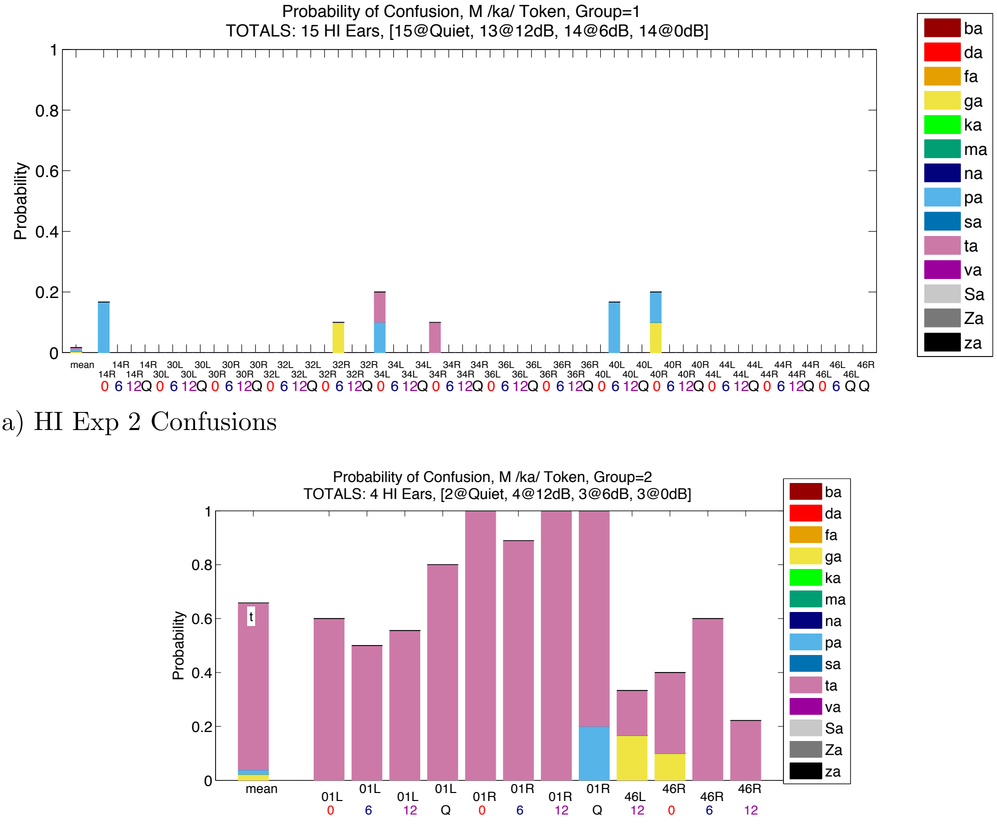
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

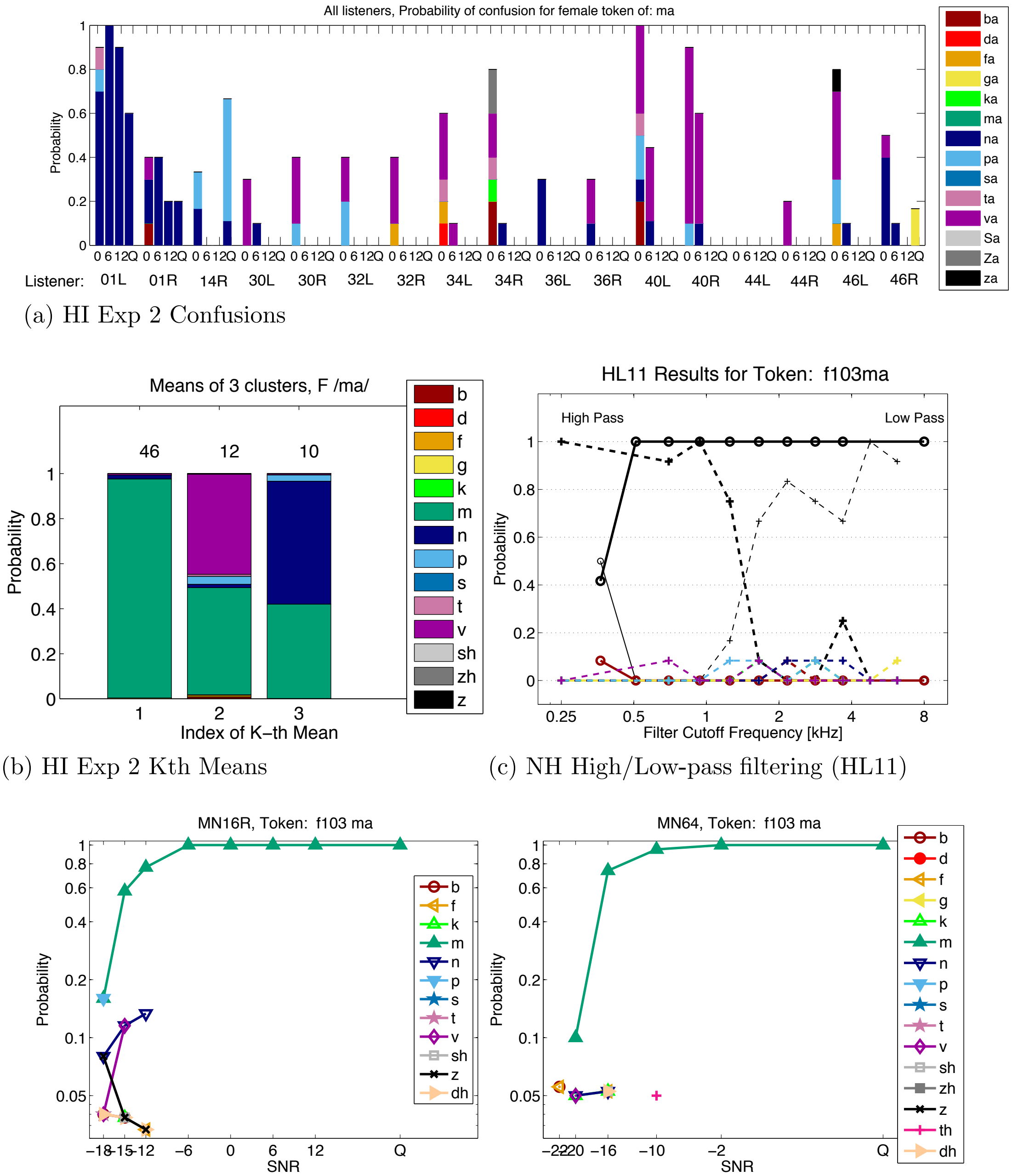
Figure D.17: Overview of data collected for the /kɑ/ token from talker m111. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.18: Each subplot shows the kth cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m111 kɑ.



(a) HI Exp 2 Confusions

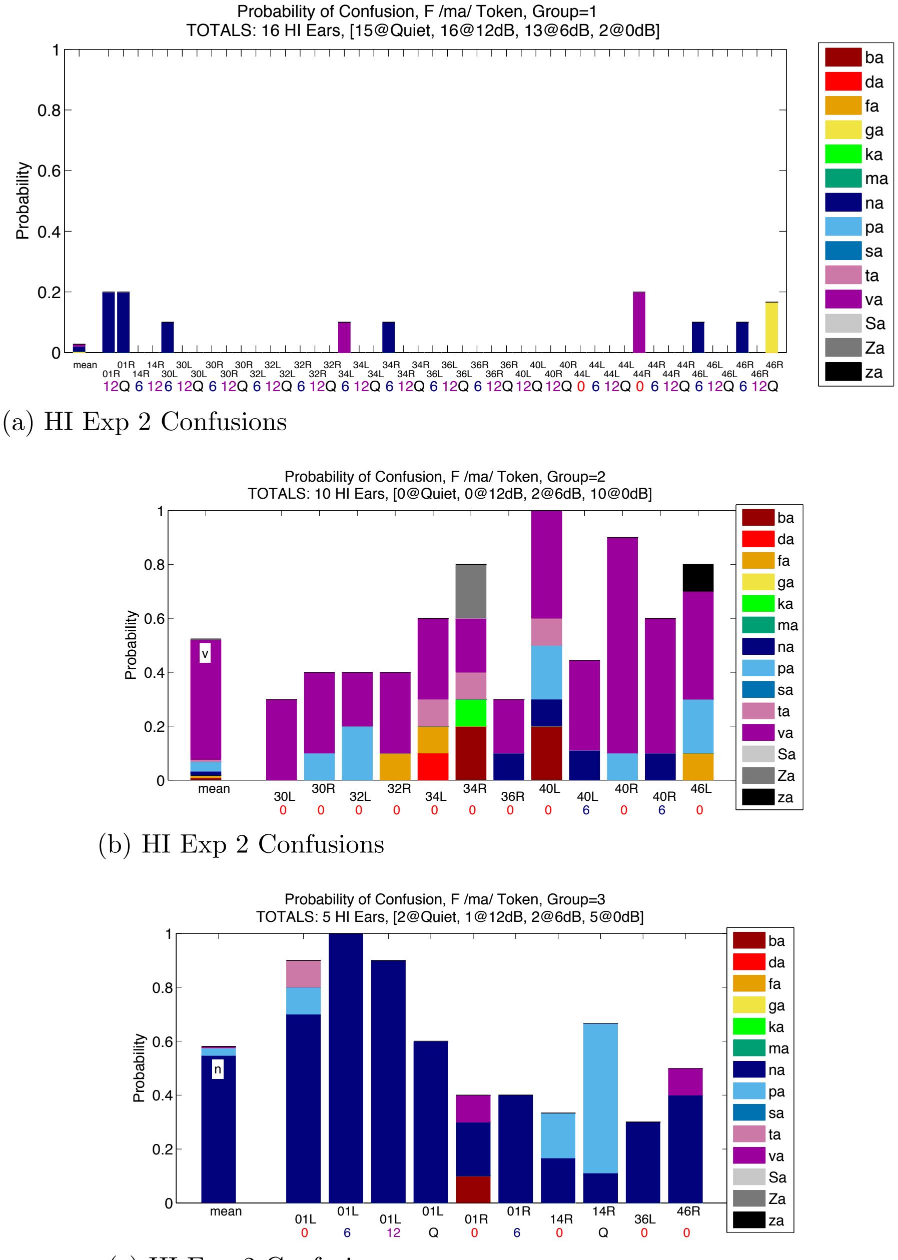
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.19: Overview of data collected for the /mɑ/ token from talker f103. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.

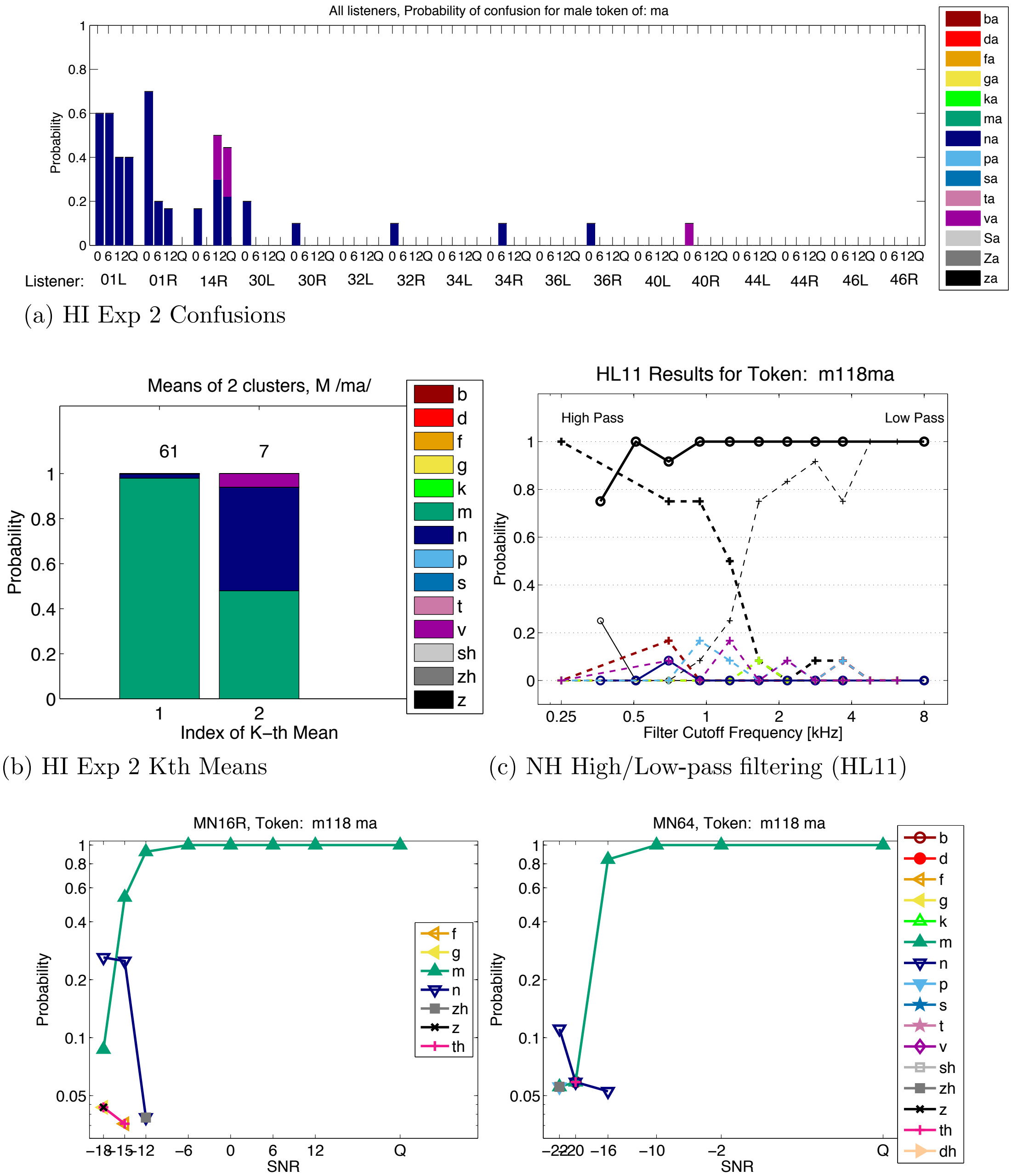


(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

Figure D.20: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f103 ma.



(a) HI Exp 2 Confusions

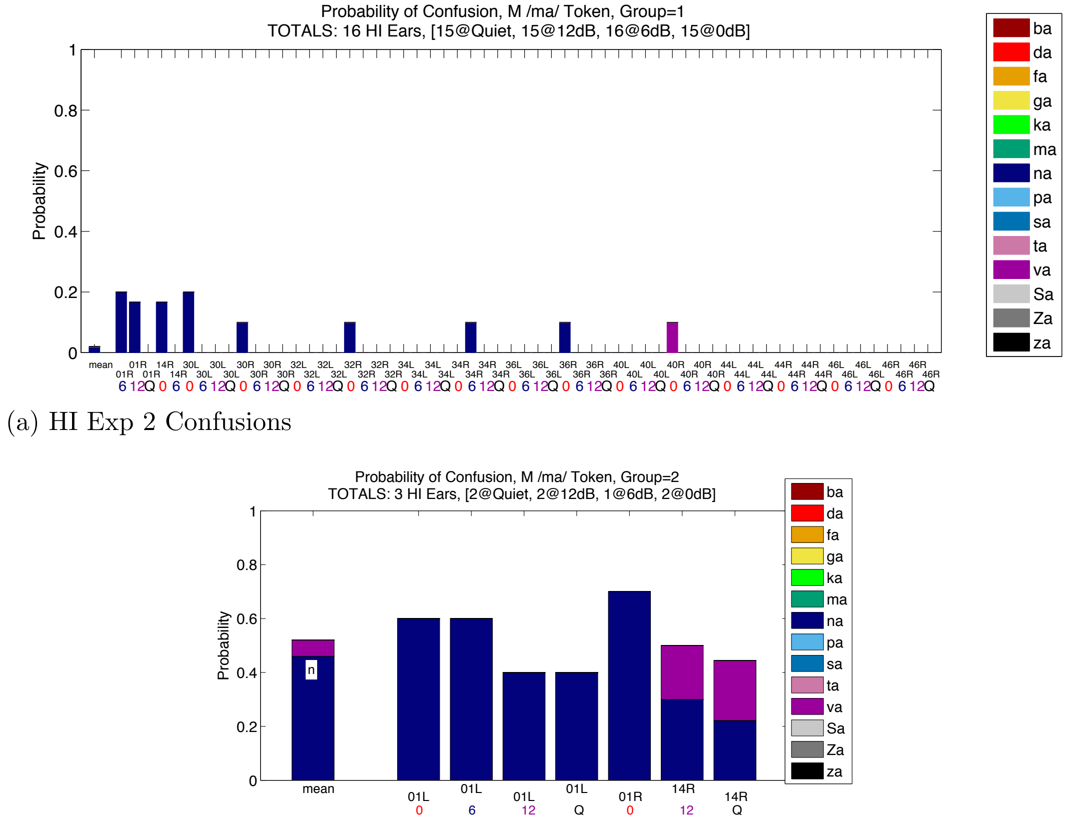
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

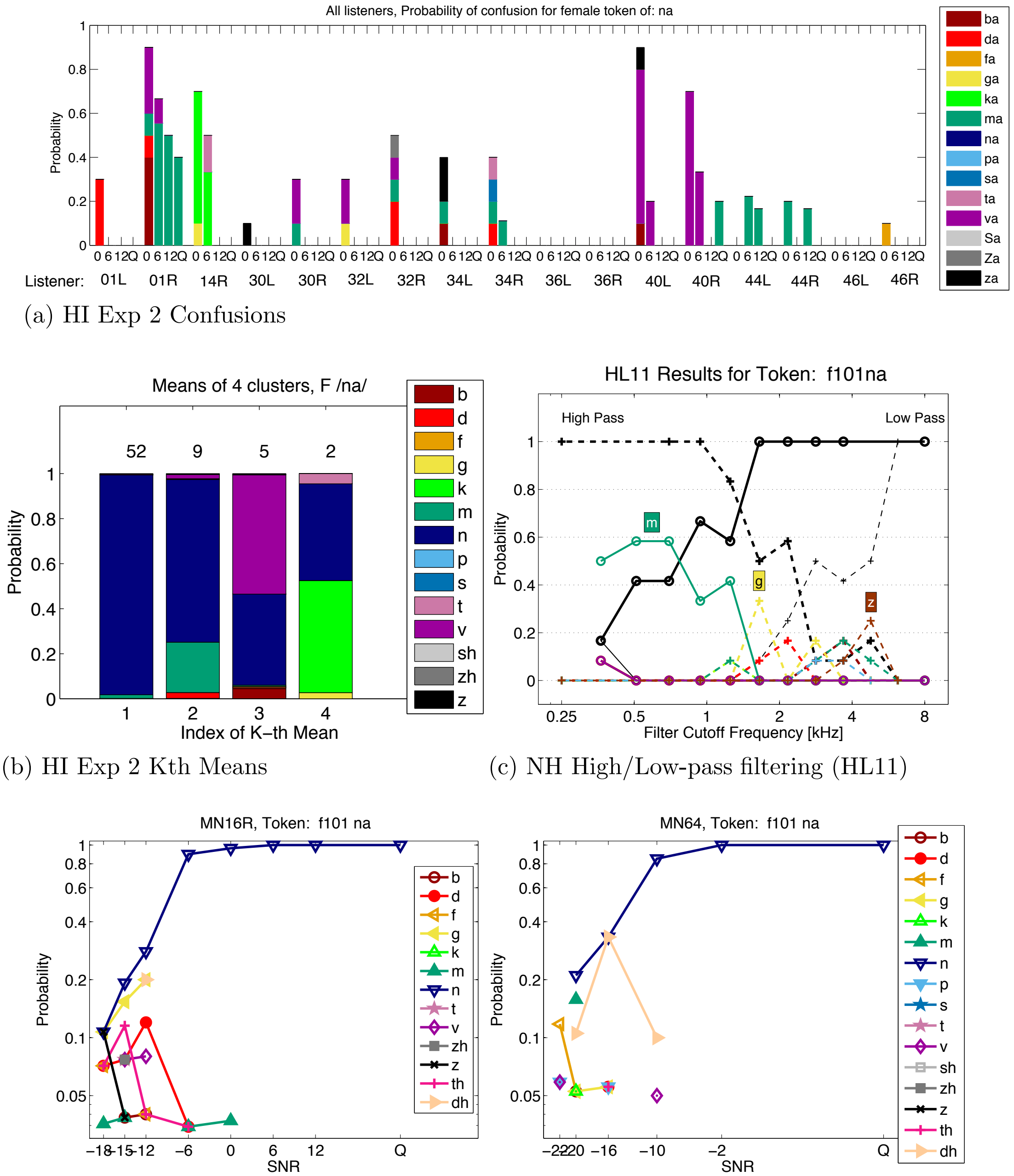
Figure D.21: Overview of data collected for the /mɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.22: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m118 mɑ.



(a) HI Exp 2 Confusions

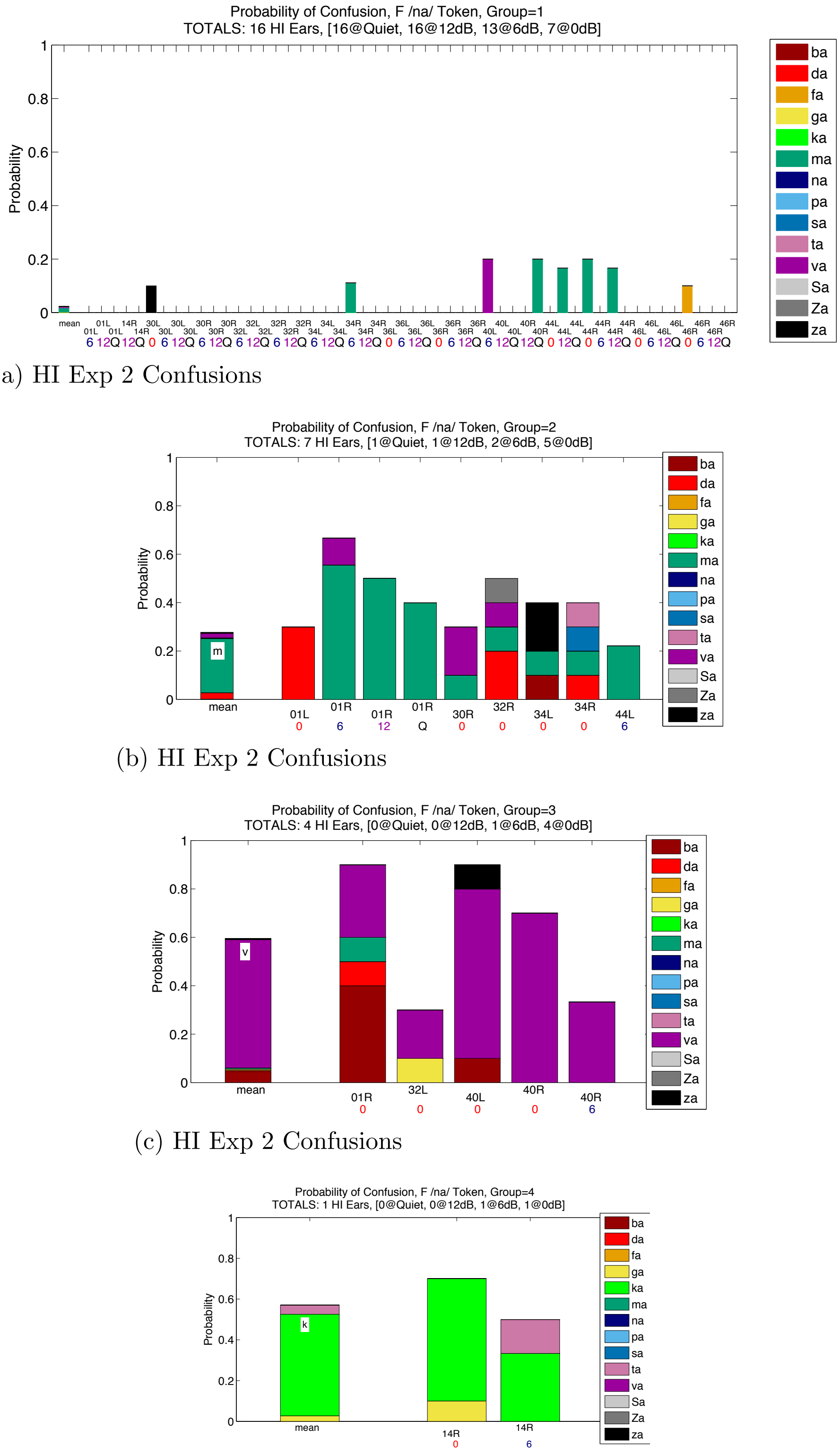
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.23: Overview of data collected for the /nɑ/ token from talker f101. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



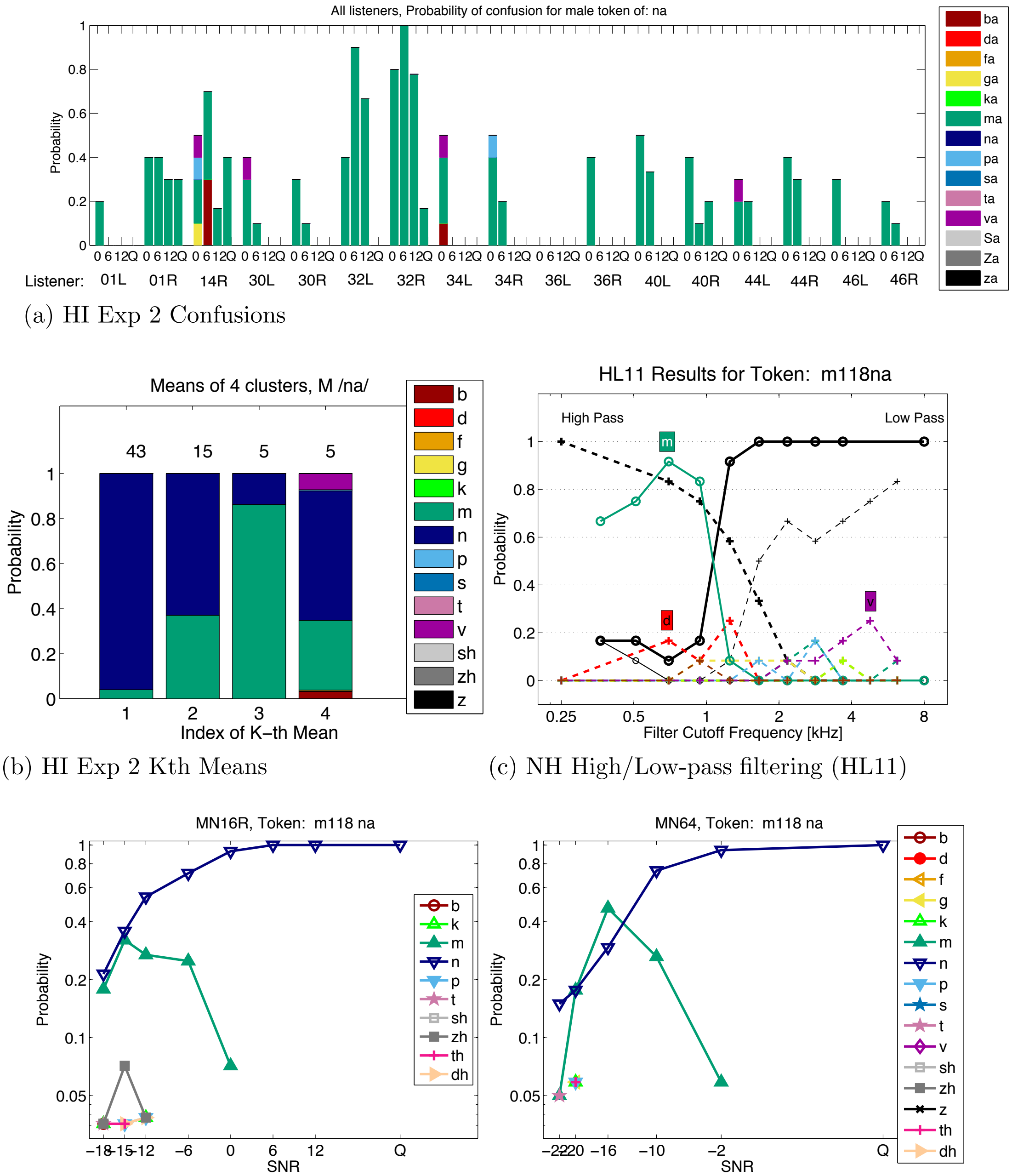
(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

Figure D.24: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f101 nɑ.



(a) HI Exp 2 Confusions

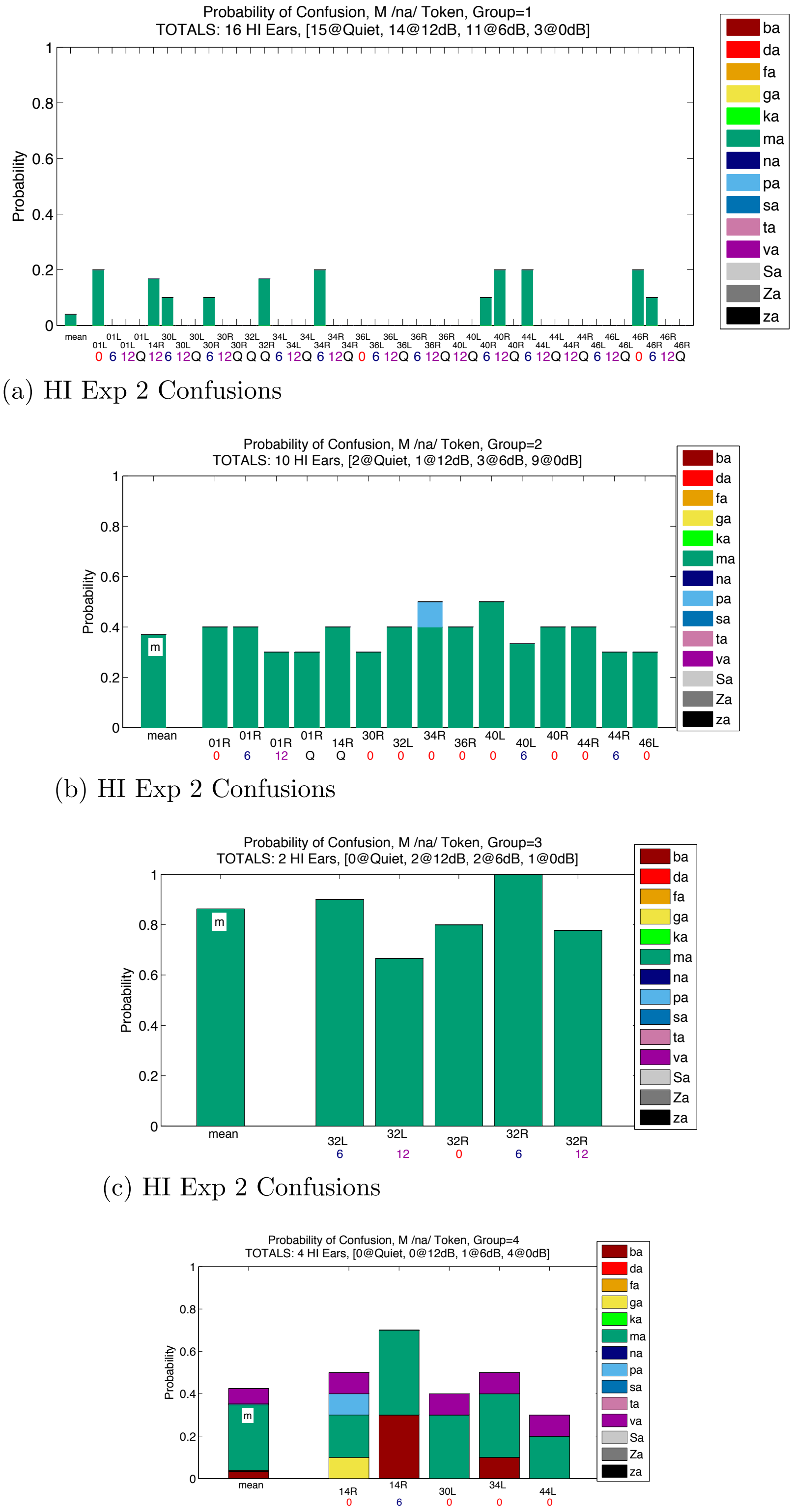
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.25: Overview of data collected for the /nɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



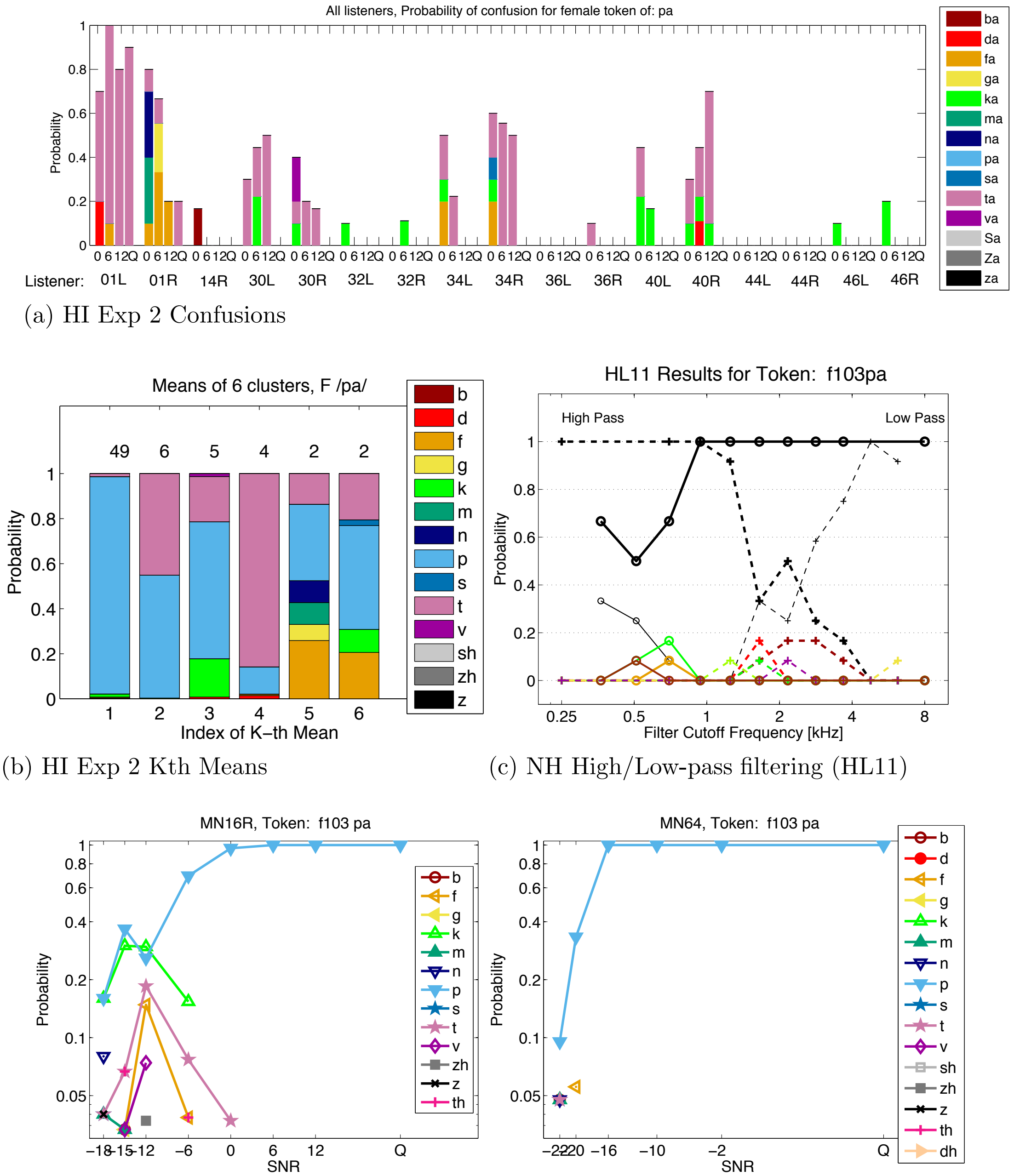
(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

Figure D.26: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m118 nɑ.



(a) HI Exp 2 Confusions

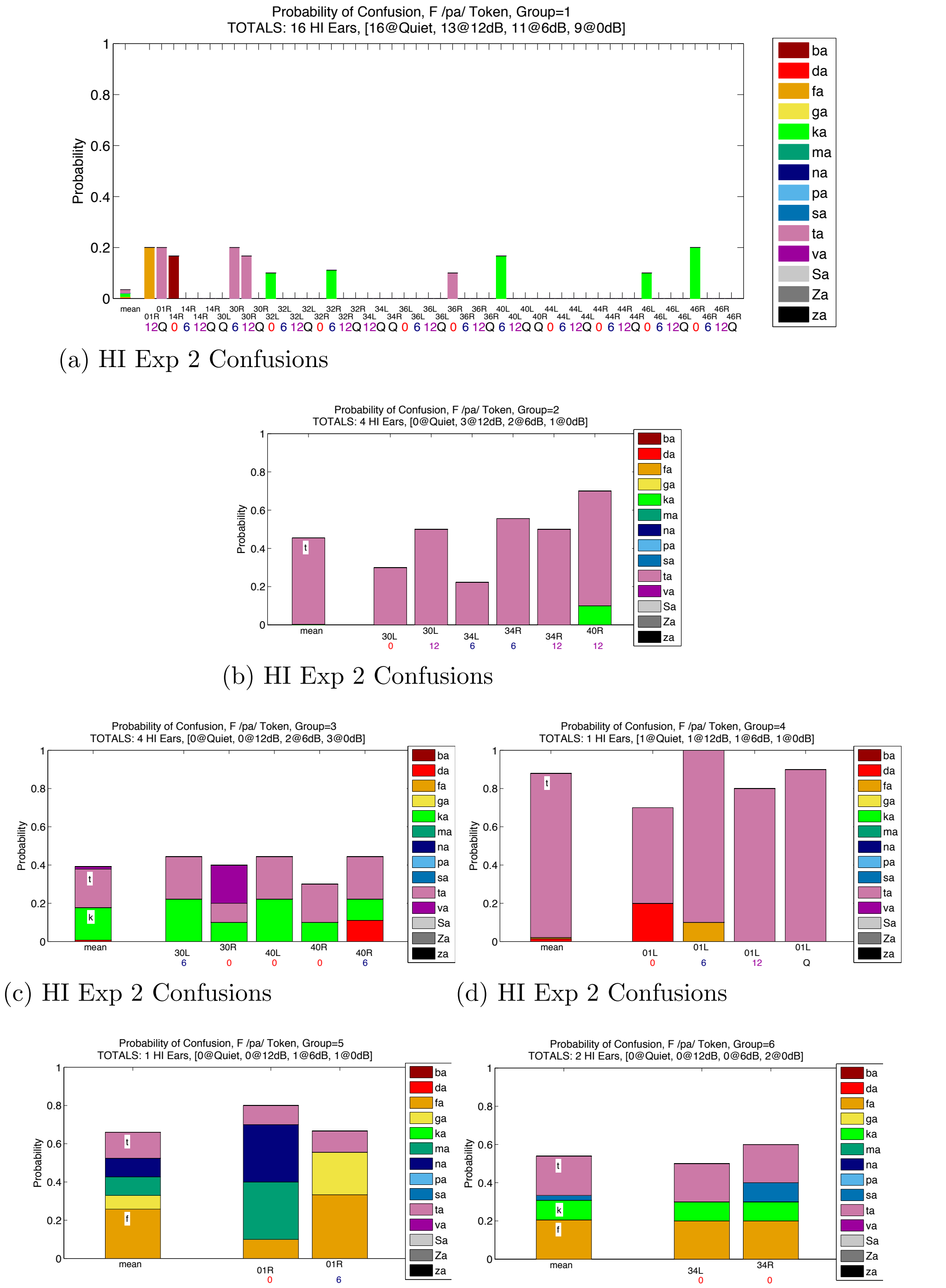
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.27: Overview of data collected for the /pɑ/ token from talker f103. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

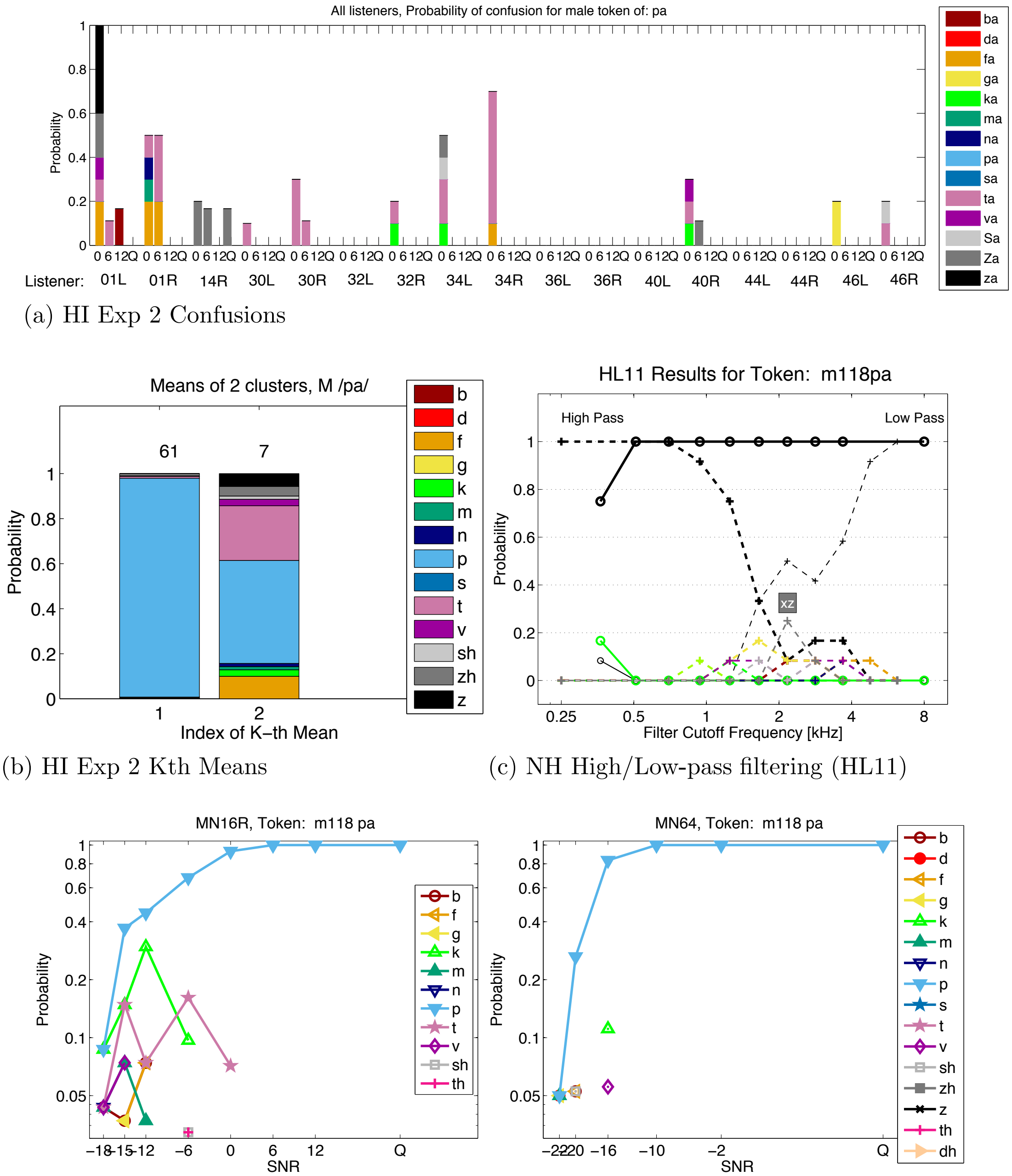
(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

(e) HI Exp 2 Confusions

(f) HI Exp 2 Confusions

Figure D.28: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f103 pɑ.



(a) HI Exp 2 Confusions

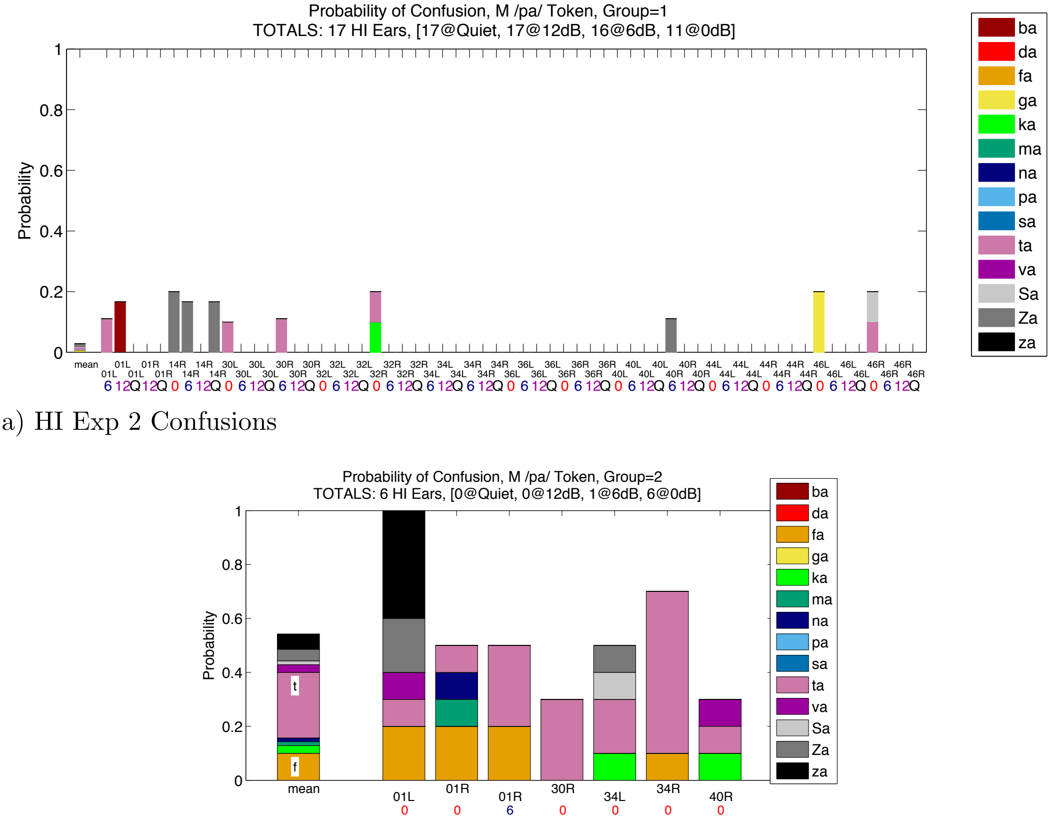
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

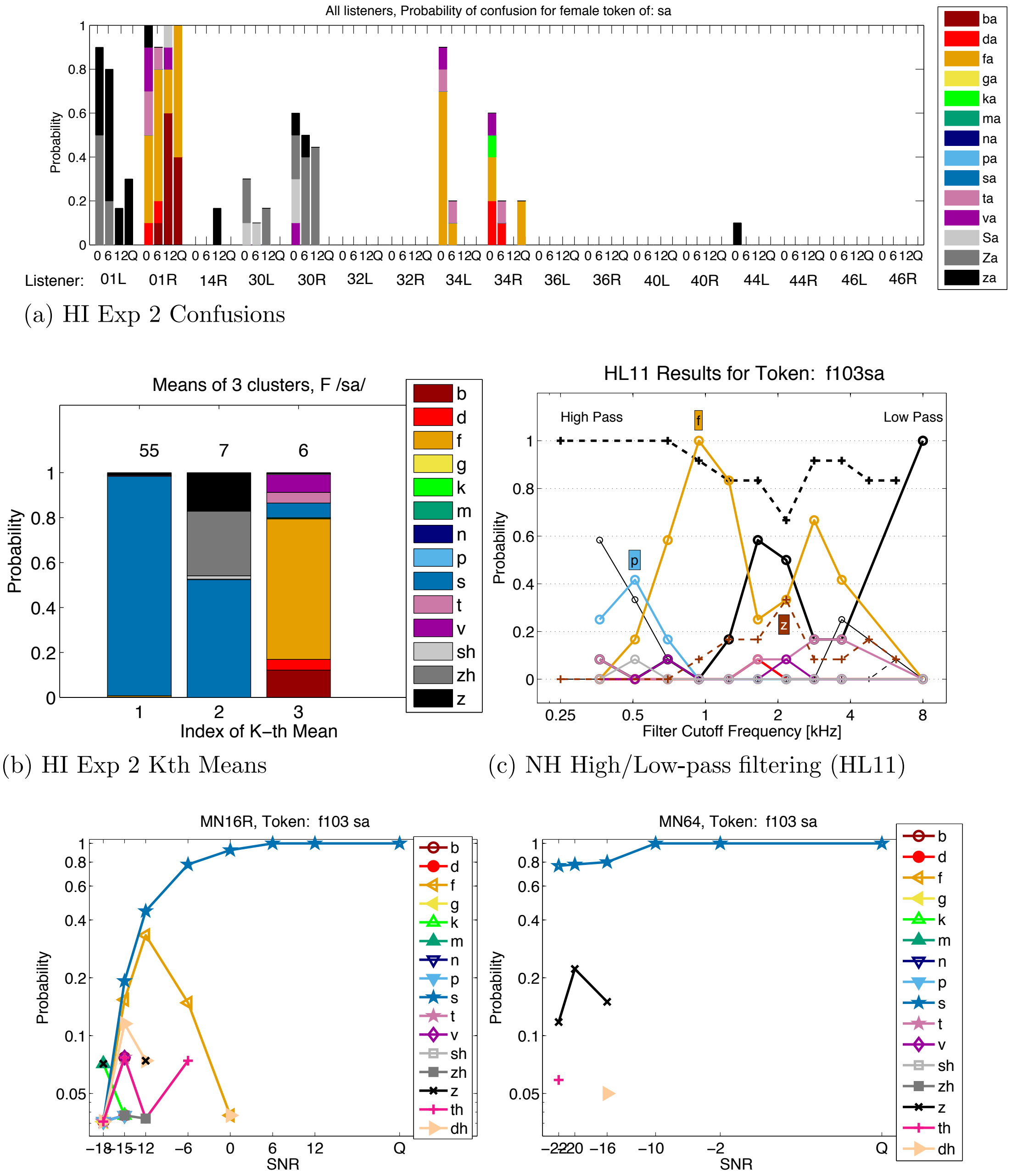
Figure D.29: Overview of data collected for the /pɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.30: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m118 pɑ.



(a) HI Exp 2 Confusions

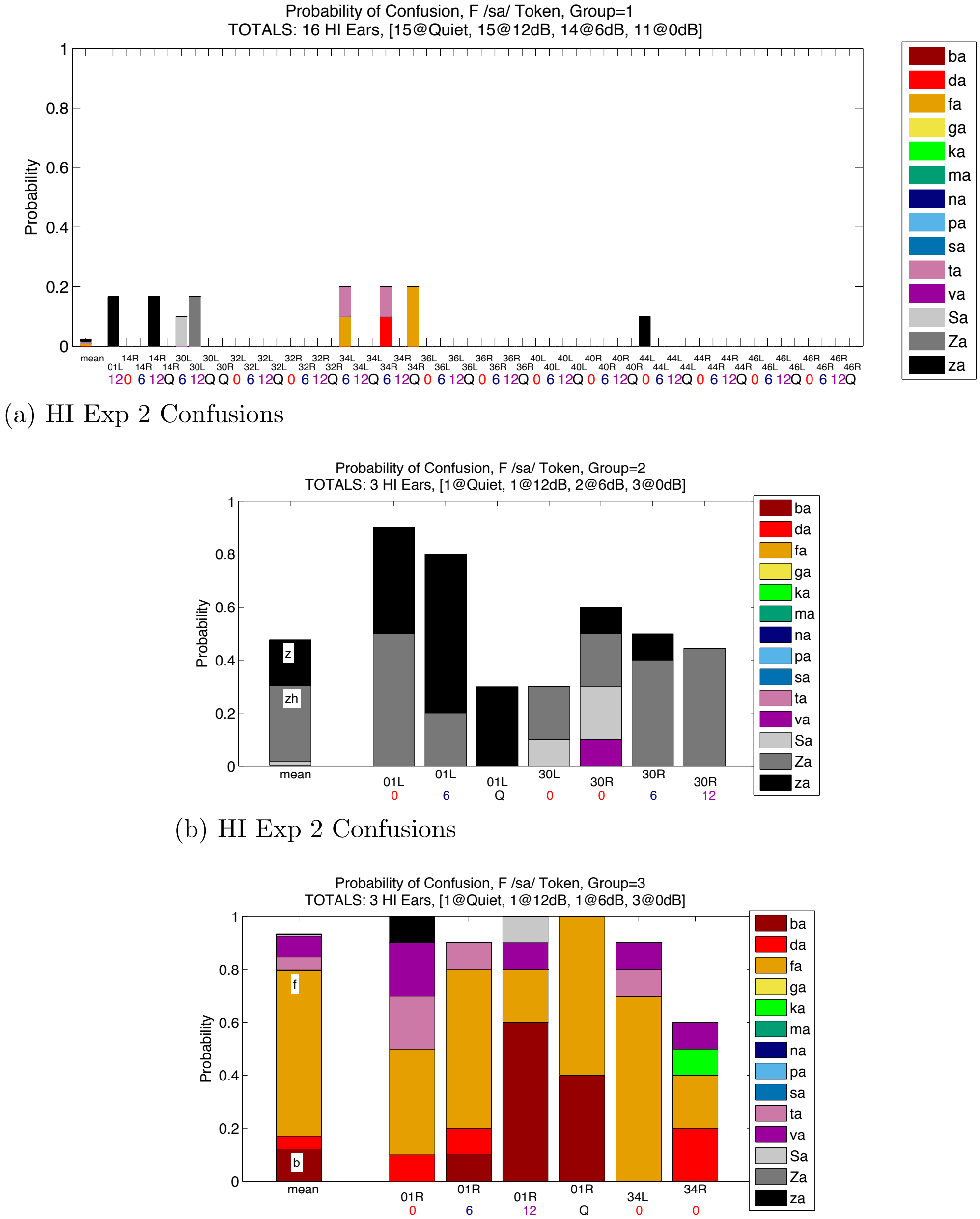
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.31: Overview of data collected for the /sɑ/ token from talker f103. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.

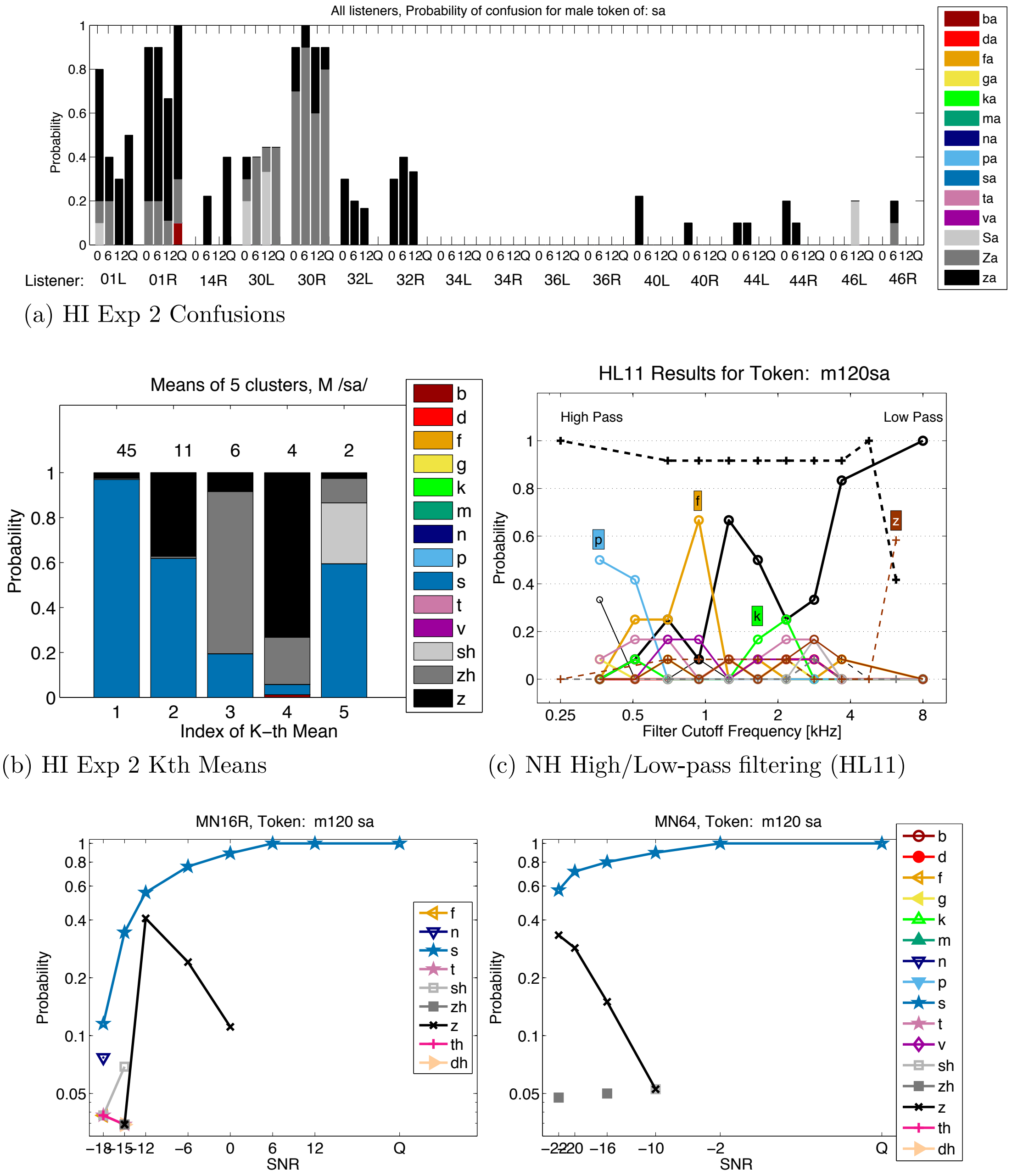


(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

Figure D.32: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f103 sa.



(a) HI Exp 2 Confusions

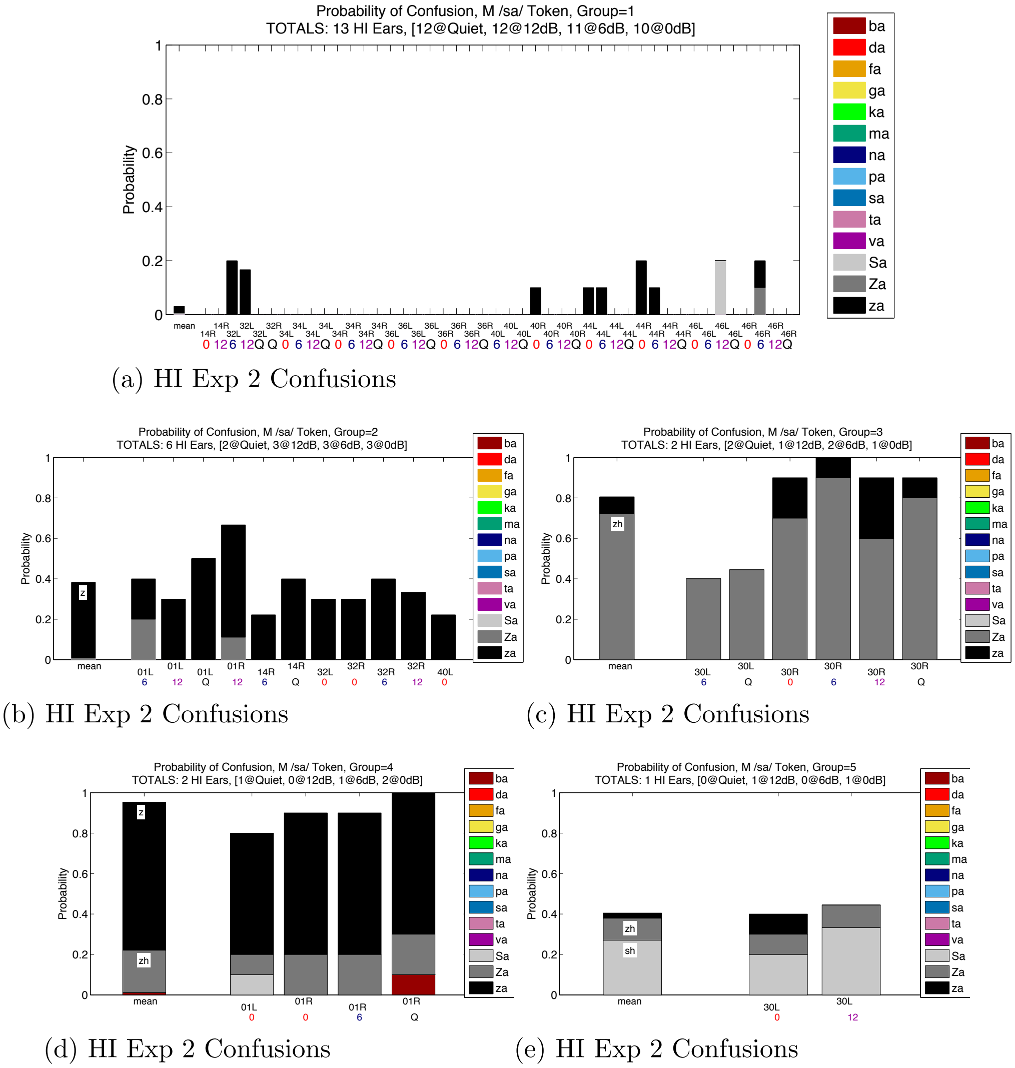
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.33: Overview of data collected for the /sɑ/ token from talker m120. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

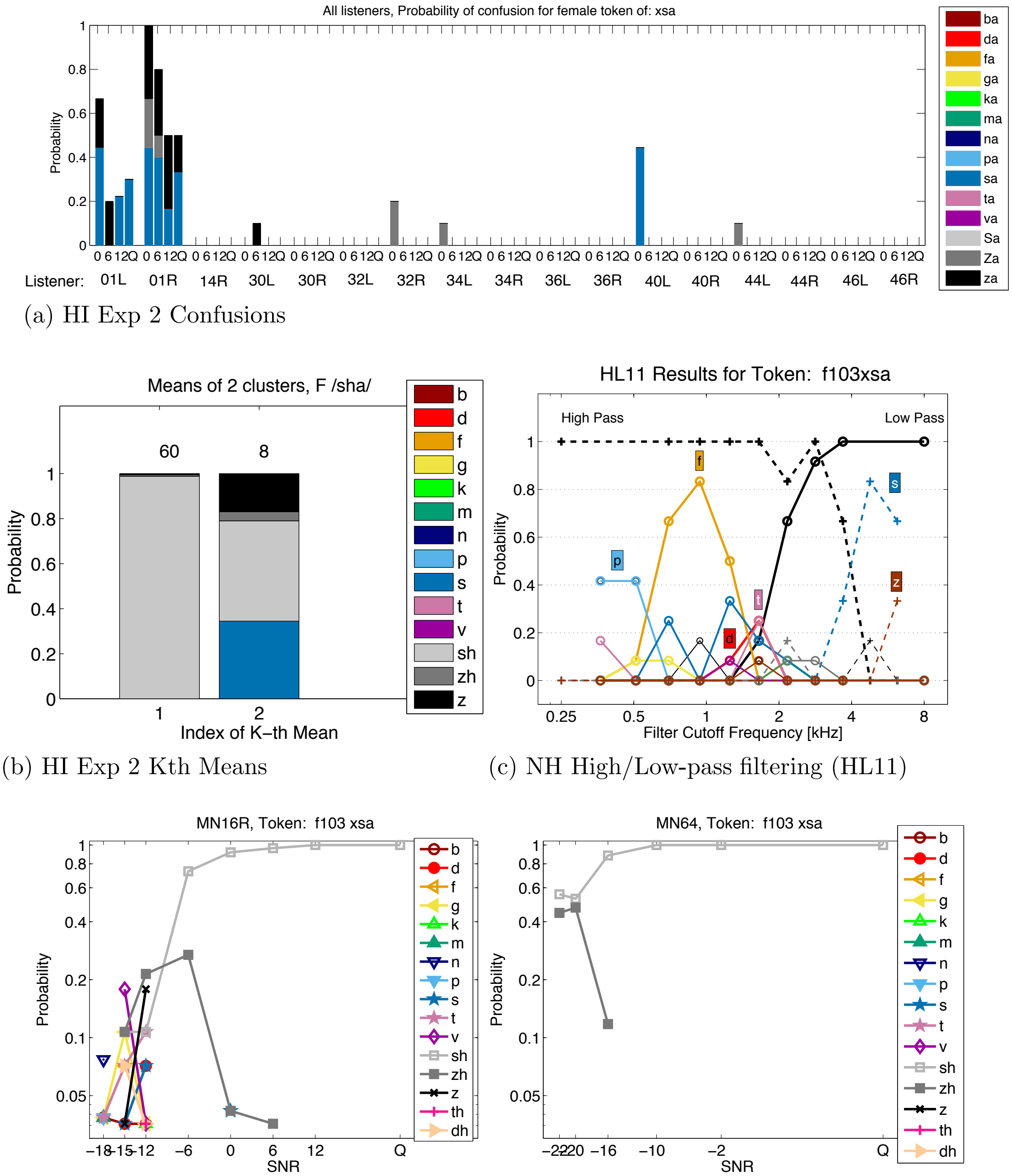
(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

(e) HI Exp 2 Confusions

Figure D.34: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group = k in each title). Data for m120 sa.



(a) HI Exp 2 Confusions

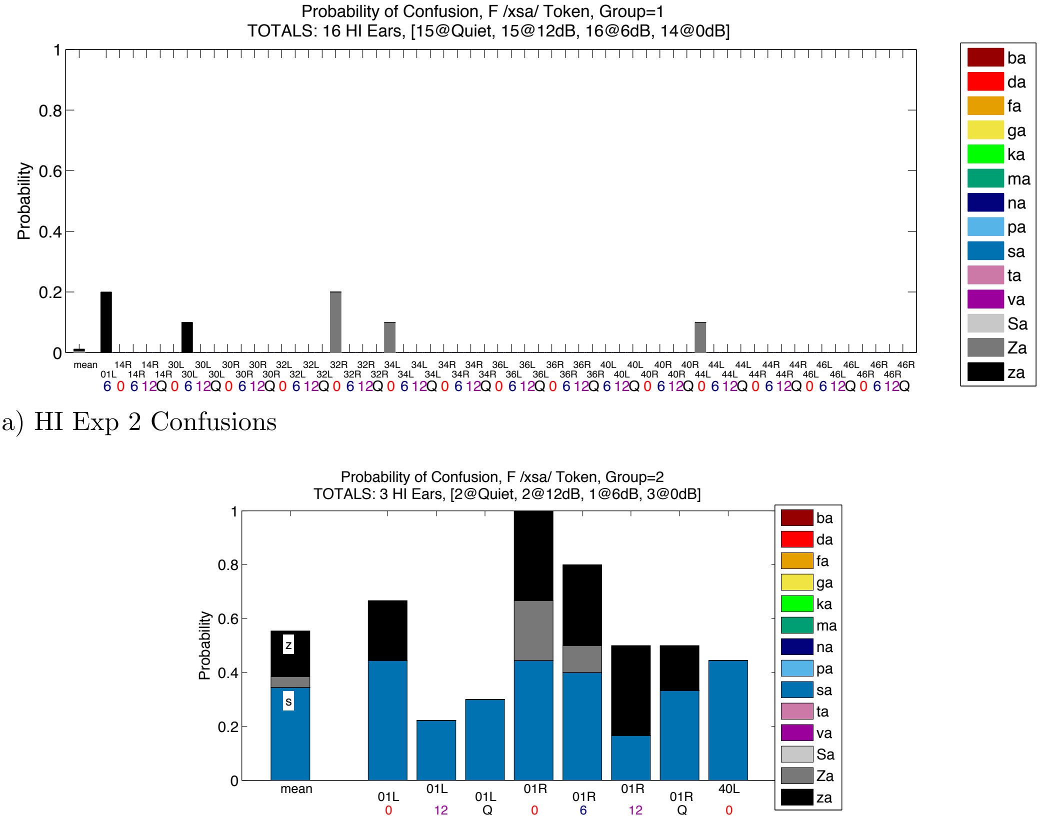
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

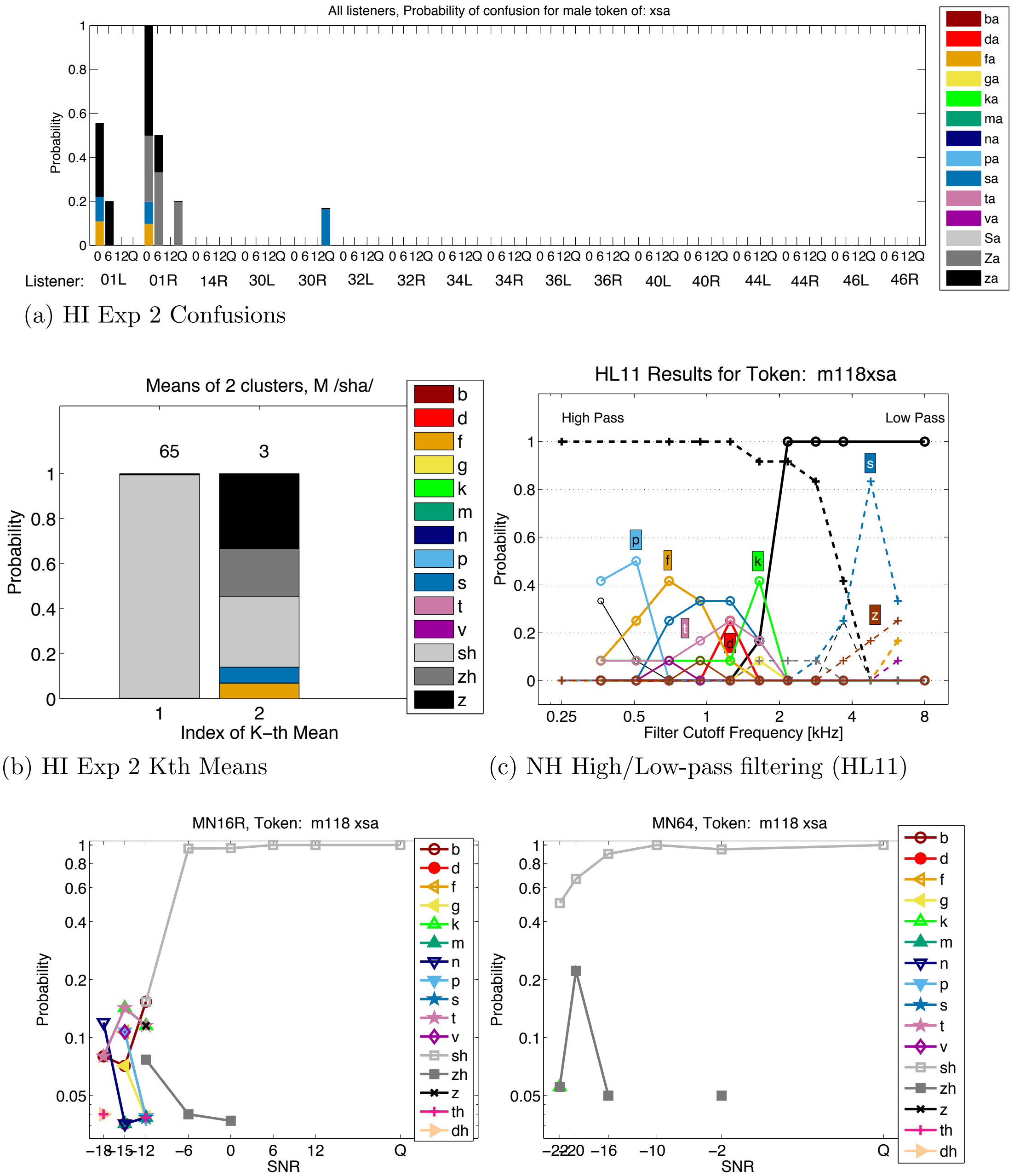
Figure D.35: Overview of data collected for the /ʃɑ/ token from talker f103. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.36: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f103 ʃɑ.



(a) HI Exp 2 Confusions

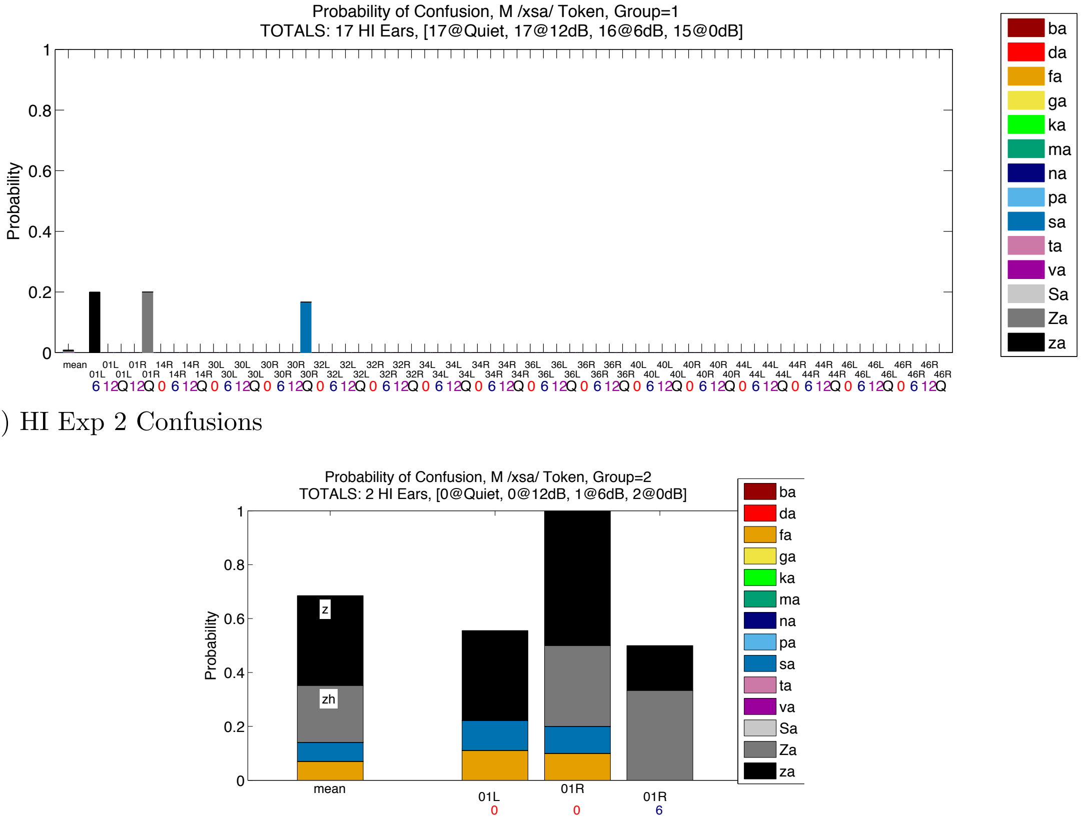
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

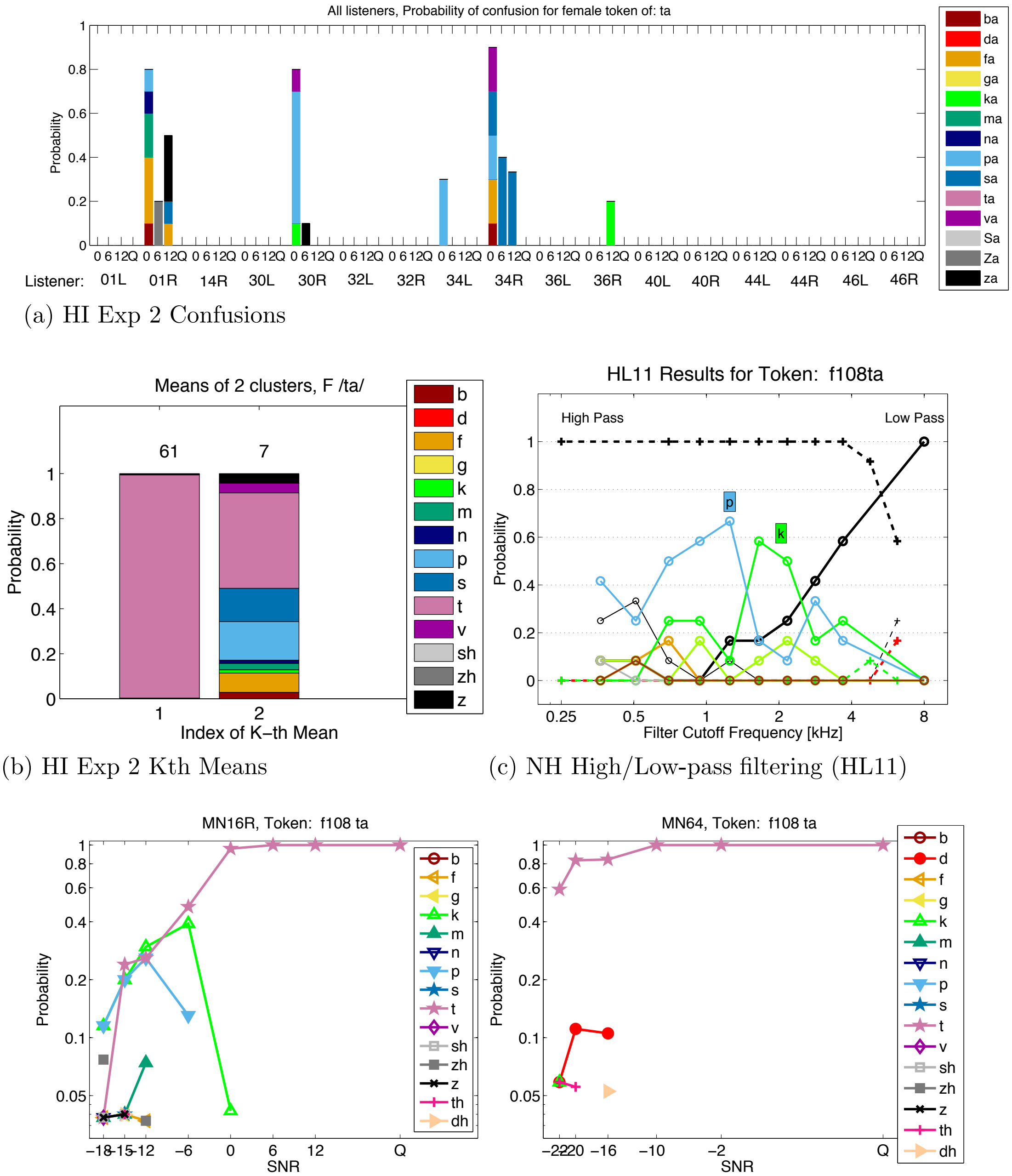
Figure D.37: Overview of data collected for the /ʃɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.38: Each subplot shows the kth cluster mean as well as the collection of data points in that cluster (k indicated by Group = k in each title). Data for m118 ʃɑ.



(a) HI Exp 2 Confusions

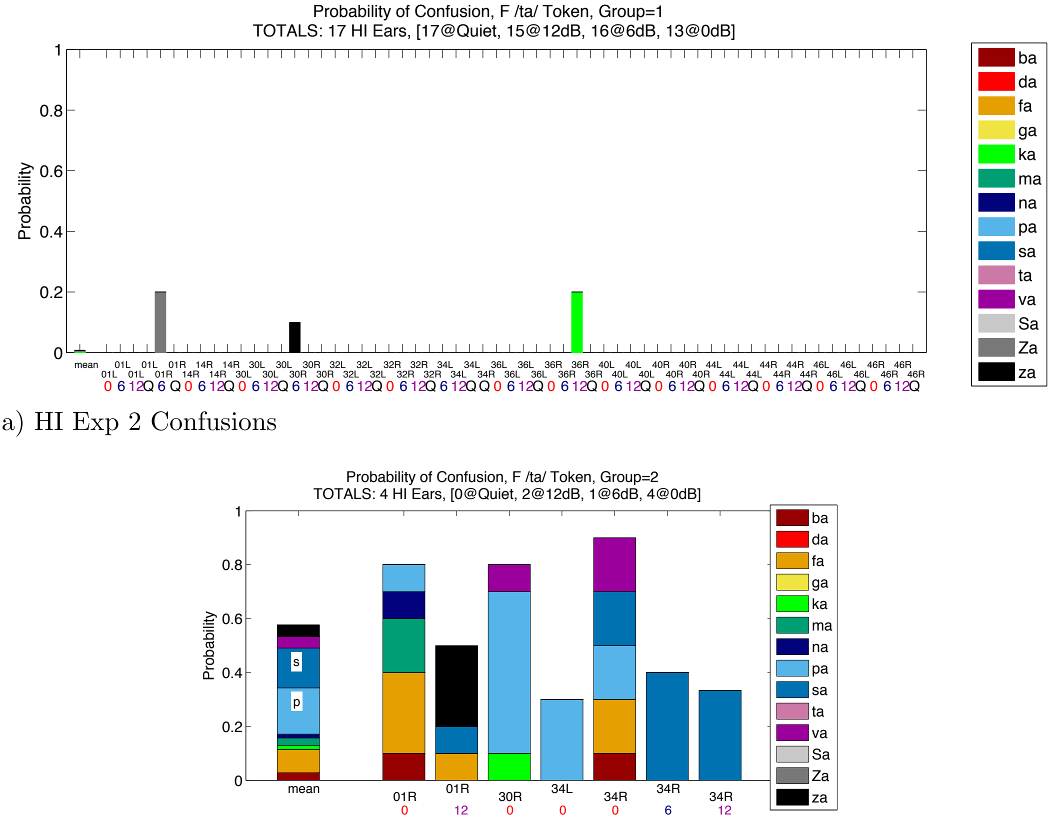
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

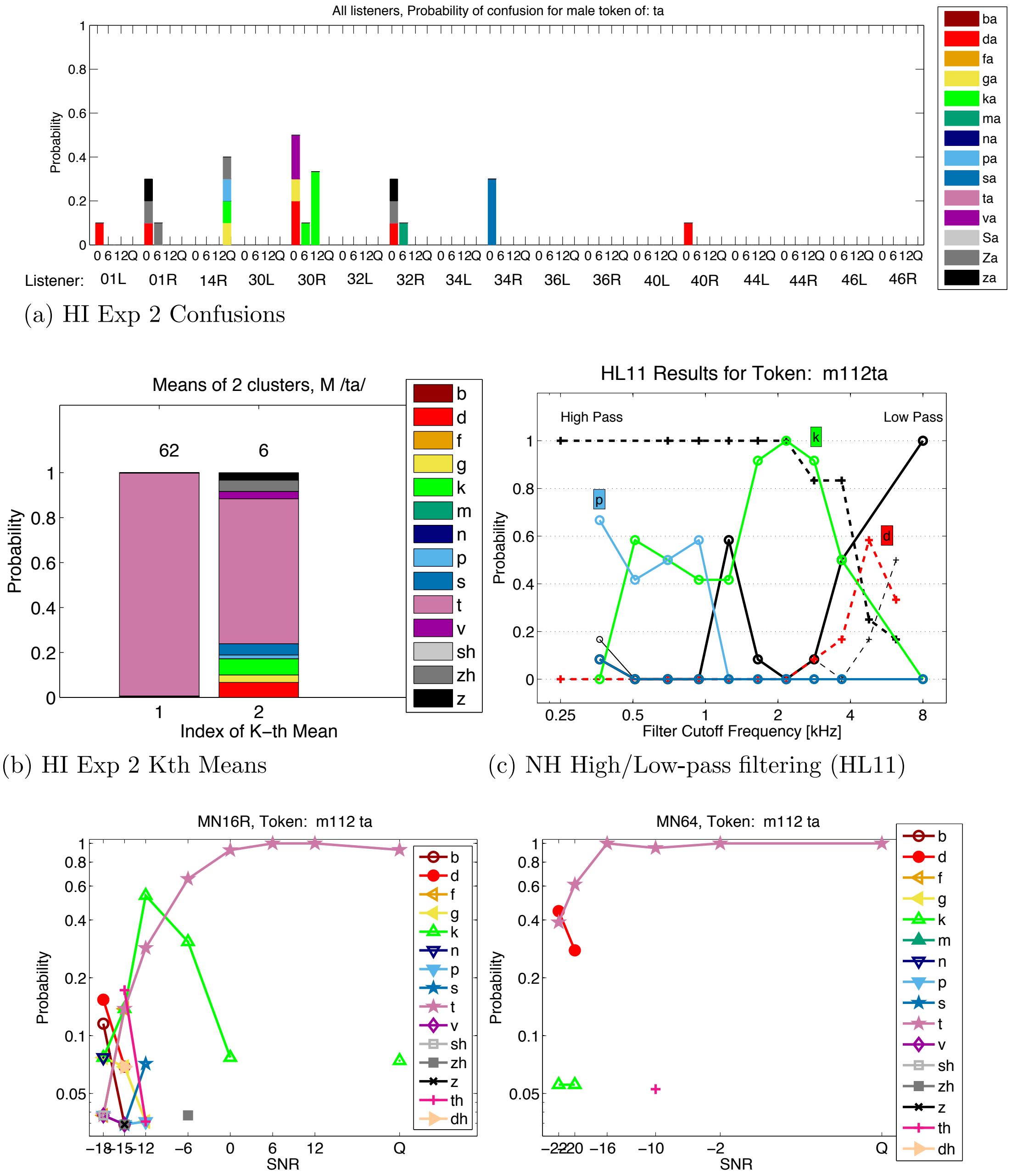
Figure D.39: Overview of data collected for the /tɑ/ token from talker f108. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.40: Each subplot shows the kth cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f108 tɑ.



(a) HI Exp 2 Confusions

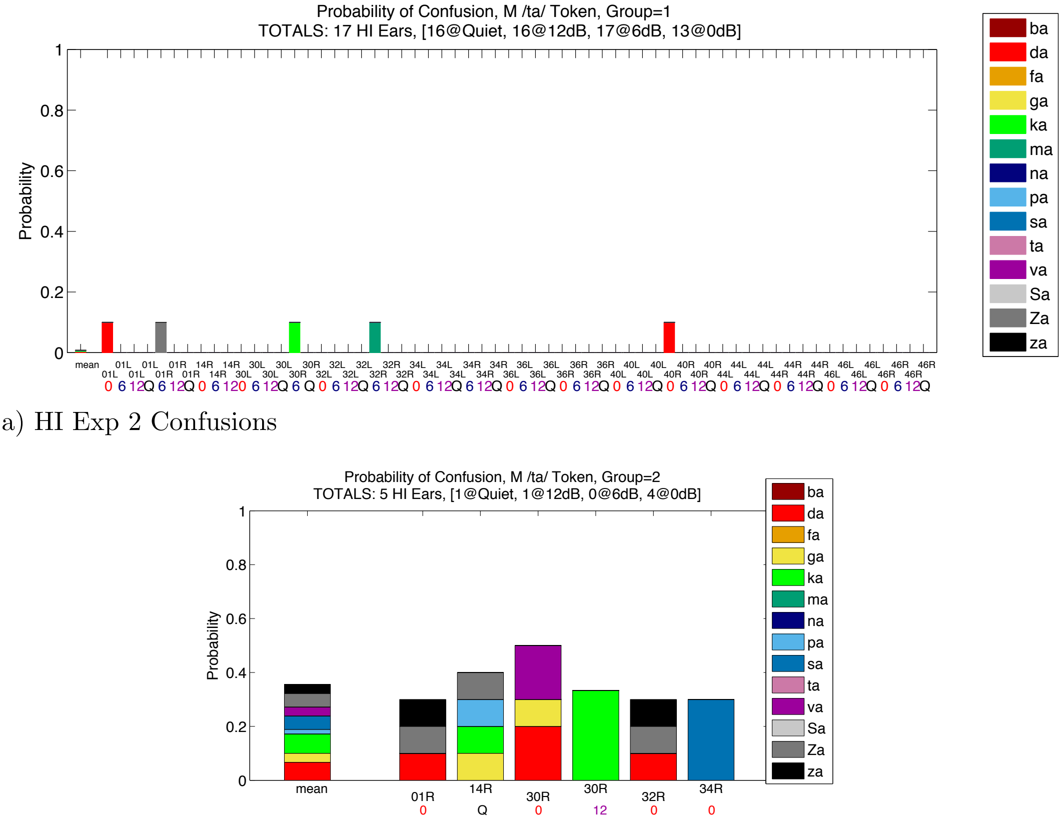
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

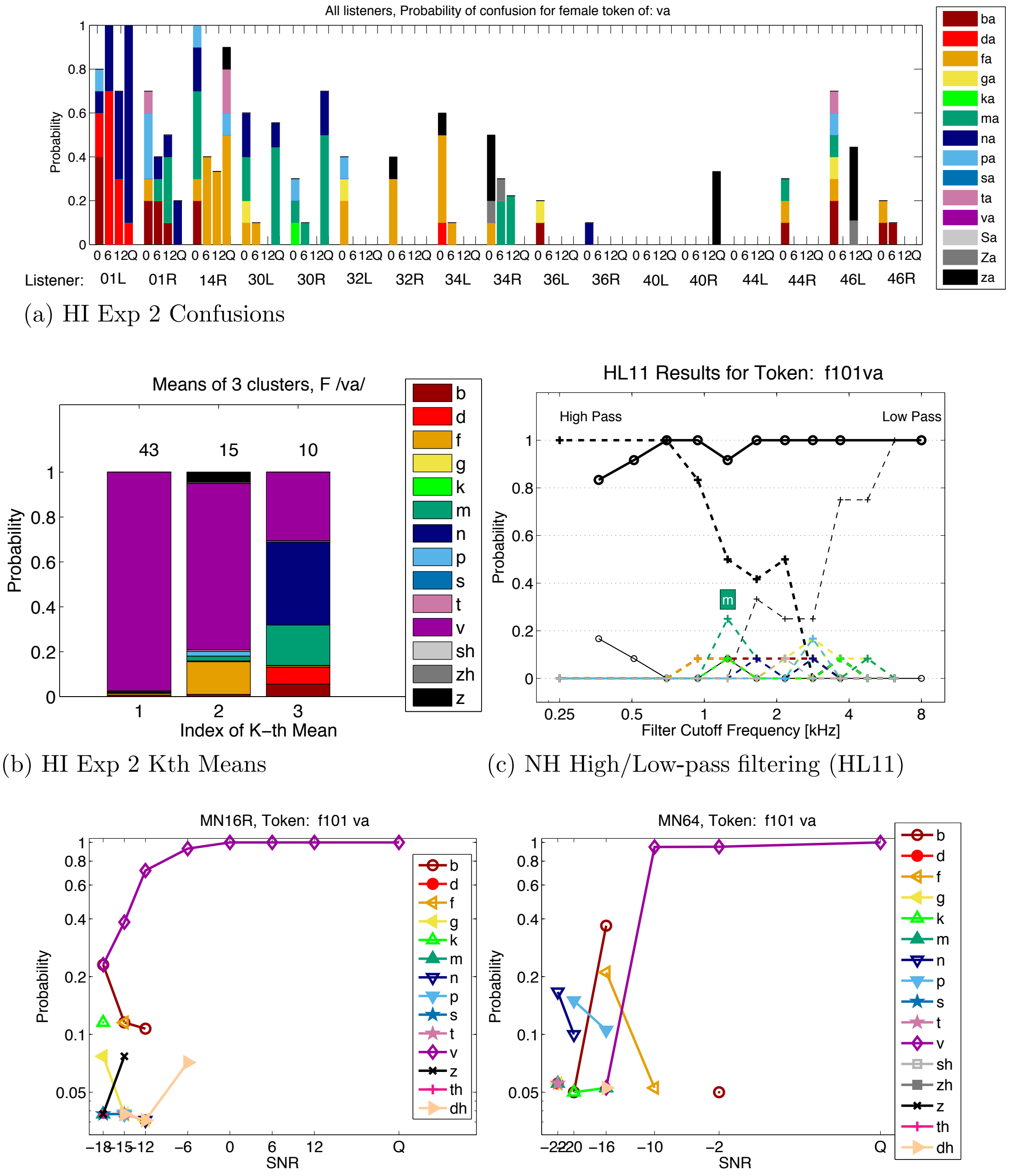
Figure D.41: Overview of data collected for the /tɑ/ token from talker m112. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

Figure D.42: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m112 tɑ.



(a) HI Exp 2 Confusions

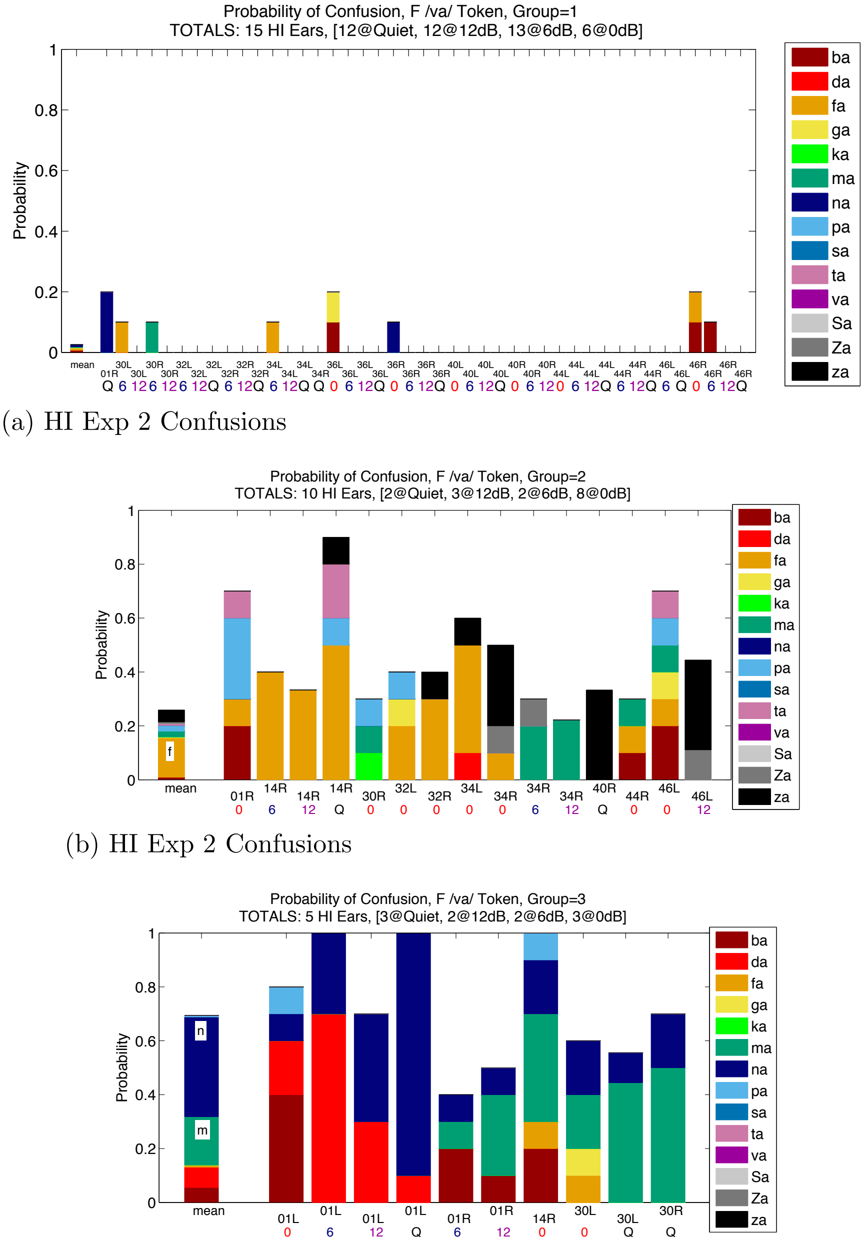
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.43: Overview of data collected for the /va/ token from talker f101. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.

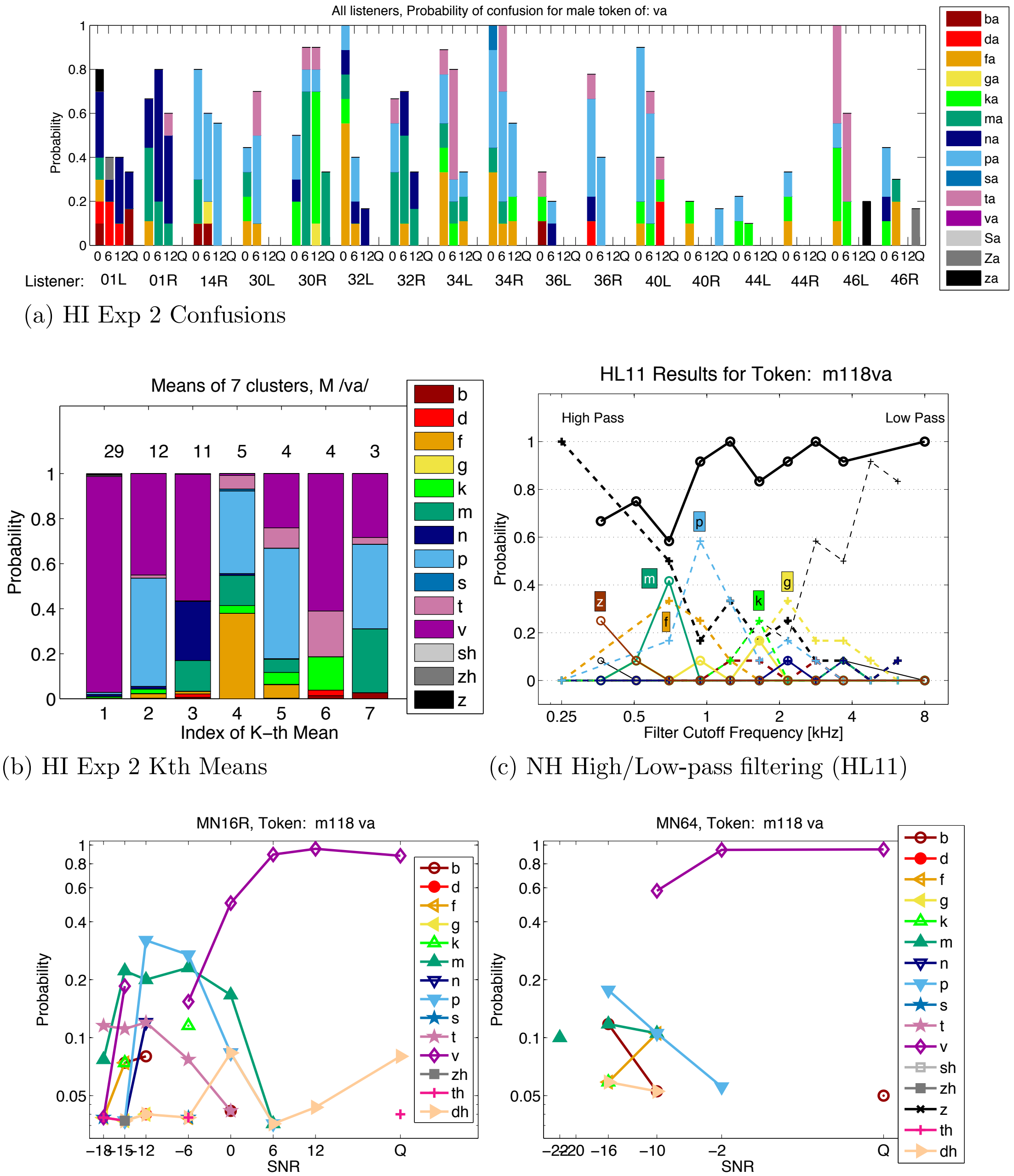


(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

Figure D.44: Each subplot shows the kth cluster mean as well as the collection of data points in that cluster (k indicated by Group = k in each title). Data for f101 vɑ.



(a) HI Exp 2 Confusions

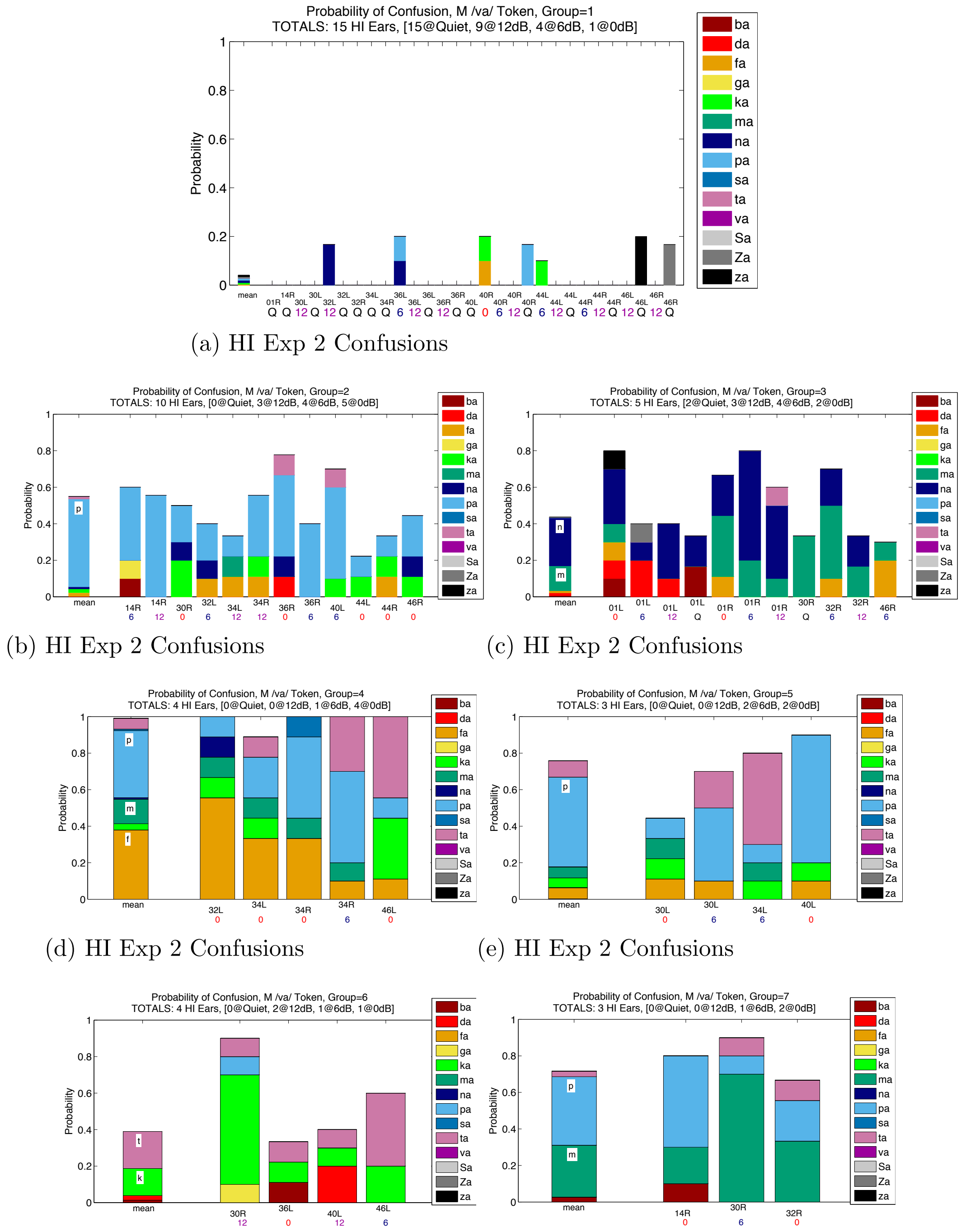
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.45: Overview of data collected for the /vɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

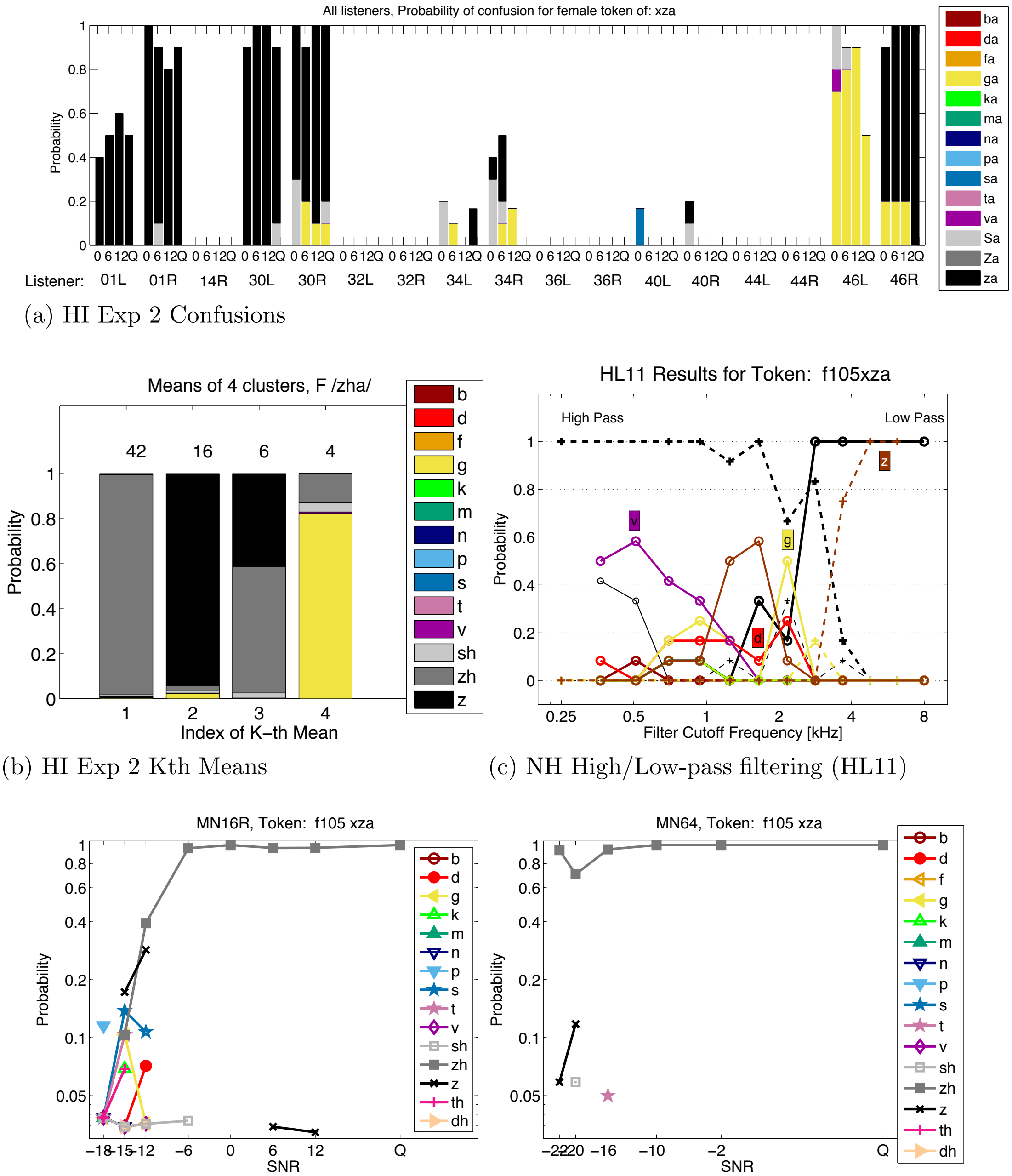
(d) HI Exp 2 Confusions

(e) HI Exp 2 Confusions

(f) HI Exp 2 Confusions

(g) HI Exp 2 Confusions

Figure D.46: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m118 vɑ.



(a) HI Exp 2 Confusions

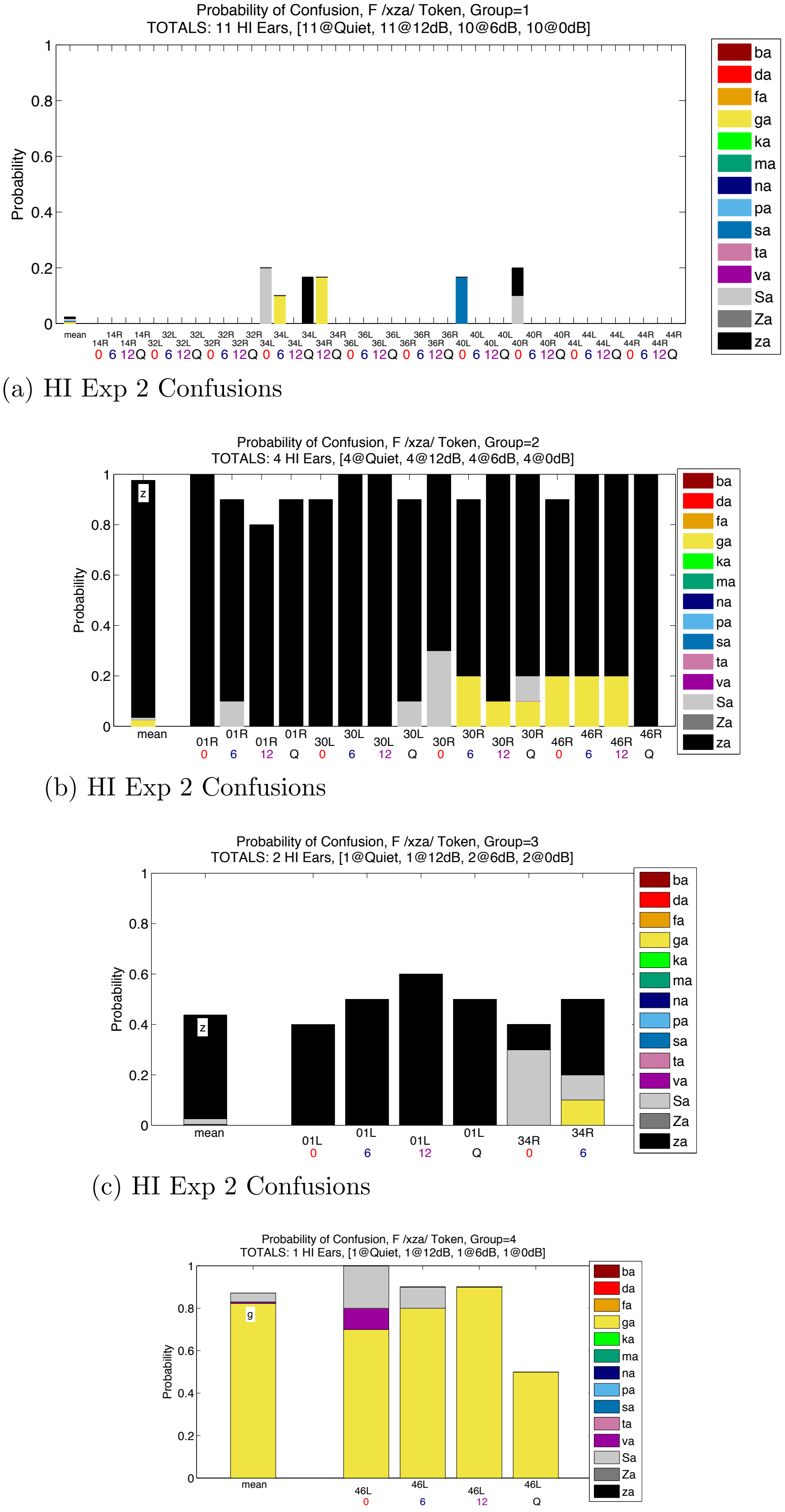
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.47: Overview of data collected for the /ʒɑ/ token from talker f105. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



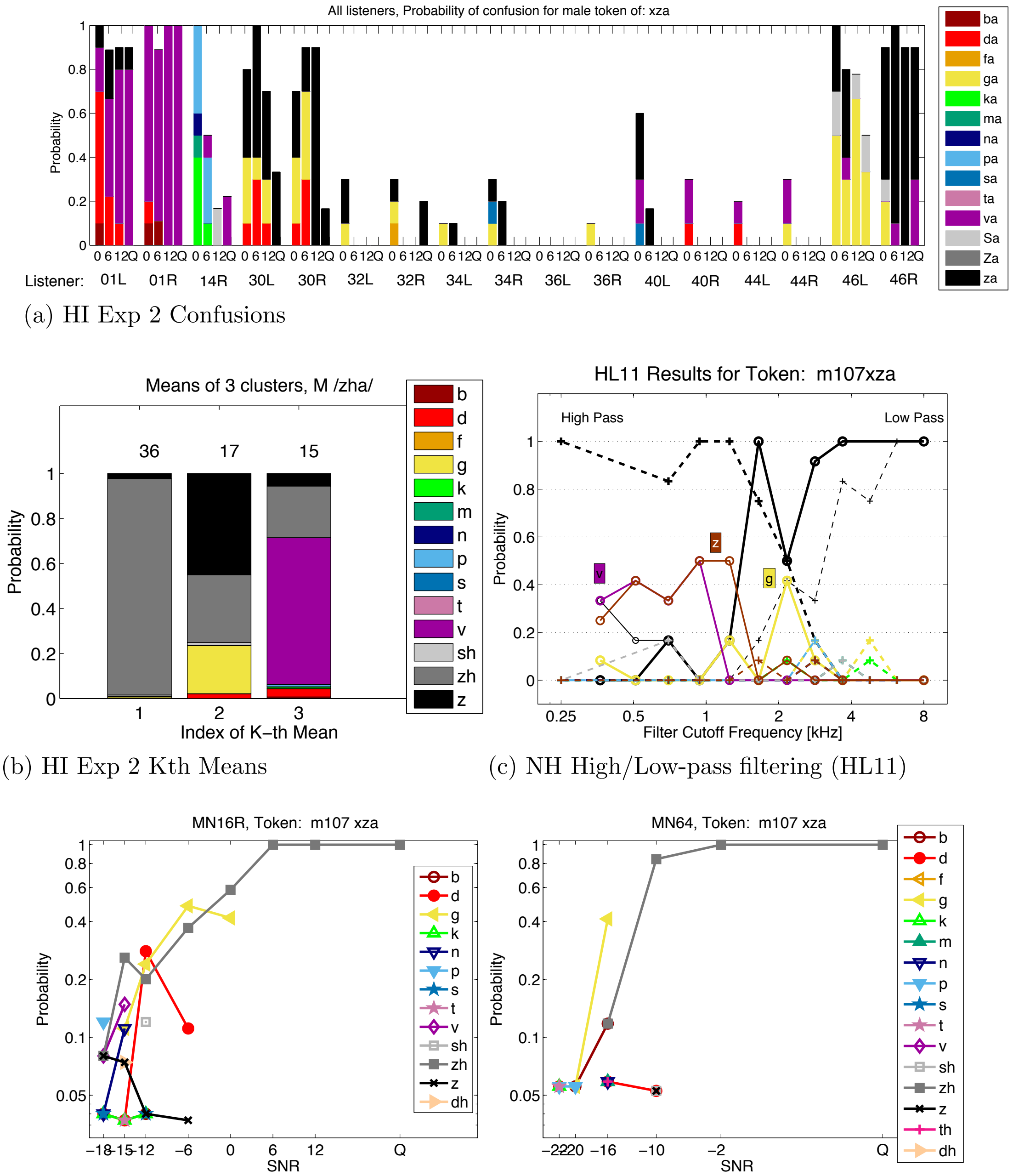
(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

Figure D.48: Each subplot shows the kth cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f105 ʒɑ.



(a) HI Exp 2 Confusions

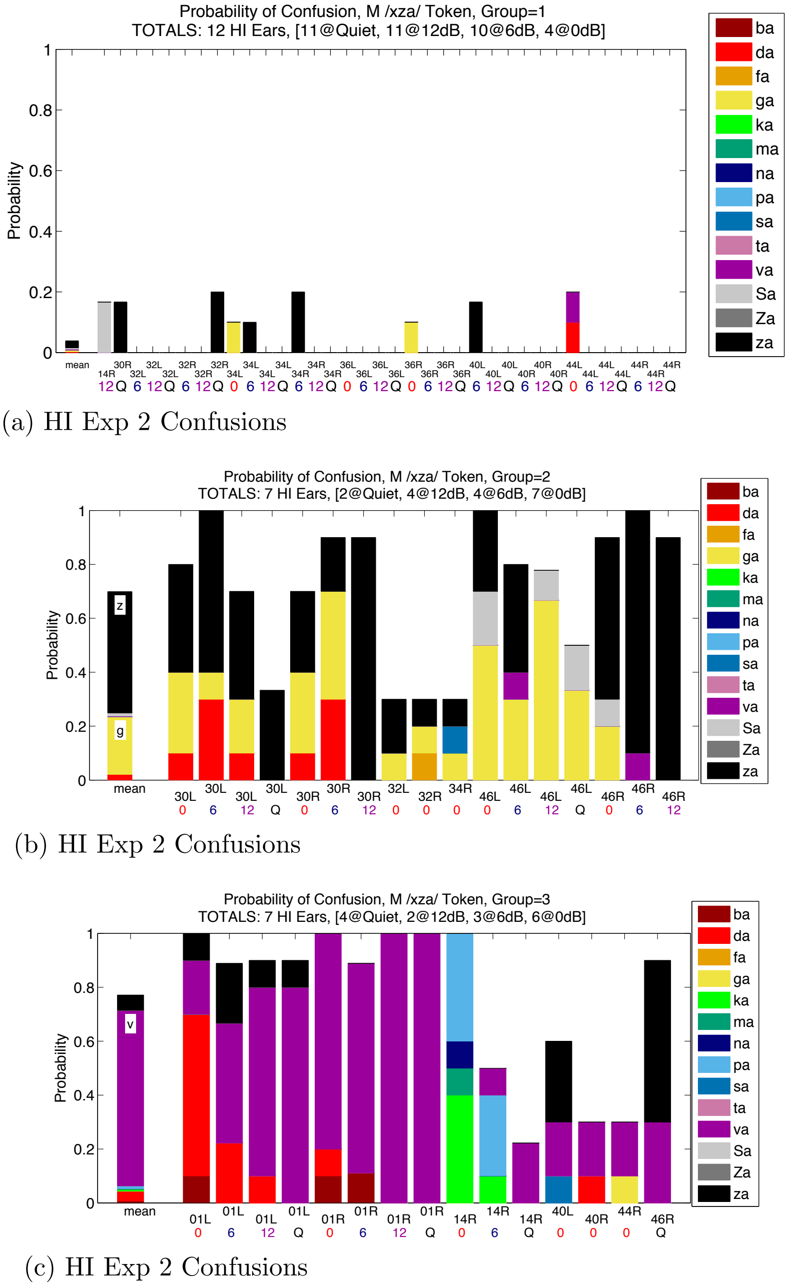
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.49: Overview of data collected for the /ʒɑ/ token from talker m107. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.

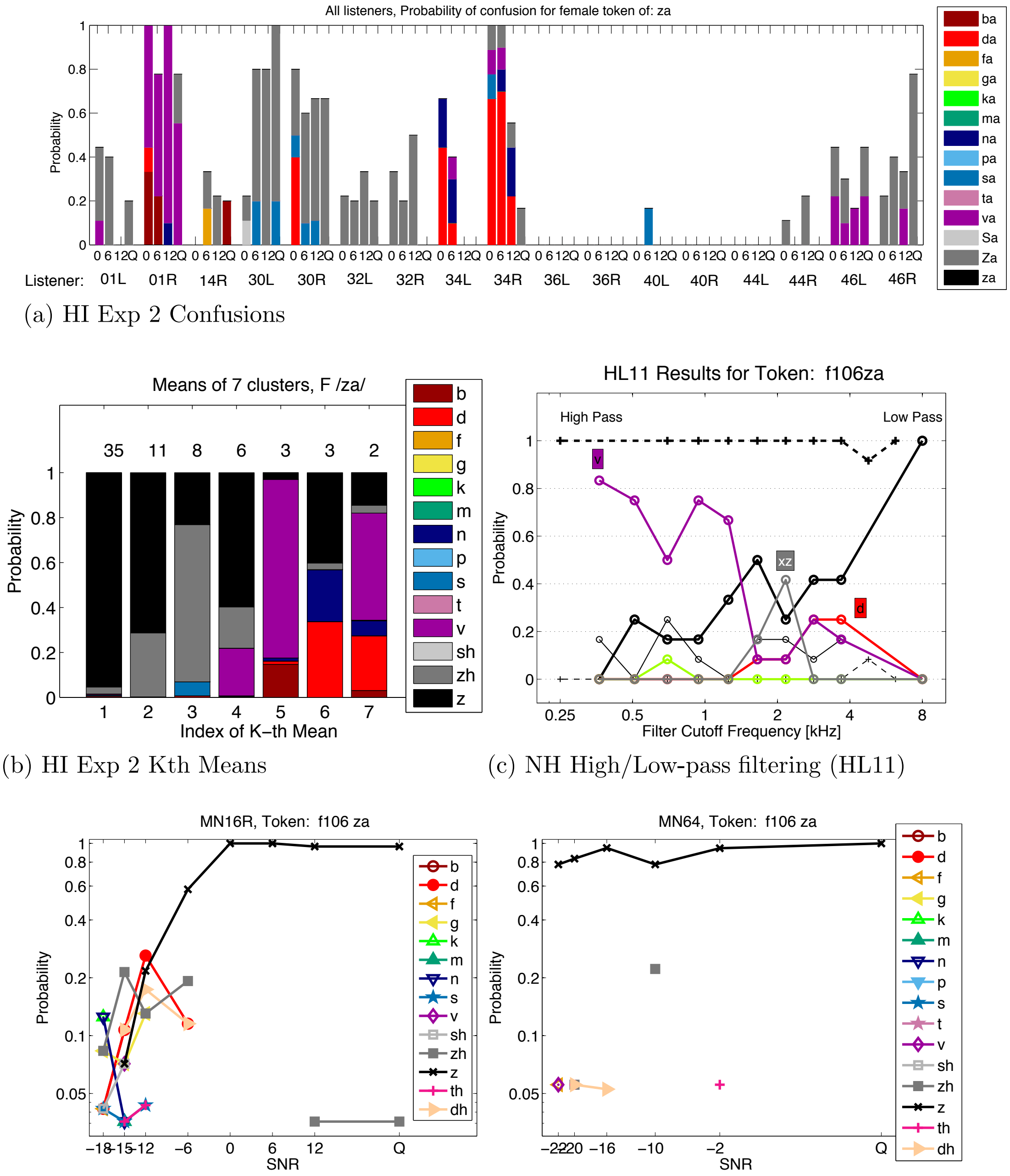


(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

Figure D.50: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m107 ʒɑ.



(a) HI Exp 2 Confusions

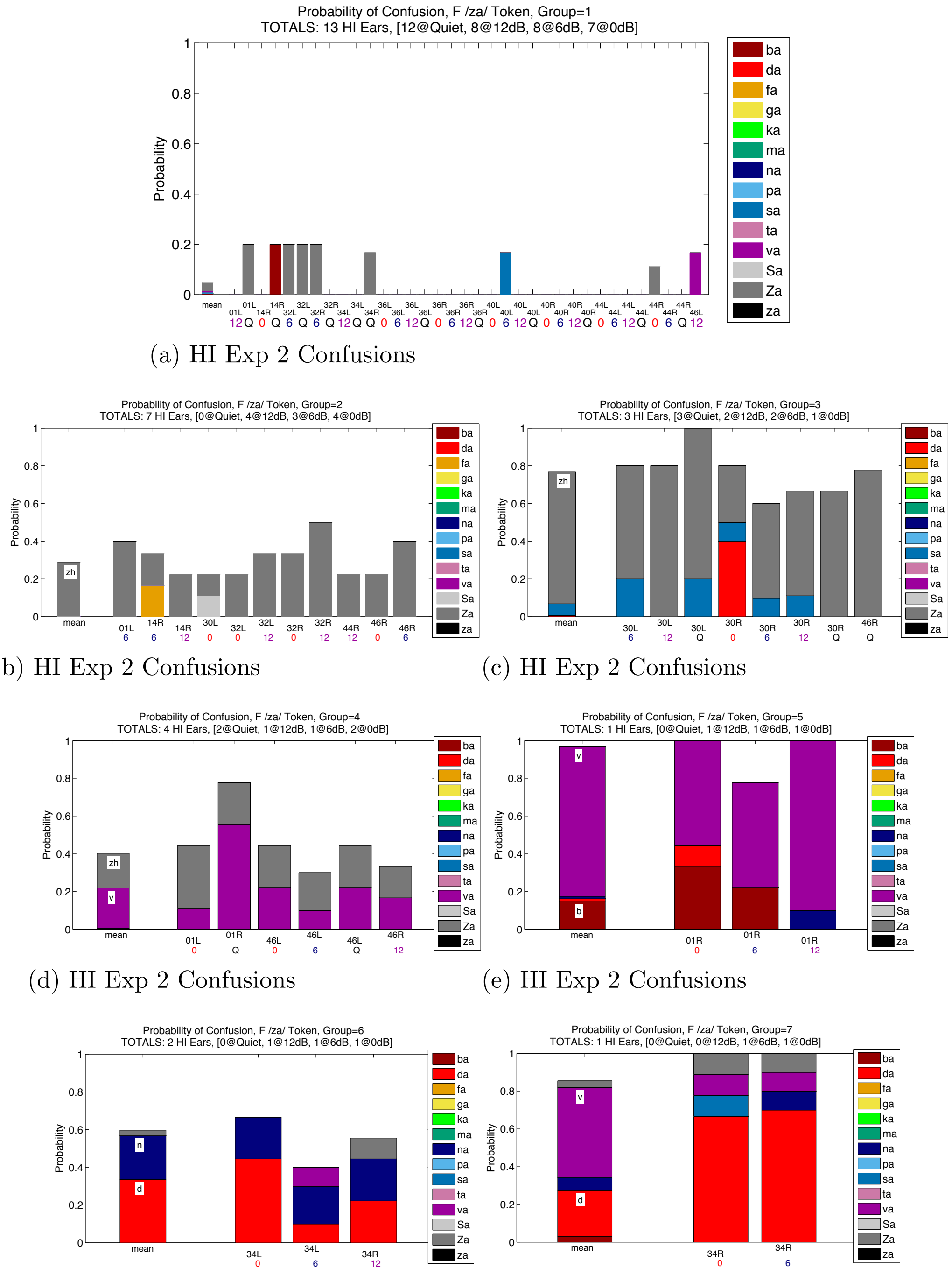
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.51: Overview of data collected for the /zɑ/ token from talker f106. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

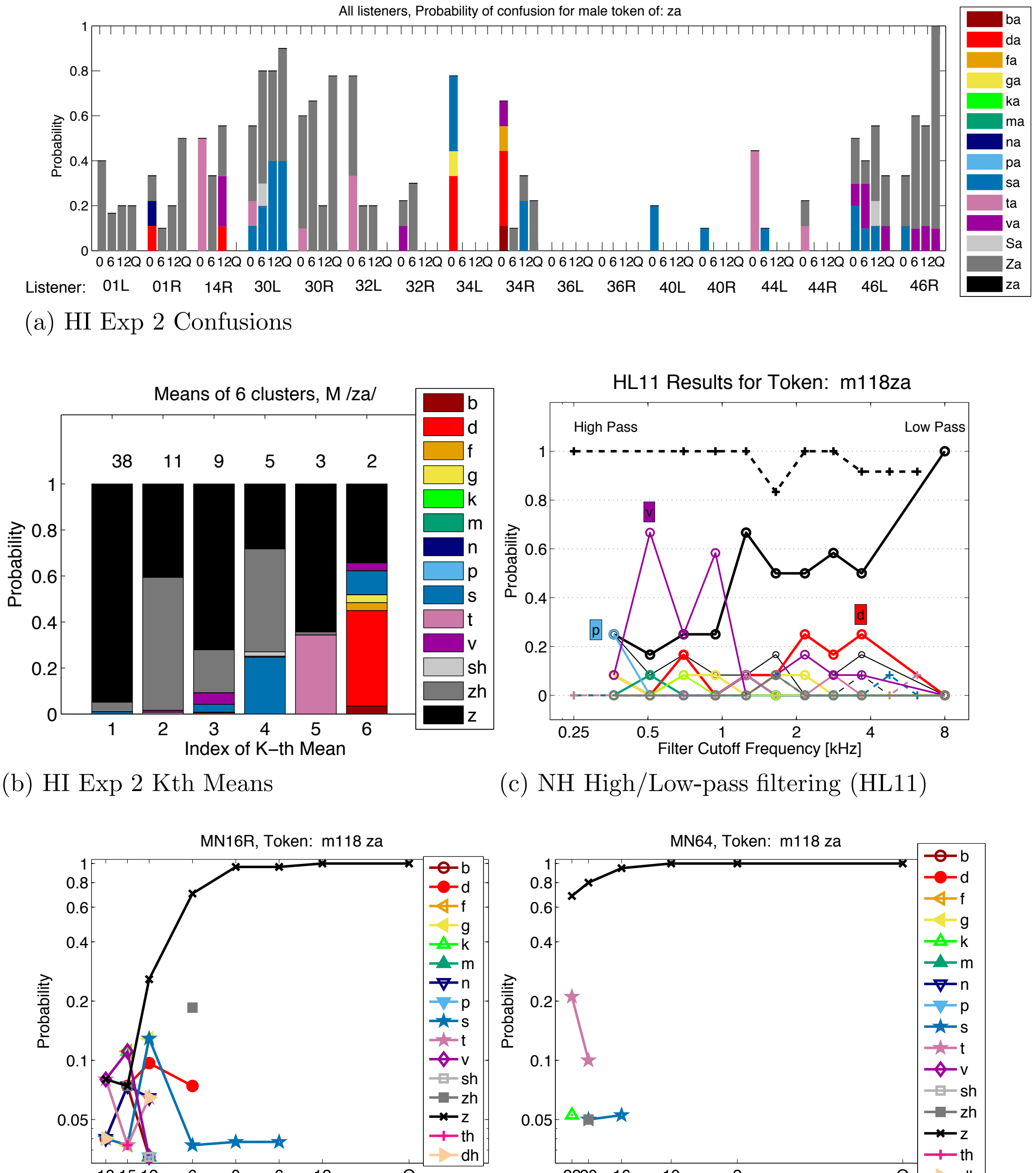
(d) HI Exp 2 Confusions

(e) HI Exp 2 Confusions

(f) HI Exp 2 Confusions

(g) HI Exp 2 Confusions

Figure D.52: Each subplot shows the k^{th} cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for f106 zɑ.



(a) HI Exp 2 Confusions

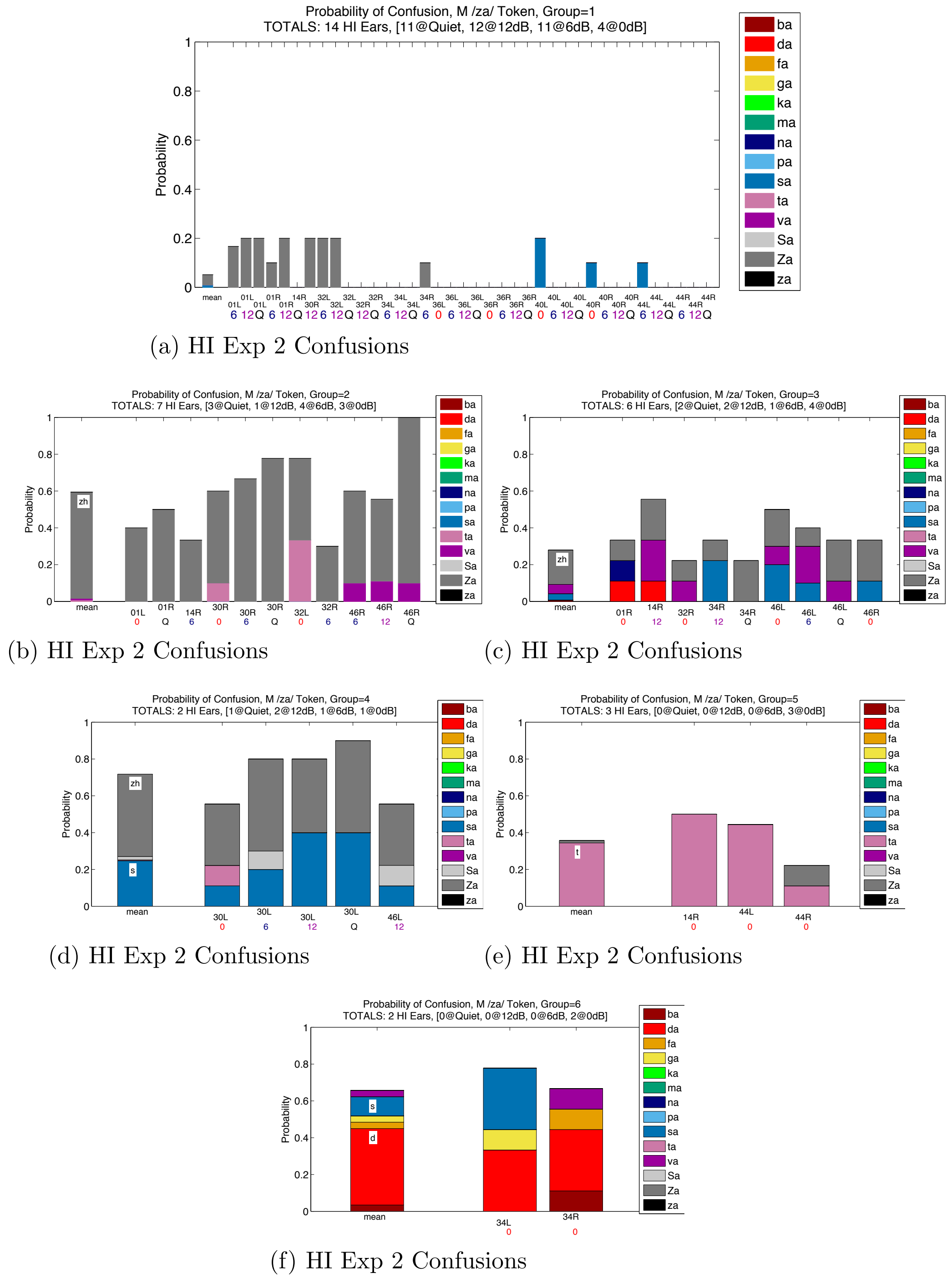
(b) HI Exp 2 Kth Means

(c) NH High/Low-pass filtering (HL11)

(d) NH + White Noise (MN16R)

(e) NH + Speech-Weighted Noise (MN64)

Figure D.53: Overview of data collected for the /zɑ/ token from talker m118. (a) The raw consonant confusion data for all HI ears and SNRs shown as stacked bars. (b) The k-means cluster means. (c) NH high/low-pass filtering results. (d, e) Confusion patterns for the NH noise-masking results, in WN and SWN.



(a) HI Exp 2 Confusions

(b) HI Exp 2 Confusions

(c) HI Exp 2 Confusions

(d) HI Exp 2 Confusions

(e) HI Exp 2 Confusions

(f) HI Exp 2 Confusions

Figure D.54: Each subplot shows the k$^{\text{th}}$ cluster mean as well as the collection of data points in that cluster (k indicated by Group $= k$ in each title). Data for m118 zɑ.

APPENDIX E

DISTRIBUTION OF NH SNR$_{90}$ VALUES IN SWN FOR ALL TOKENS

In this appendix, the SNR$_{90}$ distributions for each consonant are shown in Figs. E.1–E.14. Each figure shows the results for 56 tokens, computed from the results of Phatak's MN64 experiment (Phatak and Allen 2007). This experiment tested perception at -22, -20, -16, -10, -2 dB SNR of speech-weighted noise and in quiet. SNR$_{90}$ values above -2 dB are computed by setting quiet to 15 dB SNR and linearly interpolating, therefore they are approximations. The tokens in the "Bad Token" category never reach ≥90% correct perception. Highly robust tokens, which never show scores <90%, are in the "< -22 dB" category.

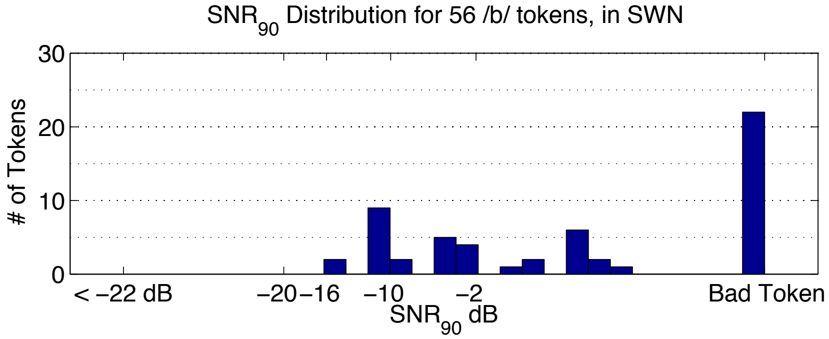


Figure E.1: SNR$_{90}$ distributions for 56 tokens of /b/, in SWN.

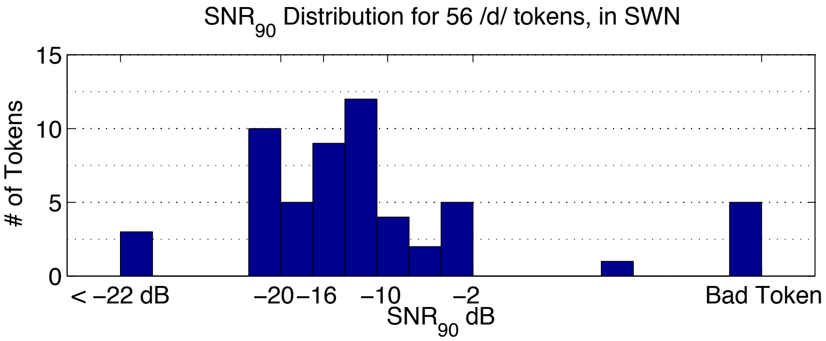


Figure E.2: SNR$_{90}$ distributions for 56 tokens of /d/, in SWN.

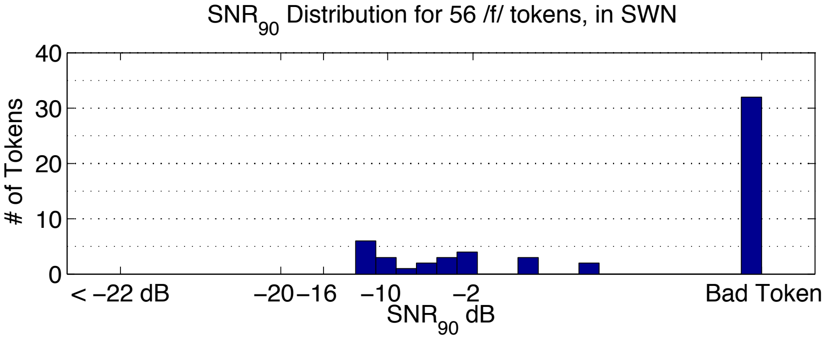


Figure E.3: SNR$_{90}$ distributions for 56 tokens of /f/, in SWN.

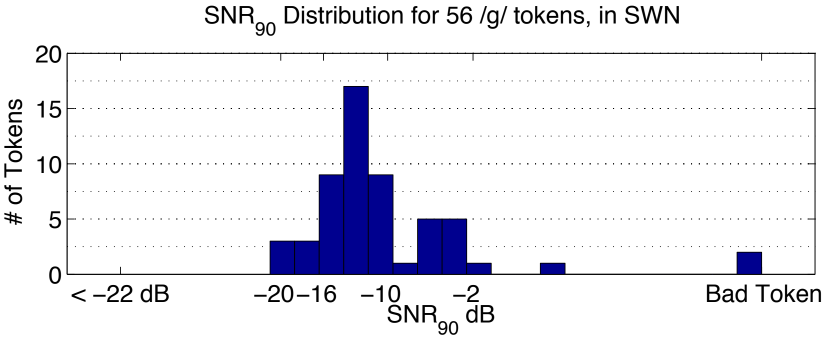


Figure E.4: SNR$_{90}$ distributions for 56 tokens of /g/, in SWN.

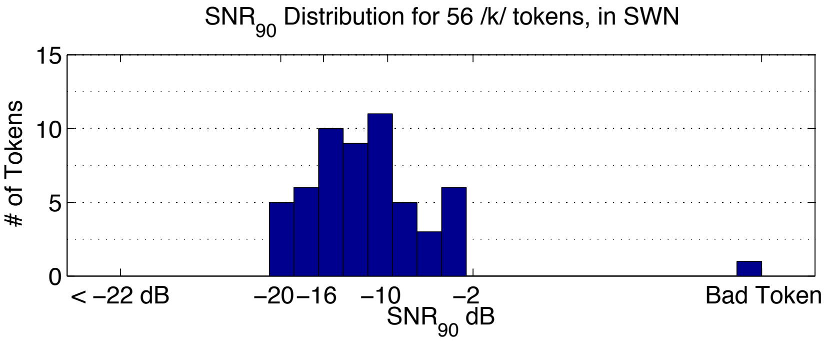


Figure E.5: SNR$_{90}$ distributions for 56 tokens of /k/, in SWN.

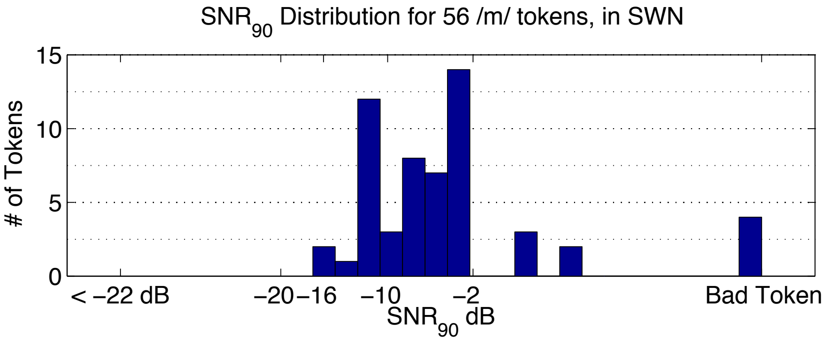


Figure E.6: SNR$_{90}$ distributions for 56 tokens of /m/, in SWN.

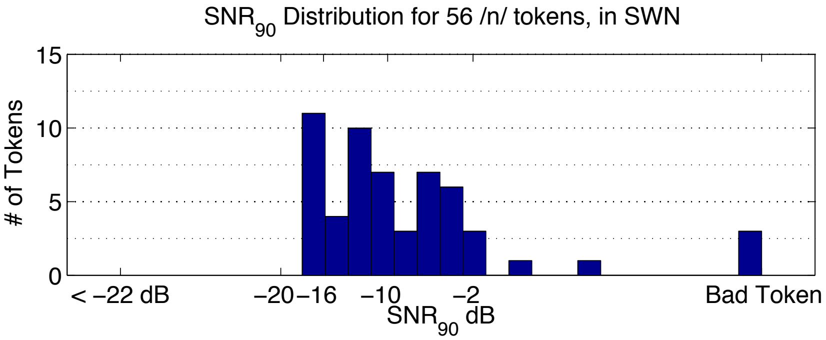


Figure E.7: SNR$_{90}$ distributions for 56 tokens of /n/, in SWN.

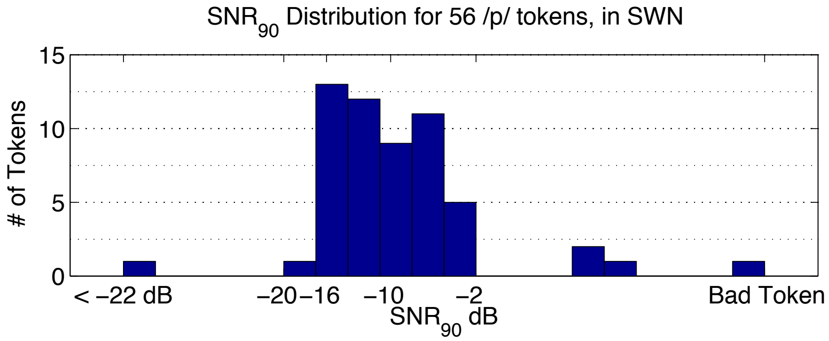


Figure E.8: SNR$_{90}$ distributions for 56 tokens of /p/, in SWN.

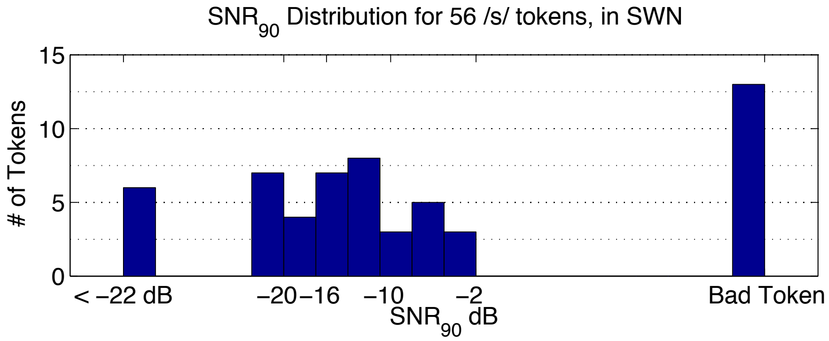


Figure E.9: SNR$_{90}$ distributions for 56 tokens of /s/, in SWN.

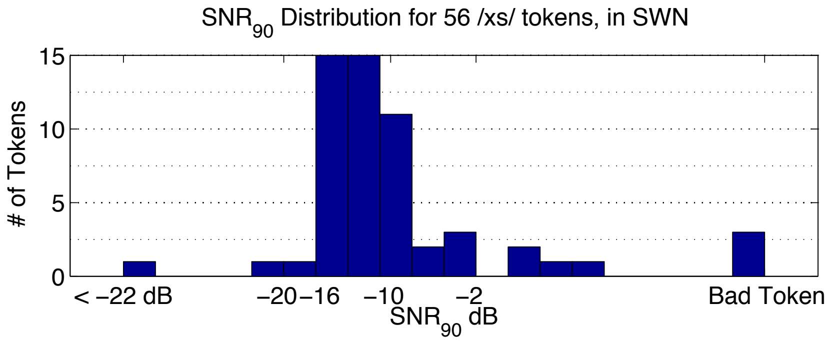


Figure E.10: SNR$_{90}$ distributions for 56 tokens of /ʃ/, in SWN.

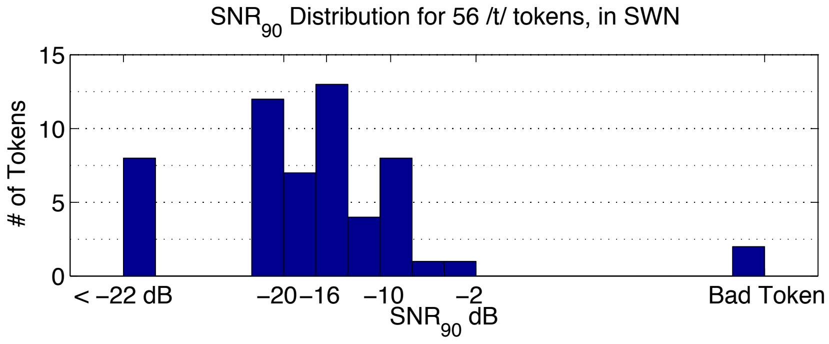


Figure E.11: SNR$_{90}$ distributions for 56 tokens of /t/, in SWN.

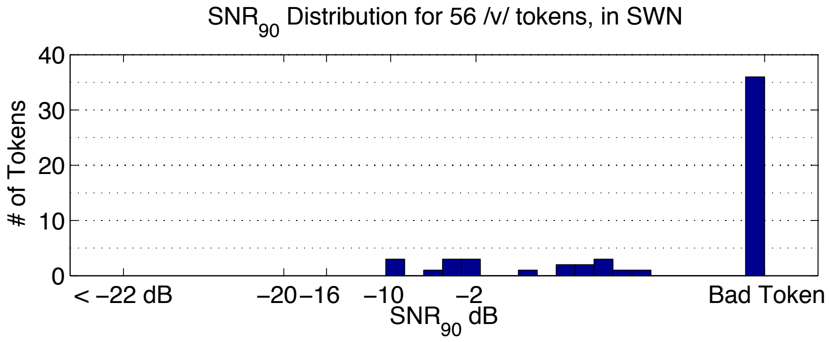


Figure E.12: SNR$_{90}$ distributions for 56 tokens of /v/, in SWN.

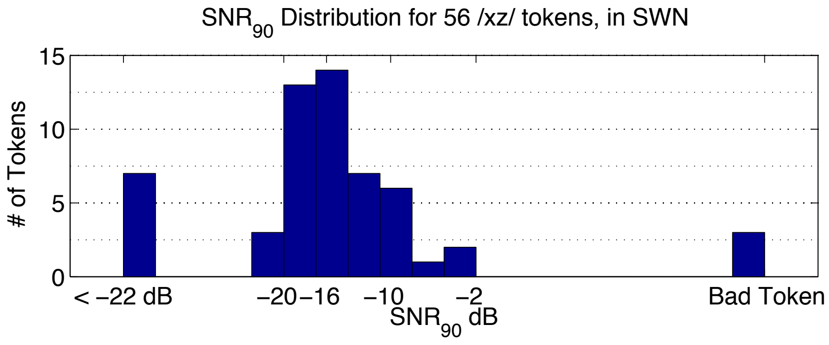


Figure E.13: SNR$_{90}$ distributions for 56 tokens of /ʒ/, in SWN.

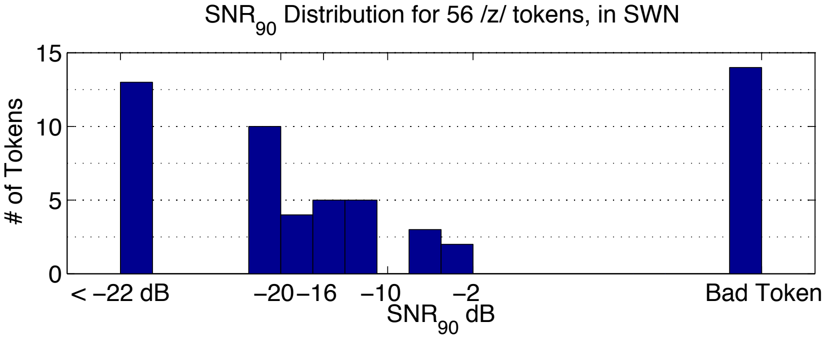


Figure E.14: SNR$_{90}$ distributions for 56 tokens of /z/, in SWN.

REFERENCES

J.B. Allen. Consonant recognition and the articulation index. *The Journal of the Acoustical Society of America*, 117:2212, 2005.

J.B. Allen. How do humans process and recognize speech? *Speech and Audio Processing, IEEE Transactions on*, 2(4):567–577, 1994.

American National Standards Institute. Calculation of the articulation index. *ANSI, S3.5*, 1969.

S.R. Baum and S.E. Blumstein. Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *The Journal of the Acoustical Society of America*, 82:1073, 1987.

S. Behrens and S.E. Blumstein. On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *The Journal of the Acoustical Society of America*, 84:861, 1988.

C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens, and A.S. House. Reduction of speech spectra by analysis-by-synthesis techniques. *The Journal of the Acoustical Society of America*, 33(12):1725–1736, 1961.

R.C. Bilger and M.D. Wang. Consonant confusions in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research*, 19 (4):718, 1976.

S.E. Blumstein, K.N. Stevens, and G.N. Nigro. Property detectors for bursts and transitions in speech perception. *The Journal of the Acoustical Society of America*, 61:1301, 1977.

A. Boothroyd. Auditory perception of speech contrasts by subjects with sensorineural hearing loss. *Journal of Speech and Hearing Research*, 27(1): 134, 1984.

A. Boothroyd and S. Nittrouer. Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, 84:101, 1988.

A.W. Bronkhorst, A.J. Bosman, and G.F. Smoorenburg. A model for context effects in speech recognition. *The Journal of the Acoustical Society of America*, 93:499, 1993.

A.W. Bronkhorst, T. Brand, and K. Wagener. Evaluation of context effects in sentence recognition. *The Journal of the Acoustical Society of America*, 111(6):2874–2886, 2002.

T.Z. Burkle, D. Kewley-Port, L. Humes, and J.H. Lee. Contribution of consonant versus vowel information to sentence intelligibility by normal and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 115(5):2601, 2004.

F.S. Cooper, P.C. Delattre, A.M. Liberman, J.M. Borst, and L.J. Gerstman. Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America*, 24(6):597–606, 1952.

R.M. Cvengros. A verification experiment of the second formant transition feature as a perceptual cue in natural speech. Master's thesis, University of Illinois, Urbana-Champaign, 2011.

P.C. Delattre, A.M. Liberman, and F.S. Cooper. Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, 27(4):769–773, 1955.

H. Dillon. *Hearing Aids*. Thieme Medical Publishers, New York, 2001.

R.A. Dobie. The AMA method of estimation of hearing disability: A validation study. *Ear and Hearing*, 32(6):732, 2011.

M.F. Dorman, M. Studdert-Kennedy, and L.J. Raphael. Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Attention, Perception, & Psychophysics*, 22 (2):109–122, 1977.

W.A. Dreschler. Phonemic confusions in quiet and noise for the hearing-impaired. *International Journal of Audiology*, 25(1):19–28, 1986.

J.R. Dubno and D.D. Dirks. Evaluation of hearing-impaired listeners using a nonsense-syllable test I. test reliability. *Journal of Speech and Hearing Research*, 25(1):135, 1982.

J.R. Dubno, D.D. Dirks, and D.E. Morgan. Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, 76:87, 1984.

D.A. Fabry and D.J. Van Tasell. Masked and filtered simulation of hearing loss: Effects on consonant recognition. *Journal of Speech and Hearing Research*, 29(2):170, 1986.

H. Fletcher and R.H. Galt. The perception of speech and its relation to telephony. *The Journal of the Acoustical Society of America*, 22(2):89, 1950.

H. Fletcher, J.B. Allen, and W.D. Ward. ASA edition of speech and hearing in communication. *The Journal of the Acoustical Society of America*, 100 (2):685–685, 1996.

P. Fousek, F. Grezl, H. Hermansky, and P. Svojanovsky. New nonsense syllables database-analyses and preliminary ASR experiments. In *Eighth Annual Conference of the International Speech Communication Association*, 2004.

N.R. French and J.C. Steinberg. Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19:90, 1947.

S. Furui. On the role of spectral transition for speech perception. *The Journal of the Acoustical Society of America*, 80:1016, 1986.

S. Gordon-Salant. Consonant recognition and confusion patterns among elderly hearing-impaired subjects. *Ear and Hearing*, 8(5):270, 1987.

C. Halpin and S.D. Rauch. Clinical implications of a damaged cochlea: Pure tone thresholds vs information-carrying capacity. *Otolaryngology-Head and Neck Surgery*, 140(4):473–476, 2009.

G. Hamerly and C. Elkan. Learning the k in k-means. *Advances in Neural Information Processing Systems*, 16:281, 2004.

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

K.S. Harris. Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech*, 1(1):1–7, 1958.

M.S. Hedrick and R.N. Ohde. Effect of relative amplitude of frication on perception of place of articulation. *The Journal of the Acoustical Society of America*, 94:2005, 1993.

J.M. Heinz and K.N. Stevens. On the properties of voiceless fricative consonants. *The Journal of the Acoustical Society of America*, 33(5):589–596, 1961.

W. Herd, A. Jongman, and J. Sereno. An acoustic and perceptual analysis of /t/ and /d/ flaps in American English. *Journal of Phonetics*, 38(4): 504–516, 2010.

J.D. Hood and J.P. Poole. Improving the reliability of speech audiometry. *British Journal of Audiology*, 11(4):93–102, 1977.

G.W. Hughes and M. Halle. Spectral properties of fricative consonants. *The Journal of the Acoustical Society of America*, 28(2):303–310, 1956.

L.E. Humes et al. Understanding the speech-understanding problems of the hearing impaired. *Journal of the American Academy of Audiology*, 2(2): 59–69, 1991.

A. Jongman. Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America*, 85:1718, 1989.

A. Jongman, R. Wayland, and S. Wong. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263, 2000.

C.A. Kamm, D.D. Dirks, and T.S. Bell. Speech recognition and the Articulation Index for normal and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 77(1):281–288, 1985.

A. Kapoor and J.B. Allen. Perceptual effects of plosive feature modification. *The Journal of the Acoustical Society of America*, 131:478, 2012.

M.C. Killion and G.I. Gudmundsen. Fitting hearing aids using clinical prefitting speech measures: An evidence-based review. *Journal of the American Academy of Audiology*, 16(7):439–447, 2005.

M.C. Killion and P.A. Niquette. What can the pure-tone audiogram tell us about a patients SNR loss. *Hear J*, 53(3):46–53, 2000.

E.J. Kreul, D.W. Bell, and J.C. Nixon. Factors affecting speech discrimination test difficulty. *Journal of Speech, Language and Hearing Research*, 12 (2):281, 1969.

S.G. Kujawa and M.C. Liberman. Adding insult to injury: Cochlear nerve degeneration after temporary noise-induced hearing loss. *The Journal of Neuroscience*, 29(45):14077–14085, 2009.

K. Kurowski and S.E. Blumstein. Acoustic properties for place of articulation in nasal consonants. *The Journal of the Acoustical Society of America*, 81: 1917, 1987.

D.L. Lawrence and V.W. Byers. Identification of voiceless fricatives by high frequency hearing impaired listeners. *Journal of Speech, Language and Hearing Research*, 12(2):426, 1969.

F. Li. Perceptual cues of consonant sounds and impact of sensorineural hearing loss on speech perception. Ph.D. dissertation, University of Illinois, Urbana-Champaign, 2010.

F. Li and J.B. Allen. Manipulation of consonants in natural speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):496–504, 2011.

F. Li, A. Menon, and J.B. Allen. A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *The Journal of the Acoustical Society of America*, 127:2599, 2010.

F. Li, A. Trevino, A. Menon, and J.B. Allen. A psychoacoustic method for studying the necessary and sufficient perceptual cues of fricative consonants in noise. *The Journal of the Acoustical Society of America*, 132:2663, 2012.

A.M. Liberman, P.C. Delattre, and F.S. Cooper. Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1(3):153–167, 1958.

B.E. Lobdell. Models of human phone transcription in noise based on intelligibility predictors. Ph.D. dissertation, University of Illinois, Urbana-Champaign, 2009.

B.E. Lobdell, J.B. Allen, and M.A. Hasegawa-Johnson. Intelligibility predictors and neural representation of speech. *Speech Communication*, 53(2):185–194, 2011.

K. Maniwa, A. Jongman, and T. Wade. Perception of clear fricatives by normal-hearing and simulated hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 123:1114, 2008.

G.A. Miller and P.E. Nicely. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27: 338, 1955.

M.A. Mines, B.F. Hanson, and J.E. Shoup. Frequency of occurrence of phonemes in conversational English. *Language and Speech*, 21(3):221–241, 1978.

S. Nittrouer. Learning to perceive speech: How fricative perception changes, and how it stays the same. *The Journal of the Acoustical Society of America*, 112:711, 2002.

E. Owens. Consonant errors and remediation in sensorineural hearing loss. *Journal of Speech and Hearing Disorders*, 43(3):331, 1978.

R.D. Patterson, I. Nimmo-Smith, D.L. Weber, and R. Milroy. The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. *The Journal of the Acoustical Society of America*, 72:1788, 1982.

S.A. Phatak. Phone confusion analysis and its applications. Ph.D. dissertation, University of Illinois, Urbana-Champaign, 2007.

S.A. Phatak and J.B. Allen. Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America*, 121(4): 2312–2326, 2007.

S.A. Phatak, A. Lovitt, and J.B. Allen. Consonant confusions in white noise. *The Journal of the Acoustical Society of America*, 124:1220, 2008.

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

R.E. Remez, P.E. Rubin, D.B. Pisoni, and T.D. Carrell. Speech perception without traditional speech cues. *Science*, 212(4497):947–949, 1981.

R.J. Roeser, M. Valente, and H. Hosford-Dunn. *Audiology: Diagnosis.* Thieme Medical Publishers, 2007.

C.H. Shadle and S.J. Mair. Quantifying spectral characteristics of fricatives. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1521–1524. IEEE, 1996.

R.V. Shannon, F.G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, et al. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

R. Singh and J.B. Allen. The influence of stop consonants perceptual features on the articulation index model. *The Journal of the Acoustical Society of America*, 131(4):3051–3068, 2012.

M.W. Skinner. Speech intelligibility in noise-induced hearing loss: Effects of high-frequency compensation. Ph.D. dissertation, Washington University School of Medicine, 1976.

M.W. Skinner and J.D. Miller. Amplification bandwidth and intelligibility of speech in quiet and noise for listeners with sensorineural hearing loss. *International Journal of Audiology*, 22(3):253–279, 1983.

G.F. Smoorenburg. Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram. *The Journal of the Acoustical Society of America*, 91:421, 1992.

S.D. Soli. Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America*, 70:976, 1981.

K.N. Stevens and S.E. Blumstein. Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, 64: 1358, 1978.

K.N. Stevens, S.E. Blumstein, L. Glicksman, M. Burton, and K. Kurowski. Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *The Journal of the Acoustical Society of America*, 91:2979, 1992.

R. Sweetow and C.V. Palmer. Efficacy of individual auditory training in adults: A systematic review of the evidence. *Journal of the American Academy of Audiology*, 16(7):494–504, 2005.

B. Taylor. Predicting real-world hearing aid benefit with speech audiometry: An evidence-based review. Ph.D. dissertation, Central Michigan University, Mount Pleasant, 2006.

A. Trevino and J.B. Allen. Individual variability of hearing-impaired consonant perception. In *Seminars in Hearing*, volume 34, page 74. Thieme Medical Publishers, 2013a.

A. Trevino and J.B. Allen. Within-consonant perceptual differences in the hearing impaired ear. *The Journal of the Acoustical Society of America*, 134:607, 2013b.

D.F. Vysochanskij and Y.I. Petunin. Justification of the 3σ rule for unimodal distributions. *Theory of Probability and Mathematical Statistics*, 21:25–36, 1980.

B.E. Walden and A.A. Montgomery. Dimensions of consonant perception in normal and hearing-impaired listeners. *Journal of Speech and Hearing Research*, 18(3):444, 1975.

B.E. Walden, S.A. Erdman, A.A. Montgomery, D.M. Schwartz, and R.A. Prosek. Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech, Language and Hearing Research*, 24(2):207, 1981.

B.E. Walden, L.L. Holum-Hardegen, J.M. Crowley, D.M. Schwartz, and D.L. Williams. Test of the assumptions underlying comparative hearing aid evaluations. *Journal of Speech and Hearing Disorders*, 48(3):264, 1983.

M.D. Wang and R.C. Bilger. Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, 54 (5):1248–1266, 1973.

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. *Journal of Speech and Hearing Research*, 21(1):5, 1978.

D.H. Whalen. Perception of the English /s/–/ʃ/ distinction relies on fricative noises and transitions, not on brief spectral slices. *The Journal of the Acoustical Society of America*, 90:1776, 1991.

F.G. Zeng and C.W. Turner. Recognition of voiceless fricatives by normal and hearing-impaired subjects. *Journal of Speech, Language and Hearing Research*, 33(3):440, 1990.

P.M. Zurek and L.A. Delhorne. Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. *The Journal of the Acoustical Society of America*, 82(5):1548–1559, 1987.