EFFECTS OF NAL-R AMPLIFICATION ON CONSONANT SPEECH PERCEPTION IN HEARING-IMPAIRED LISTENERS

BY

CHRISTOPH SCHEIDIGER

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Adviser:

Associate Professor Jont B. Allen

Abstract

This thesis investigates speech perception in *hearing impaired* (*HI*) subjects. Psychoacoustic experiments in different conditions were undertaken. In particular, two *consonant vowel* (*CV*) identification experiments in masking noise were conducted at various *signal-to-noise ratios* (*SNRs*) with 16 HI ears. In one of the experiments, the CVs were presented with a uniform gain; in the other experiment, a spectral compensation (i.e. NAL–R) for the individual hearing loss was provided. In both gain conditions, the subjects were instructed to adjust the presentation level to their *most comfortable loudness* (*MCL*), which is contrary to the common approach of adjusting the presentation level depending on the *pure tone thresholds* (*PTTs*) and the *long-term average speech spectrum* (*LTASS*) (Zurek and Delhorne (1987), Posner and Ventry (1977)). The data demonstrated that the MCL approach led to consistent responses in all subjects. Based on these results, a more rigorous definition of audibility based on entropy and the Miller and Nicely (1955) confusion groups is proposed. Furthermore, the effectiveness of NAL–R for CV perception was investigated by comparing the confusion matrices of the two experiments. In general, the error and entropy decreased with NAL–R. The average error decreased from 20.1% ($\sigma = 3.7$) to 16.3% ($\sigma = 2.8$). It was also shown that, with the help of NAL–R, the tested ears became more consistent in their responses for a given token. However, for 15.1% of the *token*[1]*-ear pairs* (*TEPs*), the entropy and error increased with NAL–R. It was shown that these 15.1% of the TEPs contained all ears and a large variety of tokens. A method based on the *Hellinger Distance* (*HD*) was introduced

[1] In this document a token is defined as a recorded sound (i.e. CV). One consonant (e.g. /p/) can have many tokens.

that enabled comparison of rows of confusion matrices and to calculate distances between responses. With this method, the highly individual problems of the 15.1% of the TEPs were further investigated and compared to the results obtained in normal hearing subjects. In conclusion, it is argued that speech testing — using the proposed methods and experiments as described in this thesis — can deliver valuable and reliable information about individual hearing loss that goes beyond what can be achieved using pure tone thresholds.

Acknowledgments

I am thankful for Professor Dr. Jont B. Allen's guidance as a mentor and advisor during my time as a research assistant at the University of Illinois at Urbana-Champaign (UIUC). I have learned a lot during the last two years and was blessed to get to work with someone with as much experience as Dr. Allen. My professional and personal life will always be influenced by what I learned during our countless personal meetings.

I had the remarkable opportunity to work with smart, nice and helpful colleagues in the Human Speech Recognition (HSR) group at UIUC. Roger, Andrea, Cliston, Sarah, Noori, Corinne, and Chloe are wonderful and bright people with whom I enjoyed discussions during our work at the HSR lab in the Beckman Institute.

Also invaluable were the discussions with my close friend at UIUC, Marcel Schuck, and my roommate, Johannes Traa; they provided me with an outsider perspective on my work. I will always cherish the memories of the discussion about work and life we had over food.

Even though the distance made contact more difficult, the biggest emotional support came from my family. Knowing of their unconditional love and support helped me to manage difficult times. I am very grateful for Erin, who has been supportive in so many ways and has also been a great role model as a committed scientist.

Table of Contents

List of Abbreviations

3DDS	Three–dimensional deep search. Psychoacoustic method to identify consonant cues.
AI	Articulation Index, a tool to predict speech phoneme identification based on the SNR in the critical bands.
ANOVA	Analysis of variance
CLP	Consonant-loss profile
CM	Confusion-matrix
CV	Consonant-vowel
CVC	Consonant-vowel-consonant
DR	Direct realism, theory of speech perception based on gestural transmission of speech
FG	Flat-gain, gain independent of frequency
HD	Hellinger distance
HL	Hearing loss
HSR	Human speech recognition
HTL	Hearing threshold level
IT	Information transmitted
LDL	Loudness discomfort level
LTASS	Long term average speech spectrum
NAL-R	The revised prescriptive procedure to fit a hearing aid based on PTAs from the National Acoustics Laboratory Australia (NAL)

NH	Normal hearing
MCL	Most comfortable loudness
MT	Motor theory, theory of speech perception based on gestural transmission of speech
PTA	Pure tone average
PTT	Pure tone threshold
RD	Reading disabled
REG	Real-ear gain
RG	Response group
S/B	Speech to babble ratio
SNHL	Sensorineural hearing loss. Most SNHLs are due to abnormalities in the hair cells in the cochlea.
SNR	Signal-to-noise ratio, usually in dB, defined as $10 \cdot log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$, where P is the power
SNR90	Signal-to-noise ratio at which NH listeners perceive an utterance with 90% accuracy
SPL	Sound pressure level, measures the sound pressure in dB with a reference $p_{ref} = 20\ \mu$P
SRT	Speech recognition threshold, SNR at which 50% of the words of a sentence are recognized correctly
SWN	Speech weighted noise
TEP	Token-ear pair
WDRC	Wide dynamic range compression

Chapter 1

Introduction

The goal of this thesis is to analyze speech perception in *hearing impaired* (HI) ears as a function of the speech token and the *signal-to-noise ratio* (SNR). The approach for the *human speech recognition* (HSR) group at the University of Illinois at Urbana-Champaign is to look at *consonant-vowel* (*CV*) recognition tasks in the HI, relative to *normal hearing* (NH) subjects. CVs are chosen to minimize the influence of higher-order processing in the auditory pathway, thus limiting the influence of cognitive abilities (e.g. memory, semantics) (Miller et al. (1951)).

For this work, 16 HI ears were tested under two conditions; *flat-gain* (FG) and spectral correction, based on the prescriptive procedure NAL-R. In both conditions the subjects listened to 28 CV tokens[1] (2 tokens per consonant) at 4 different SNRs, with the presentation level at the subject's *most comfortable loudness* (MCL). All eight subjects (16 HI ears) had mild-to-moderate hearing loss, as described in Chapter 3 (Methods). Less extensive data on 48 HI ears under the two above-described conditions was also collected, to allow for test-retest analysis. The tokens used for these experiments have been previously used in NH experiments and have been analyzed using the *three–dimensional deep search* (3DDS) method (see Table 1.1). In NH experiments these tokens had zero error at the tested SNRs (Singh and Allen (2012)).

A secondary aim is to address the definition and verification of audibility in speech perception experiments. Lastly, we will show that by working at the token level, speech as a test for hearing loss and hearing aid evaluation can deliver more detailed insights than the commonly used pure tones, in contrast to previous conclusions (Walden et al. (1983), Zurek and Delhorne (1987)).

[1] In this document a token is defined as a recorded sound (i.e. CV). One consonant (e.g. /p/) can have many tokens.

Table 1.1: Experiments performed by the HSR group with both normal-hearing and hearing-impaired subjects in chronological order.

Year	Experiment	Students	Details	Publications
2004	MN04(MN64)	Phatak & Lovitt	Repeat Miller Nicely (SWN)	Phatak & Allen (2007)
2005	MN05WN	Phatak & Lovitt	Replicate MN04 (WN)	Phatak et al (2008)
2005	MN05SWN	Phatak & Lovitt	MN64 more subjects(SWN)	
2005	HIMCL05	Yoon & Phatak	CVs in 10 HI ears @ MCL in WN	Phatak et al (2009)
2006	HINALR05	Yoon	CVs in 10 HI ears with NALR@MCL	Yoon et al. (2011)
2006	Verification	Regnier	/ta/	Regnier & Allen (2008)
2006	CV06SWN	Phatak	9C+8V SWN	
2006	CV06WN	Regnier	9C+8V WN	
2007	CV06	Pan	9 Vowels	2 unpublished MSs
2007	HL07	Li	Hi/Lo pass	Li & Allen (2009)
2008	TR07/08	Li	Furui86/3 vowels	Allen & Li (2009)
2009	3DDS	Li	Plosives & Fricatives	Li et al. (2010); Li et al. (2011); Li et al.(2012)
2009	Verification	Abhinauv	Modify + Remove primary burst	JASA 2012
2009	Verification	Cvengros	Modify burst & F2 transition	JASA, Rejected
2009	MN64 high error	Singh	High error sounds in PA07	JASA, April 2012
2010	HIMCL10-I/-III	Woojae Han	46 HI ears with N=4/Consonant	EH rejected
2010	HI10NALR-II/-IV	Woojae Han	17 HI ears with N=20/Consonant	
2011	HL11	Trevino	High/Low filter CVs of HI10	JASA 2013

1.1 Consonant Perception in Hearing Impaired Subjects

One of the primary goals of hearing aids is to help HI subjects understand speech. Many studies, however, show that hearing aid wearers are often unsatisfied with their devices (Dillon (2012)). One of the reasons for this mismatch could be the way hearing aids are adjusted to the individual wearer. For many years, starting with Knudsen and Jones (1935), *pure tone thresholds* (PTTs) have been used for fitting hearing aids, while speech tests have been found to be ineffective. Speech has therefore not been used as a diagnostic tool (Dobie (2011), Walden et al. (1983)). This may mainly be due to the large variability in speech, and the type of analysis applied to speech perception tests. When the individual characteristics of the speech tokens are not known, the use of speech has proven to be limited. Pure tones remain a popular fitting procedure, even though it has been shown that people with similar pure tone thresholds may differ significantly in their ability to perceive speech (Halpin and Rauch (2009), Kamm et al. (1985), Killion and Niquette (2000), Roeser and Valente (2007), Skinner (1976), Skinner and Miller (1983), Smoorenburg (1992), Walden and Montgomery (1975)).

An often used measure for the ability of speech perception is the *speech reception threshold* (SRT), which is based on speech with context. However, context is another pitfall, since it involves higher-order auditory processing, such that cognitive abilities may play a role (Allen (2005a)).

The present work avoids these drawbacks and seeks to gain information about individual *sensorineural hearing losses* (SNHL) by looking at the ability of the subjects to discriminate individual well-studied CV tokens.

The above mentioned variability in speech has often been pointed out in consonant cue literature (Baum and Blumstein (1987), Dorman et al. (1977), Herd et al. (2010), Jongman et al. (2000), Kurowski and Blumstein (1987), Li (2010)). For NH subjects, this variability does not seem to matter, since acoustically different tokens of the same consonants can be easily identified, even in noisy conditions (Singh and Allen (2012)). However, this same variability strongly affects HI listeners (Trevino and Allen (2013b)). We shall show that this variability (i.e., token variability) has an effect on the benefit that an individual hearing impaired ear gets from a prescribed gain.

In order to investigate the hearing loss with CV token recognition, a solid understanding of the individual tokens used in the test is required. This understanding can be gained by testing NH subjects at the token level.

Consonant cue research with NH subjects has a long history, which is briefly reviewed below. A more detailed review may be found in Chapter 2.

1.2 Consonant Cues

The above mentioned variability is one of the main issues when studying consonant cues. With the invention of the speech vocoder at Bell Labs in the 1930s, most early researchers used synthetic speech for their experiments, to *avoid* the variability problem. The first work was done at the Haskin Labs during the 1950s (Cooper et al. (1952), Delattre et al. (1955), Liberman et al. (1954), Liberman (1957), Bell et al. (1961)). The clear disadvantage of using synthetic speech is that it only contains what is synthesized; i.e. in order to find perceptual cues, one first needs to encode the signal cues (assuming they are the right cues) and then the importance of the assumed cues needs to be verified empirically. This represents a principal limitation of the approach. During the initial years, the synthetic speech was so poorly produced that the subjects needed to be trained before participating (Delattre et al. (1955)). Such training may lead the subjects to listen to cues they normally would not listen to. Later studies used natural speech (Baum and Blumstein (1987), Behrens and Blumstein (1988), Jongman et al. (2000)); however, their findings were based again on inspection of the signal and not on human perception. Therefore, the question of relevant perceptual cues was not addressed.

With this history in mind (Allen (2005a)), the HSR group developed the *3-Dimensional Deep Search* (3DDS) method (Li (2010)). With this method, the perceptual cue region for a specific token is identified by analyzing the results of three different psychoacoustic experiments: A consonant vowel noise masking experiment (similar to Miller and Nicely (1955)), a high and low-pass experiment (similar to French and Steinberg (1947)) and a time truncation experiment (similar to Furui (1986)). These three experiments, together with the *articulation index* AI-gram (Lobdell et al. (2011), Lobdell (2009), Régnier and Allen (2008)), a spectrogram-like representation of speech that is built to represent only the audible parts of the signal, allowed for the identification of the necessary and sufficient cues in CVs. The results of the 3DDS method for plosives[2] are published in Li et al. (2010) and for fricatives[3] in Li et al. (2012).

1.3 Roadmap of Thesis

In Chapter 2 the literature for speech perception research is reviewed. The review is split up in three sections. Section 2.1 reviews the literature for consonant perception research in normal hearing subjects, which explains how the results of early research led to different theories of speech perception. A subsection explains how some researchers think speech is transmitted gesturally, another describes the work of researches arguing for the importance of acoustic cues in the signal. In Section 2.2, research about consonant perception in HI subjects is reviewed. Section 2.3 provides an overview of the history of hearing aid fitting.

Next, Chapter 3 gives detailed information about the methods of the experiments analyzed in this thesis. It also explains the novel methods that are used to analyze the confusions matrices (CM) generated by the experiments. These include metric space methods based on the Hellinger distance.

Chapter 4 presents the results found by the methods described in Chapter 3. It addresses how audibility can be defined in a more rigorous way, based on a token entropy vs. token error classification scheme. It analyzes the impact of spectral

[2] Plosives, also known as stop consonants, are consonants where the air flow is completely blocked by the tongue (e.g. /t/) or by the lips (e.g. /b/) for awhile, such that the air flow ceases. For nasals (i.e. /m/ and /n/) the vocal tract is also blocked but the airflow continues through the nasal cavity.

[3] For fricatives the vocal tract is only blocked partially, the narrow constriction causes frication, such as in /s/.

compensation on token perception.

Chapter 5 summarizes the main results and discusses their significance. It discusses the five important points this thesis makes: (i) Testing the subjects at their *most comfortable loudness* (MCL) provided an audible level. (ii) Audibility is ill defined in the existing literature and not adequately formulated for reliable consonant vowel (CV) recognition experiments. The novel definition of audibility based on token entropy is shown to fix these short-comings. (iii) NALR lowers the token entropy which, according to our new definition of audibility, proves that NAL-R at MCL makes the sounds more audible. (iv) For more than 1/7 of the token-SNR conditions (15%), NAL-R makes CV recognition worse. Those occasions are distributed over all tested tokens and ears, which means they represent very specific problems of individual ears. (v) The *Hellinger distance* (HD) has proven to be a powerful measure for the analysis of *confusion matrices* (CMs) as it allows one to measure distances between token responses. Based on the HD, CM data can be clustered and visualized.

Appendix A discusses the *probabilistic latent semantic indexing* (PLSI) algorithm and its use in analyzing CM data. The PLSI algorithm can be viewed as a matrix factorization similar to *singular value decomposition* (SVD). The algorithm allows one to factorize (by means of a iterative constraint optimization) a CM into three probabilistic matrices that reveal information about the confusion groups and the listeners' distributions to these groups. This section investigates the possibility that the PLSI is a better choice for clustering CMs than the k-means algorithm. Appendix A provides some arguments along with a direct comparison of the two algorithms.

Chapter 2

Background

The following literature review was in part inspired by the work of Cvengros and Allen (2011) and Humes et al. (1990). It is split in three sections: consonant perception in NH listeners, speech perception in HI listeners and prescriptive procedures.

2.1 Consonant Perception in Normal Hearing

There has been a long history of research on consonant cues (Wright (2004)). The earliest may be tracked back to the experiments at the Haskins Laboratory in the 1950s. The field started to look for time-frequency cues in the acoustical signal that are perceptually relevant. The Haskins Laboratory developed a speech synthesizer called the *Pattern Playback*, which played back drawn spectrograms. Several classic studies that had a major influence on the field were conducted with this method. Synthetic speech was applied to stop consonants (Blumstein et al. (1977)), fricatives (Heinz and Stevens (1961); Hughes and Halle (1956)), nasals (Liberman (1957)). Remez et al. (1981) went further and used sine wave speech in order to investigate the ability of humans to perceive speech information in non-speech signals. However, synthetic speech has often been criticized. With the advancement of digital signal processing, modified natural speech was eventually used more often (Miller and Nicely (1955), Blumstein and Stevens (1979), Hazan and Simpson (1998), Li et al. (2010)).

The variability of the acoustical cues found at Haskins Laboratory led some to think the invariability of speech transmission has to be found somewhere other than in the signal. Theories, such as the *Motor Theory* (MT) and *Direct Realism* (DR), argue for a gestural perception of speech, abandoning the idea of invariant acoustical cues in the signal. However, the search for acoustic cues has continued. The following review is therefore split into two sections; one looks at the research that led to the theories of gestural speech perception, the other describes studies that have been done to find the perceptual relevant cues in the acoustic signal.

It all started with the Haskins experiments, which were intended as acoustical cue research but led to the popular gestural speech theories. They are therefore the foundation of both branches. Liberman et al. (1954) analyzed spectrograms of naturally spoken CVs and picked potential speech cue candidates by hand. They noticed that speech is composed of smaller building blocks such as narrow band *bursts*, resonances (*formants*) and *transitions* from the bursts to the formants (Delattre et al. (1955), Liberman et al. (1957) and Liberman (1957)). In order to isolate the cues, they then used the above described Pattern Playback method to conduct perceptual experiments. They first started with the stop consonant (e.g. /b/, /d/, /g/) burst (Cooper et al. (1952), Liberman (1996)). The burst feature was tested by synthesizing CVs with different burst frequencies and durations. With the methods they used the synthesized speech was poor: Liberman himself admitted the stimuli were far from easily understandable speech ((Liberman, 1996, p. 12)). In their experiments they found that the burst duration and frequency were well correlated to the consonant response of the subjects. The results are summarized in Figure 2.1.

Furthermore, Cooper et al. found that the context vowel has an effect on how the sound is perceived. As seen in Figure 2.1, a burst with a center frequency of 1.5 [kHz] is perceived as a /g/ followed by a /ɑ/ and as a /b/ if followed by a /o/. It can be concluded that high-frequency bursts make listeners perceive the consonant /d/, mid-frequency bursts cause listeners to hear /g/, whereas low-frequency bursts cue a /b/ (compare 3DDS results, Figure 2.5a and Figure 2.5b on p. 19).

After researching the burst features, the researchers at Haskins Laboratory considered a feature called the *F2 transition* (Liberman et al. (1954)), which is the transitional region between the burst and the second formant (F2) of the context vowel. In the experiment, Liberman et al. varied the slope of the transition and

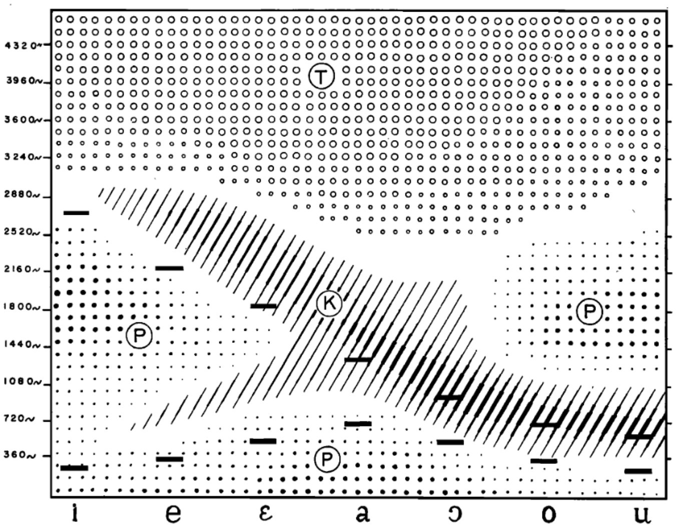


Figure 2.1: The burst timing and frequency leads to perception of different stop consonants. The preceding vowel (context vowel) indicated on the abscissa affects the perception of the consonant, an effect known as *co-articulation*. See Cooper et al. (1952).

asked one group of subjects to respond with /b,g,d/ (voiced), while the other group had the choice to pick from /p,k,t/ (unvoiced). Based on this study alone, they concluded that the F2 transition is a more robust cue for stop consonants than the burst. As a result of these synthetic speech sounds the burst was discarded at Haskins as the relevant acoustic cue.

2.1.1 Gestural Speech Perception

For both of the above mentioned Haskins Laboratory studies, a wide variability of acoustic features (i.e. bursts, F2 transitions) was found to cue the same consonant. Together with the co-articulation effect (dependence on context vowel), this variability made the people at the Haskins Labs look somewhere else for the invariant part of speech transmission. This led Liberman to *Motor Theory* (MT), where it is believed that articulation *gestures* are the invariant for perception relevant cues and not the speech signal itself.

The MT was further developed by Liberman and Mattingly (1985, 1989). They argued that speech is transmitted gesturally rather than acoustically; which means that the place of articulation (i.e. place in the vocal tract where obstruction occurs), manner of articulation (i.e. configuration and interaction of the body parts involved in shaping the vocal tract) and voicing (i.e. oscillation of the vocal cords) are the invariant cues rather than the tempo-spectral properties of the signal. The receiver receives speech by using his own motor system for speech production to infer what was said. It is important to point out that according to MT one's neural motor system is responsible for perceiving speech, but not in perceiving environmental sounds. In their final theory, gestures not necessarily referred to the actual shape of the vocal tract, but rather to the higher level neural commands that are transmitted from the brain to the muscles in the vocal tract. The physical evidence for this theory is slim.

Direct Realism (DR) also proposes a gestural transmission of speech. It is argued that a universal theory for perception would favor gestural perception, since gestures are what we perceive visually, haptically and so on. The differences between DR and MT are subtle, but important. Fowler, the founder of DR, writes that MT and DR disagree in all points but the gestural transmission (Fowler (1996)). She assumes that DR does not involve the use of the perceiver's motor system and thus also does not involve the necessary distinction between speech and other auditory perception, which makes DR a simpler and more universal theory than MT.

Direct Realism argues that auditory perception is real, meaning that humans perceive real objects (i.e. the vocal tract). The perception is also direct in the sense that humans directly perceive the vocal tract movements rather than properties in the acoustic signal (Fowler (1986)), much as humans perceive edges instead of photons (visual system) or structure instead of skin deformation (haptics). In an analogous way, Fowler argued that our auditory system uses sound to directly infer the vocal tract shape. So whereas MT argues that a listener perceives the neural pattern, Direct Realism assumes that the vocal tract position itself is received (Fowler (1996)). One might ask how the auditory system makes this transformation.

Often the concept of coarticulation (Liberman et al. (1967)) is used to support

DR, as well as the McGurk-effect[1] (McGurk and MacDonald (1976)), which both emphasize the directness of speech transmission (Galantucci et al. (2006), Fowler, 1986, 1996). While MT and DR are appealing theoretically and logically, they lack sound experimental evidence. They for example fail to appreciate the importance of Miller and Nicely (1955) or the articulation index (AI) model. The concept of co-articulation which is support these theories is under debate as well and might only be a result of the synthetic speech used at Haskins Laboratory.

2.1.2 Acoustical Speech Perception

Even though commonly accepted speech perception theories (i.e. Motor Theory and Direct Realism) are based on the early Haskins Laboratory work, there has been some criticism on the validity of their results. Especially the lack of a link of acoustic properties of the signal and perception has been criticized ((Blumstein and Stevens, 1980, p. 648), Remez et al. (1981), (Cvengros and Allen, 2011, p. 11)). Ohala strongly criticizes the gestural theories in the paper with the descriptive title "Speech perception is hearing sounds, not tongues." His theoretical argument is based on phonological evidence. He argues that if speech sounds had evolved to be distinct articulatorily it would be evident in the development of languages and of what sounds they use to encode the messages. However this is not what is found: "language code units" (phonemes) are selected and developed for their acoustic properties. For example many of the obstruents[2] are similar in their production and only make use of a small range of the possible vocal tract configurations, yet consonants in general and obstruents particularly, make up an important part of a majority of languages. Ohala believes that languages would have evolved differently if the articulation gestures were transmitted.

[1] Phenomenon that demonstrates interaction between visual and auditory components in speech perception: If a sound is played and the visual components of another sounds are seen, this can lead to the perception of a third sound.

[2] Obstruents are consonants formed by blocking (obstructing) airflow; the obstruction leads to increased air pressure in the vocal tract. In phonetics obstruents are part of a speech classification scheme, the opposite group being sonorants, which are produced with a continuous airflow.

Fowler (1996) replied to his criticism in an equally strong way and addressed each of Ohala arguments, but also clearly distinguished DR from MT. The theoretical conflict remains unresolved.

Cole and Scott (1974) provided a different explanation of how humans perceive speech. They argued that consonant-vowel signals can be split up in three parts: (i) an invariant part, (ii) the transition and (iii) the envelope. In some phonemes[3] one of the three parts is sufficient for recognition. In general, however, having only one part limits the accuracy and produces a small confusion group. According to them, the role of the signal envelope is to smooth together phonemes in conversational speech. Therefore, the effect of changing a consonant in a sentence could be reduced by keeping the envelope of the replaced consonant.

Blumstein et al. (1977) The multiplicity in the Haskins Lab results led Blumstein et al. to think that perception was based on a combination of several acoustic features, which they referred to as *integrated cues*. They used the so called *aversion effect*, to verify their hypothesis. It was shown by Eimas and Corbit (1973) that, if a consonant is repeated to a subject multiple times before the actual identification test, the subject is less likely to perceive the presented consonant. Blumstein et al. (1977) used this effect in their synthetic speech experiment to test the strength of acoustic features. Their experiment used three different stimuli for which they expected different degrees of adaptation: Stimulus (i) with burst and transition in agreement, stimulus (ii) with burst and transition in disagreement, and stimulus (iii) no burst but only an F2 transition. They expected the stimulus (i) to show the greatest aversion effect, whereas the stimuli (ii) and (iii) were expected to show moderate adaptation in the subjects. Their hypothesis was true, adaptation was the

[3] Phones represent the basic set of sounds that can be used to describe most languages. They are usually written in brackets (i.e. [ ]). Every language will choose to use only a subset of the phones. The set of unique sound categories that a language uses are called the **phonemes** of the language. Two sounds are considered to be parts of different phonemes if they make a distinction between two words. The words mat and pat for example have different meanings; therefore, we can conclude /m/ and /p/ are different phonemes in English. Phonemes are written between slashes. See (Gold et al., 2000, p. 310).

strongest if both F2 transition and burst were in agreement. Thereby they confirmed the importance of F2 transition and burst for consonant perception.

Stevens and Blumstein (1978) In a second synthetic speech experiment the probability for burst-only recognition was found to be low (18%), while F2-transition-only accuracy was at 81% and the burst plus F2 transition accuracy (reference condition) was 90%. Solely based on these numbers they concluded that the F2 transition feature is a sufficient cue for synthetic plosive consonants.

Stevens and Blumstein (1978) suggested, based on their experiment (with synthetic speech), that it is neither the burst nor the F2 transition that matters for recognition. Instead, they hypothesized that it is the spectral slope of the consonant onset. They proposed that the cue is due to the release and closure of a stop consonant and not due to time sequence of events. The cue was not dependent of the preceding vowel in their experiments. In their opinion, the transition just fulfills the purpose of smoothing the spectrum between the consonant and the vowel and, therefore, their purpose is to avoid a new onset ((Stevens and Blumstein, 1978, p. 1367)). They confirmed their results with natural speech in Blumstein and Stevens (1979).

Remez et al. (1981) used a very different approach to investigate speech perception. They used sine waves to generate totally voiced signals such as “Where were you a year ago?” This sine-wave speech lacked the transitional and onset cues that are generally assumed to be important (therefore the title of the paper “Speech perception without traditional speech cues”). The experiment investigated whether or not the speech was still intelligible. When the subjects were not told that they were listening to speech, only 5/31 subjects thought the signal resembled speech and only two out of the five were able to understand the speech. On the other hand, when people were told that the signal was computer generated speech, 9 out of 31 were able to transcribe it correctly, 10 did not recognize a sentence at all and the remaining 12 were able to transcribe part of “Where were you a year ago?” If the phrase was given to the subjects beforehand, most subjects claimed that they actually heard the phrase; however, they found the stimuli to sound unnatural. The paper was concluded with the statement that listeners do not need the transitions and onsets as cues in order to perceive speech. This conclusion seems both strong and questionable, given their data

Table 2.1: The 11 acoustic features that were used in Dubno and Levitt (1981)

Parameter	**Abrev**	**Units**	**Description**
Vowel peak frequency	VPF	[Hz]	Frequency at most intense spectral peak (vowel portion)
Consonant spectral peak	CF1	[Hz]	Frequency at most intense spectral peak (consonant portion)
Consonant spectral peak	CF2	[Hz]	Frequency at sedcond most intense spectral peak (consonant portion)
Origin of second formant transition	ORIG	[Hz]	Frequency at onset of change in steady-state formant
Magnitude of second formant transition	MAG	[Hz]	Change in frequency from onset of transition to onset of consonant
Direction of second formant transition	DIR	–	Direction of change in frequency from start at end of transition (rising, falling, no change)
Overall consonant-noise bandwidth	BW	[Hz]	Derived from upper and lower cutoff frequencies at points 3 dB below average level of power spectrum
Crossover frequency	XF	[Hz]	Frequency above which the consonant spectrum is more intense than the noise spectrum
Total energy of consonant	CE	[dB] re: least intense consonant	rms energy averaged over consecutive 30.72 [ms] time windows corresponding to onset and offset of consonant
Total energy vowel	VE	[dB] re: least intense vowel	rms energy averaged over consecutive 30.72 [ms] time windows corresponding to onset and offset of vowel
Consonant-to-noise ratio	C/N	[dB]	rms energy in consonant portion and noise portion, converted to [dB]; energy (noise) subtracted from energy (consonant)
Vowel-to-noise ration	V/N	[dB]	rms energy in consonant portion and noise portion, converted to [dB]; energy (noise) subtracted from energy (vowel)
Consonant duration	CD	[ms]	Time form start of consonant to end of consonant
Vowel duration	VD	[ms]	Time from start of vowel to end of vowel
Closure duration	CLD	[ms]	Time from onset of stop closure to plosive release

Dubno and Levitt (1981) conducted an extensive experiment with 91 natural speech tokens. They looked at 11 acoustic features (see Table 2.1) in both quiet and at + 5 [dB] SNR in *speech-weighted noise* (SWN). Both vowel-consonants (VC) and consonant-vowels (CV) spoken by a male talker were used in the experiment. The stimulus level was raised from 20 to 54 [dB SPL], the strength of the 11 acoustic features were measured along with the perceptional accuracy, the strength of the features was then correlated to the perceptional data. The results showed the highest correlation in quiet for the consonant energy (CE), consonant duration (CD), and the origin of the second formant transition (ORIG). In the SWN condition the consonant-to-noise (C/N), consonant spectral peak frequency (CF1) and the consonant duration (CD) had the highest correlation with the perception scores. The importance of the features were consonant dependent.

Dubno et al. (1987) In response to Stevens and Blumstein (1978) and Blumstein and Stevens (1979) Dubno et al. (1987) investigated the importance of the consonant onset spectra as a perceptual cue. Using synthetic stimuli they investigated the duration of the onset spectra/voicing and discovered that it needed to be longer than 2 [cs] to have > 87% accuracy with NH subjects.

Turner et al. (1992) used synthetic stimuli and presented them at different SNRs. Each stimulus was presented in two forms: (i) with full duration and (ii) only the beginning of the consonant (i.e., first 4 [cs]). The results showed no differences between the truncated and the original signal. Their results thus provided evidence for Stevens and Blumstein (1978) and Blumstein and Stevens (1979) hypotheses, while it provides evidence against Liberman et al. (1967), since the truncated consonants did not include any vowel information.

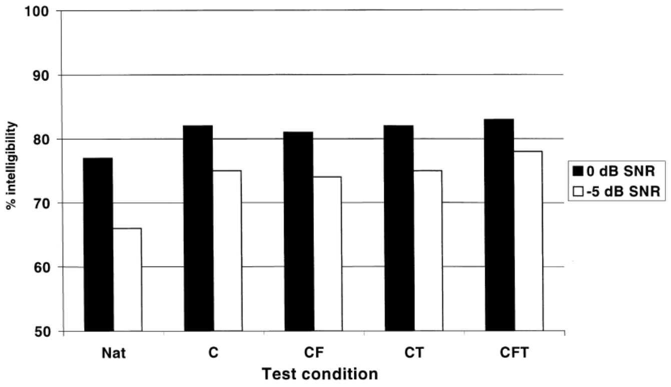


Figure 2.2: Results for the natural and the different enhancement conditions of the CVC experiment in Hazan and Simpson (1998) at 0 [dB] and -5 [dB SNR]. It is noticeable that the C (amplified wide-band consonantal energy, i.e. burst for plosives, friction for fricatives or nasal portion for nasals) condition brought a significant improvement. The CT condition does not improve the results any further; it even decreases the performance slightly. See Hazan and Simpson (1998).

Hazan and Simpson (1998) took a different approach: rather than removing potential cues from the natural speech signal (i.e. *vowel-consonant-vowel* (VCV)), they enhanced them. They worked with four different modifications:

C the region of highest wide-band consonantal energy release was amplified by 6 [dB] (Fricatives and Nasals) and 12 [dB] (Plosives)

CT both the consonantal region and the transition to the formant was enhanced

CF the consonantal region (for plosives and fricatives) was filtered before being amplified

CTF the vowel onset/offset regions were also amplified in addition to the modifications of CF

The authors carried out two analyses of variance (ANOVA) for the two SNRs on the complete set of data. The effect of test condition was significant at -5 [dB SNR] [F(4,48) = 41.54; p < 0.0001] and at 0 dB SNR [F(4,48) = 16.04, p < 0.0001]. At both SNRs, significantly higher average intelligibility scores were obtained for all four enhanced conditions (C, CT, CF, CFT) than for the natural condition. However, a look at Figure 2.2 shows that the improvement from C to CT slightly decreases. C seemed to enhance the signal the most, whereas everything else only had a small effect. This suggests that the burst is more relevant than the F2 transitions.

Li and Allen (2011) 14 years later Li and Allen developed a method called three–dimensional deep search (3DDS) to find consonant cues in highly variable natural speech. 3DDS uses psychoacoustic testing in NH subjects to triangulate the cue in the three dimensional time-frequency-amplitude space. They first applied it to the 16 Miller and Nicely consonants, each paired with 3 vowels (Li and Allen (2011)).

2.1.3 AI gram

The *articulation index (AI)-gram*, as shown in Figure 2.3, is the basis for all the cue identification experiments (3DDs). It is a time-frequency representation of stimulus signals that is built to simulate human auditory processing; it takes into account the

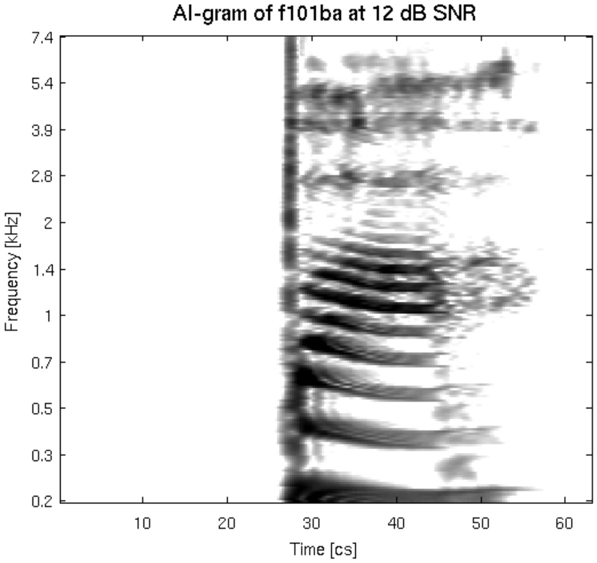


Figure 2.3: The AI gram predicts the audibility of noise masked stimuli using Harvey Fletcher's critical-band auditory model.

effect of noise masking on audibility. The base for the AI-gram is Harvey Fletcher's *Articulation Index* (AI) and his critical-band auditory model. Given a stimulus and the masking noise, the AI-gram produces an image showing the audible part of a speech signal in the time-frequency space. More information on the AI-gram can be found in Régnier and Allen (2008), Lobdell and Allen (2006) and Lobdell (2009).

In order to find the consonant cues in natural speech tokens, 3 psychoacoustic experiments are combined. Each experiment modifies the speech along one of the axes of the AI-gram (i.e. time, frequency, SNR). The perceptual data from these three experiments allow it to see where the necessary information on each axis is, by assuming that the perceptual necessary cue is lost as soon recognition drops below 90%. The combination of the three perceptual curves indicates the cue region (Figure 2.4).

The 3DDS method was applied to stop (plosive) consonants and fricatives. For the plosives the cues are bursts in the time-frequency regions as shown in Figure 2.5a. It is obvious that the *voice onset time* (VOT) plays an important role to distinguish the voiced plosives (/bɑ, dɑ, gɑ/) from their unvoiced counterparts (/pɑ, tɑ, kɑ/).

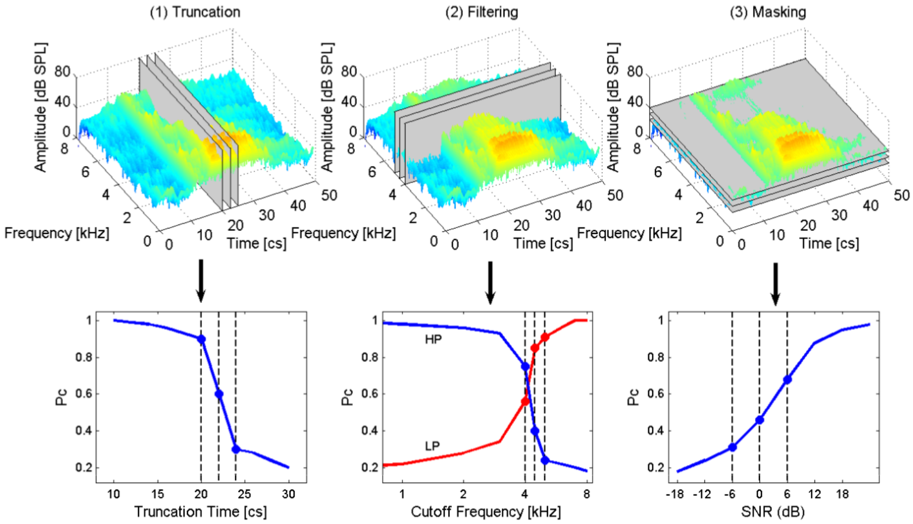
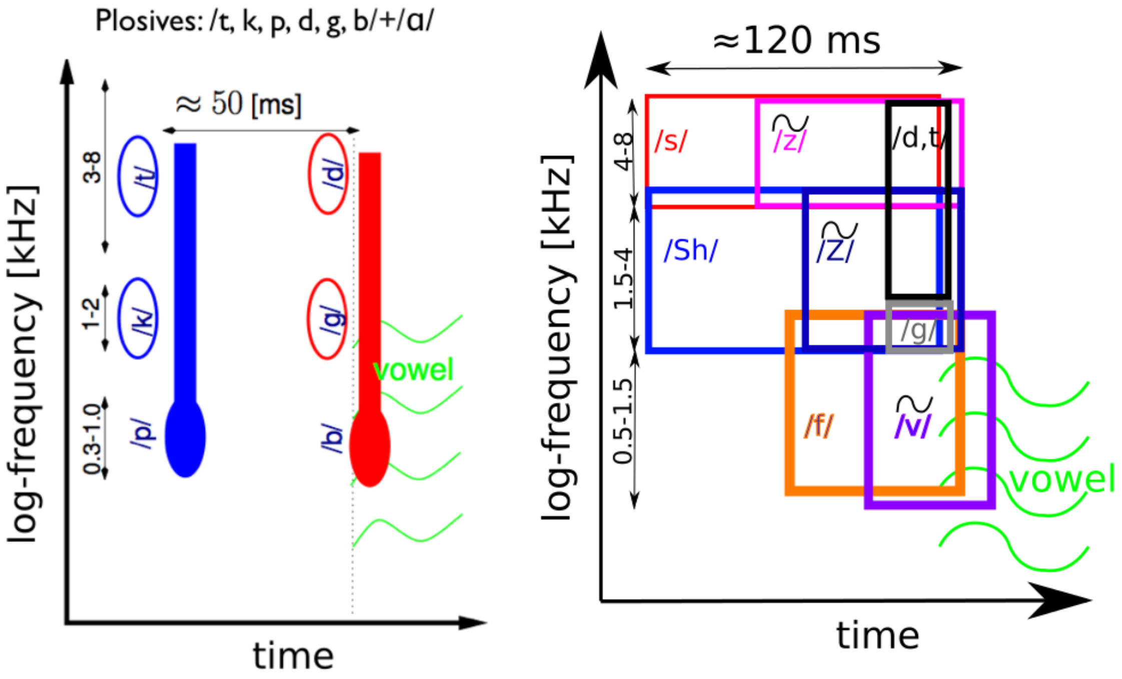


Figure 2.4: The 3DDs method triangulates the consonant cues in the time frequency space using three psychoacoustic experiments. See Li et al. (2010).

For the fricatives /s, z, ʃ, ʒ, f, v/ + /ɑ/ the 3DDS cue regions are summarized in a schematic in Figure 2.5b. The alveolar consonants /sɑ,zɑ/ have their cue region in the sustained frication no lower than 2 [kHz], while the palato-alveolar consonants /ʃɑ,ʒɑ/ have their cue region between 1.3 and 3.6 [kHz]. For the non-sibilant labiodentals /fɑ,vɑ/, the cue region is between 0.6 and 1.7 [kHz]. For the voiced sibilants, the friction noise is modulated by the pitch fundamental (FO). It was found in the high-pass experiments that the low frequency voicing energy is perceptually not necessary.

The described speech cue research left us with a rather conflicting set of conclusions. The field generally accepts that stop consonants are identified by bursts and transitions (Allen and Li (2009), Blumstein and Stevens (1980), Cooper et al. (1952), Heil (2003) and Li et al. (2010)).



(a) 3DDS results for plosives

(b) 3DDS results for fricatives

Figure 2.5: Schematic summary plots for the results found by the 3DDS method applied to plosives and fricatives.

2.2 Consonant Perception in Hearing Impaired

Consonants are more prominent in English than vowels (Mines et al. (1978)). However, the importance of consonants in speech perception is still debated. For example, while Kewley-Port et al. (2007) and Cole et al. (1996) argued that vowels are more important, Miller (1951) argued for the importance of consonants. It seems to be undebated, however, that consonants are more affected by high-frequency sensorineural hearing loss than vowels, since they consist of more high-frequency energy. Some of the early important studies on consonant perception in hearing impaired subjects are Bilger and Wang (1976), Owens (1978), Dubno et al. (1984), Boothroyd (1984), Zurek and Delhorne (1987), which are described in more detail below.

Owens used CVCs in quiet (no added noise) to test consonant perception in HIs. The study presented CVCs and the subjects had to choose which CVC was presented among four answer choices. The results showed a dependency on CVC tokens and it was noted that, as Miller and Nicely (1955) found with NH subjects, the confusion groups for consonants were rather small.

Zurek and Delhorne (1987) tested the consonant perception of both HI and noise-masked NH listeners (i.e. a hearing loss was simulated by masking noise). The noise was shaped for the NH listeners such that their noise-masked thresholds matched the pure-tone thresholds of the individual HI ears. The matching was implemented over the range of 0.125 - 4 [kHz]; hearing losses over 4 [kHz] were 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 utterances, approximated the perception of the corresponding HI ears. The majority of HI ears had a <70% probability correct rate at even the quiet condition, with 4 out of 6 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 degree of averaged error. One might conclude from this study that the "distortion factor," also known as "SNR-loss" is not a key factor. The set of conclusions is frequently at odds in all of the these studies. As a result it is difficult to identify the most important factors. Speech perception in HI subjects is a very complex problem due to the large number of interacting factors. It is not clear how to make further progress.

2.3 Prescriptive Procedures

Hearing aids often do not provide the desired benefit to the hearing impaired people (Dillon (2012)). According to Kochkin, less than 60% of hearing aid users are satisfied with their devices. Especially people with mild to moderate hearing losses are often not satisfied with the performance of their assistive listening device. For them it is more acceptable to have subtle difficulties with understanding speech than to wear a hearing aid. Up to the present day it remains unclear why two people with the same hearing loss (i.e. pure tone audiogram) benefit differently from hearing aids. We believe that this is a universally accepted truth.

The history of methods for gain prescription for hearing aids goes back to 1935, when Knudsen and Jones proposed to mirror the audiogram by subtracting a constant from the measured hearing thresholds. Later studies showed that this led to excessive gain (Dillon (2012)). Watson and Knudsen (1940) proposed the most comfortable level (MCL) as appropriate for hearing aid fitting. At the end of World War II a large study by Lybarger (1944) found that, on average, people chose about half of their hearing loss as MCL. This result came to be known as *half-gain rule*. Several of today's fitting procedures (e.g. NAL-R) are variations of this rule.

Many consider the 1940s to be the decade when the field of audiology was born (Humes (1996)). Two big reports had a major impact on the young field of audiology and fitting procedures: Davis et al. (1946) (known as Harvard report) and Radley et al. (1947) (known as the MedResCo report). The two reports were wrongly taken as evidence for a "one size fits it all" approach to hearing aid fitting. The data in the two reports suggested that a +6 [dB/octave] frequency response would be right for all potential hearing aid wearers (Humes (1996)). Based on this report, the field fitted hearing aids according to the *comparative approach* (Carhart (1946)), where 3-4 similar hearing aids were compared in different listening situations and the one that performed the best in those situations and in a battery of tests was chosen. The comparative approach remained popular until the mid to late 1980s, when the programmable hearing aid first came to market.

However, criticism of this approach became stronger from 1960 to 1983. Walden et al. (1983) systematically analyzed the 5 key assumptions that underlie the comparative approach and found them to be wrong. Many other studies during this period showed that the so called *selective amplification* (i.e. amplification depended on the individual hearing loss) resulted in better adjusted hearing aids (Shore et al. (1960); Walden et al. (1983)).

The shift from the comparative approach to selective amplification led to the creation of many different prescriptive procedures in the late 1970s and the early 1980s. Different approaches for the audiological assessment were chosen; some procedures were based on pure tone thresholds, while others were based on various measures of loudness (Allen et al. (1990)), such as the most comfortable loudness (MCL) or the loudness discomfort level (LDL). With the many emerging procedures, the question of which was the best arose. Many studies compared the different approaches and found different results. In his review, Humes (1996) came to the conclusion that many of the prescriptive procedures, even though slightly different, resulted in similar hearing aid selection, because achieved gains often did not differ much in reality. Byrne (1986) from the National Acoustic Laboratories in Australia compared six procedures (1 threshold based, 4 MCL, 1 LDL) with nonsense syllables in noise. He used a paired comparison paradigm to judge the intelligibility and quality of the stimuli. Mean nonsense-syllable identification scores showed no significant difference. Similar results were found by Sullivan et al. when they fitted hearing impaired ears with four different prescriptive procedures (1 MCL, 2 threshold, 1 special adaptive algorithm). The two above mentioned studies used simulated hearing aids (i.e. they listened to filtered sound in a booth over headphones). Humes et al. (1990) repeated a similar study with real behind the ear (BTE) hearing aids, but also did not find a significant difference in speech-recognition performance.

In summary, it can be said that the many different procedures differ in their prescribed gains and in their rationales underlying the procedure (e.g. equal loudness in all bands or normal hearing thresholds in all bands), but studies have failed to identify a difference in their efficiency to make speech more intelligible. Nevertheless it can be stated that the most evaluated and popular procedures are NAL-R (Byrne and Dillon (1986)), the MSU (Cox, 1983, 1985, 1988), the Berger method (Berger et al. (1977)), and POGO (McCandless and Lyregaard (1983) and Schwartz et al. (1988)). NAL-R, Berger and POGO only require thresholds; MSU on the other hand fits the hearing device according to threshold and loudness measurements.

All of the above mentioned formulas are for linear hearing aids. Nowadays, almost all hearing aids are *wide dynamic range compression* (WDRC) hearing aids. New formulas were developed for such devices (e.g. Allen et al. (1990); Dillon (1999)). Depending on the input level, they use different gains and are therefore non-linear hearing aids. In controlled circumstances, as the ones used for this work, where the subjects are in a booth and are able to adjust the level, a wide dynamic range compression would not be appropriate.

NAL-R was chosen as prescriptive procedure for the present experiments. NAL-R is one of the half-gain rule variations and was developed by the National Acoustic Laboratory of Australia (NAL) (Byrne and Dillon (1986)). Like all the half-gain based procedures, it is based solely on the audiogram. It is the revised version of the older prescriptive formula NAL. The goal of the NAL procedures is to prescribe the gain so that speech in all bands has the same loudness. After introducing the first version (NAL), tests in the early 1980 showed that equal loudness was not obtained in all bands, especially for steeply sloping losses (Dillon (2012)). This shortcoming was corrected based on a major evaluation by the NAL-R standard. NAL-R has been widely used in clinics and is most suitable for mild to moderate hearing losses. However, there are problems associated with pure tone based prescription. These problems may well be one of the main reasons why $\approx 40\%$ of the people are not satisfied with their hearing aid.

Chapter 3

Methods

Four experiments (Table 3.1) with hearing impaired subjects were conducted by Han (2011). Two of these experiments (Exp 1 and 3) presented CV tokens with flat-gain (FG having no frequency dependent amplification) at the subject's most comfortable loudness (MCL); the second two (Exp 2 and 4) presented tokens with a frequency dependent gain based on the subject's individual hearing loss (HL(f)) according to the NAL-R standard (see Section 3.2.2). The first two experiments in fall 2010 - each a FG experiment and a NAL-R experiment - included 27 subjects with a relatively small number of trials per subject (16 consonants x 6 utterances x 2 presentations x 6 different noise conditions=1152 at max; if a consonant had a score $< 18.75\%$ at a certain SNR, lower SNRs were skipped). Experiment 2 and 4 in spring 2011 were verification experiments with a subset of the subjects (9 subjects) from the first two experiments but with more trails per token and SNR (only 4 SNRs and 2 tokens per CV), depending on their error rate subjects had 800-1000 trials. This document focuses on the verification experiments: Exp 2, the verification flat gain experiment referred to as FG Exp in the rest of the thesis and Exp 4 the verification NAL-R experiment referred to as NAL-R Exp. The experiments were conducted in the following way.

Table 3.1: The four different experiments performed by Han (2011), they differ in the gain provided as well as in the number of subjects, tokens, CVs and SNRs.

	16 CVs, 6 tokens per CV 6 SNRs, 27 subjects	14 CVs, 2 tokens per CV 4 SNRs, 9 subjects
FG	Exp I	Exp II (FG Exp)
NAL-R	Exp III	Exp IV (NAL-R Exp)

3.1 Flat-Gain (FG) Experiment

3.1.1 Subjects

Seventeen HI ears (Table 3.2) were tested from April 2010 to May 2010. Each participant passed a middle-ear examination and was confirmed to have the same hearing level (HL) (see Figure 3.1), as measured in the previous experiment in fall 2009, which means their audibility had not changed.

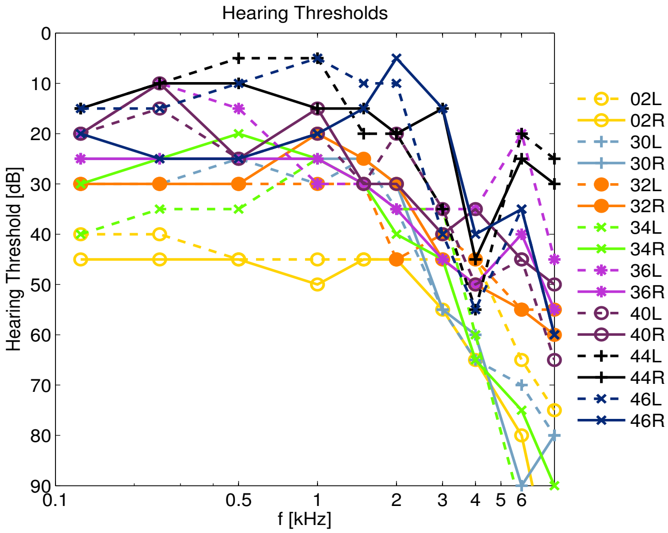


Figure 3.1: PTTs for the sixteen ears.

Table 3.2: Subjects' age, PTAs and MCLs

HI ear	Age	PTA	MCL	
			FG	NAL-R
44L	65	10	82	77
44R	65	15	78	77
46L	67	8.3	82	85
46R	67	16.6	82	86
40L	79	21.6	79, 81	
40R	79	23.3	80	80
36L	72	26.6	68	75
36R	72	28.3	70	75
30L	66	30	80	79
30R	66	26.6	80	79
32L	74	35	79	81
32R	74	26.6	77	78
34L	84	31.6	84	85
34R	84	28.3	82	85
02L	82	45	83	88
02R	82	46.6	82	89
(μ,σ)	(74,7)	(29,15)	(79,4)	(81,5)

3.1.2 Speech Stimuli

The consonant-vowel (CV) syllables consisted of 14 consonants (6 stops, 6 fricatives, and 2 nasals) followed by the /ɑ/ vowel. These CVs are known as the Miller and Nicely consonants (Miller and Nicely (1955)). Two fricatives, /θ/ and /ð/, were not used in the experiment, as they have high error, even for normal hearing (NH) subjects (Phatak and Allen (2007)). To reduce the time of administration, only 2 talkers (1 male and 1 female) were selected per consonant. The recordings from the

Linguistic Data Consortium (LDC) 2205S22 database (Fousek et al. (2004)) were used, where the CV are all spoken by native speakers of American-English and were recorded at a 16 kHz sample rate. The tokens were chosen from those for which there was less than 10% error in data of NH listeners. In total, there were $14 \times 2 = 28$ different tokens (one female and one male token per CV). All 28 tokens had zero-error for SNR ≤ -2 dB (SNR90 $\leq$ -2) across the 14 NH listeners in the Phatak and Allen (2007) study. In this document tokens are referred to as $/\text{f}\alpha/_m$ for the male token of /fɑ/.

After conducting the experiment, we realized that one token, m112 /fa/, was damaged during the signal processing due to a minor software bug. During the filtering of silent parts existing before and after the speech stimulus, the low frication energy of the /fa/ was removed (Han (2011)). Consequently, it was necessary to remove the token from our analysis, and the number of /fa/ tokens reduced from 2 to 1 in the FG experiment. The problem was fixed for the NAL-R experiment, resulting in a stimulus set with no broken tokens.

3.1.3 Procedure

The subjects had one practice session, with 14 syllables in quiet. These 14 tokens (one per CV) were different from the tokens used in the experiment, to limit learning effects. Syllable presentation was randomized over consonants, speakers, and SNRs. Three SNRs (12, 6, and 0 dB) and a quiet conditions (no added noise) were tested. The experiment consisted of two sessions, to limit the duration of the booth time and thus reduce subject fatigue.

In the first session (I), each of the 28 tokens was presented 4 times at each SNR. This resulted in 28 tokens × 4 SNRs × 4 presentations = 448 trials and the experiment took a total of 30-40 mins per ear. For each token at each SNR, the correct score percentage was calculated. The possible scores were 0% (0/4), 25% (1/4), 50% (2/4), 75% (3/4) and 100% (4/4). In the second session (II), the number of trials depended on the subject's performance in the first session, as shown in Table 3.3. Across the two sessions each token was presented between 5 and 10 times, depending on the error rate in the first session. This resulted in 10 and 20 presentations per consonant at each SNR (see Table 3.3).

Table 3.3: Number of trials in the two sessions (I,II) of the experiments per SNR and token.

	Flat-Amplification			NAL-R		
Errors (P_e)	I	II	Total	I	II	Total
0 (0%)	4	1	5	4	2	6
1 (25%)	4	2	6	4	2	6
2 (50%)	4	5	9	4	5	9
3 (75%)	4	6	10	4	6	10
4 (100%)	4	6	10	4	6	10

The rationale behind this experimental design was to increase the sample size as a function of the session I error, to obtain more data when there are more errors. Due to subject based factors (attention) the total number of trials per consonant (sum of sessions I and II) was not same for each subject. Between 800 and 1000 trials were presented to each subject in total.

3.2 NAL-R Experiment

3.2.1 Subjects

A year after the FG experiment (May 2011) all but one subject also participated in the NAL-R experiment. The subjects' thresholds were retested and all subjects had the same pure-tone hearing threshold as in the previous year, within 5 [dB] in the testing frequencies (from 1.25-8 [kHz]). All subjects again reported no history of middle ear pathology.

Table 3.4: The tokens used in the two experiments are listed along with their SNR90 in dB in parentheses. The tokens that were swapped in the NAL-R experiment are shown in red.

pa	f103 (-20)	m118 (-16)	da	f105 (-16)	m118 (-10)
ta	f108 (-16)	m112 (-20)	ga	f109 (-10)	m111 (-16)
ka	f103 (-10)	m111 (-16)	va	f101 (-10)	m118 (-2)
fa	f109 (-16)	m112 (-10)	za	f106 (-20)	m118 (-16)
sa	f103 (-16)	m120 (-10) m107 (-10)	ʒɑ	f105 (-16)	m107 (-10) m111 (-20)
ʃa	f103 (-16)	m118 (-16)	ma	f103 (-16)	m118 (-16)
ba	f101 (-10)	m112 (-2)	na	f101 (-10)	m118 (-2) m112 (-16)

3.2.2 Speech Stimuli

The speech stimuli for the NAL-R experiment were the same as in the FG experiment with a few ones swapped (see Table 3.4)

NAL-R Amplification

The difference between the two conditions is the spectral shape of the gain. In the FG experiment the gain is uniform over all frequencies. In the NAL-R condition a spectral emphasis according to a subject's hearing loss was provided. The NAL-R formula was calculated according to the following two steps (Dillon (2012)).

Step 1: Calculate, as function of HTL(f), the hearing thresholds of the ear at frequency f.

$$X(dB) = 0.15 \times (\mathrm{HTL}(500) + \mathrm{HTL}(1000) + \mathrm{HTL}(2000))\,/3 \tag{3.1}$$

Step 2: Calculate the prescribed real-ear gain REG at each frequency $f = \{0.25,\ 0.5,\ 1,\ 1.5,\ 2,\ 3,\ 4,\ 6\}\ kHz$, with $c_f = \{17,\ 8,\ 3,\ -1,\ -1,\ 1,\ 2,\ 2\}$.

$$REG_f(dB) = X + 0.31 \times \mathrm{HTL}(f) - c_f \tag{3.2}$$

3.2.3 Procedure

All test procedures were the same as in the FG experiment. After the practice session conducted with 14 tokens that are not part of the actual experiment, syllable presentation was randomized over 14 consonants, 2 speakers, and 4 SNRs. As in the FG experiment, the experiment consisted of two sessions (see Table 3.3).

3.3 Analysis

The two described experiment present a rich and extensive data set. In this thesis traditional and new methods are used to analyze the data. These different methods are described below.

3.3.1 Error Analysis

Many past studies have focused on an error analysis. Often scores are averaged over different tokens and different consonants. As shown by Trevino and Allen (2013b) this approach can be fatal, since it destroys (due to averaging) details which are critical to a detailed diagnosis of a speech loss. In this thesis, such an average error analysis is avoided.

3.3.2 Confusion Patterns

Confusion patterns (e.g. Figure 3.2) were first introduced by (Allen, 2005b, p. 2215). Instead of plotting the error versus SNR, the error is expressed in terms of the different confused consonants. In other words the confusion pattern is a row of a *confusion matrix* (CM) plotted as function of SNR. This shows which confusions dominate as a function of SNR. This allows one to observe more information than just the error plot.

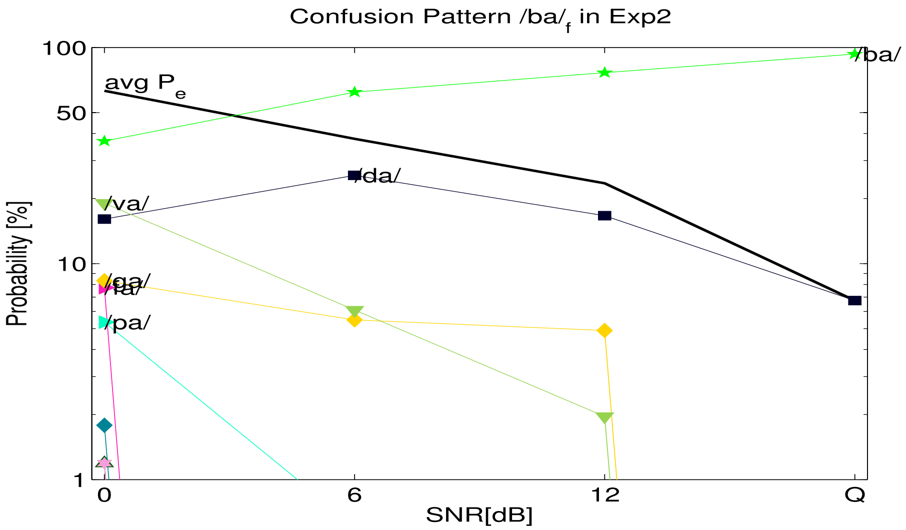


Figure 3.2: Confusion pattern for the female /bɑ/$_f$ in the FG experiment averaged across all subjects. The probability of the correct answer /bɑ/$_f$ is high at high SNRs, but only about 40% at 0dB SNR. The main confusion is /dɑ/; it exists at all SNRs. Other confusions such as /pɑ,vɑ/ get introduced at lower SNRs.

3.3.3 ANOVA

Analysis of Variance (ANOVA) is a widely used statistical method to assign variation in an experimental procedure to a number of sources. Even though the name suggests otherwise, ANOVA is a method to compare means of different groups (populations) to see if they are significantly different. The null hypothesis H_0 is the contrary of what is expected, namely: All populations have identical means. The result of an ANOVA is an F value, from the Fisher distribution, that depends on the *degrees of freedom (DF)*[1] and a p-value. The p-value as in many other tests is the probability that, if H_0 is true (i.e. no difference between the means), one will observe a difference at least as big as that actually observed in the data based on random variations (chance). If the p-value is small, that means it is unlikely that the observed difference is just due to random variation and therefore it is reasonable to assume that H_0 is not true. If the p-value is smaller than a threshold chosen beforehand, called significance level

[1] The degrees of freedom are determined by the group size and the number of groups.

α, the results can be called statistically significant. Typically α is 5%.

3.3.4 Entropy measures

Information theory and the important concept of entropy were introduced by Shannon (1948). George Miller was the first to apply *information transmitted* (IT), also known as the *mutual entropy.* Entropy, a measure of the randomness of a response, is defined as the expected value of the information $log(1/p_n)$; the CM row sum is $\sum_n p_n = 1$, where $n = 1 \dots 14$ (14 is the number of possible responses):

$$\mathcal{H}(\mathbf{p}) = \sum_{n=1}^{N} p_i \, log_2 \left(\frac{1}{p_n} \right). \tag{3.3}$$

3.3.5 Hellinger Distance

To define the Hellinger distance metric we must start by defining a norm via an inner product, defined as

$$(\sqrt{\mathbf{p}}, \sqrt{\mathbf{q}}) \equiv \sqrt{\mathbf{p}} \cdot \sqrt{\mathbf{q}} = \sum_{n=1}^{N} \sqrt{p_n}\sqrt{q_n}. \tag{3.4}$$

The definition of the Hellinger norm is

$$||p||^2 = \sqrt{\mathbf{p}} \cdot \sqrt{\mathbf{p}} = \sum_{n=1}^{N} \sqrt{p_n}\sqrt{p_n} = \sum p_n = 1. \tag{3.5}$$

The squared norm of the difference between two vectors (i.e., responses) $||\mathbf{p} - \mathbf{q}||^{\mathbf{2}}$ is the squared Hellinger distance. Geometrically it can be interpreted the following way:

$$\sqrt{\mathbf{p}} \cdot \sqrt{\mathbf{q}} = ||\mathbf{p}|| \; ||\mathbf{q}|| \; \cos(\theta) = \cos(\theta). \tag{3.6}$$

If we take the Hellinger inner product of probability vectors $\mathbf{p}$ and $\mathbf{q}$, we get:

$$\sqrt{\mathbf{p}} \cdot \sqrt{\mathbf{q}} = \sum_{n=1}^{N} \sqrt{p_n}\sqrt{q_n} = ||\mathbf{p}||\,||\mathbf{q}||\,\cos(\theta). \tag{3.7}$$

By solving for θ, we obtain the angle between the two different /ba/ tokens:

$$\theta = \arccos\left(\frac{\sqrt{\mathbf{p}} \cdot \sqrt{\mathbf{q}}}{||\sqrt{\mathbf{p}}||\;||\sqrt{\mathbf{q}}||}\right). \tag{3.8}$$

Since $||\sqrt{\mathbf{p}}||$ and $||\sqrt{\mathbf{q}}||$ are equal to one - all the square root terms are squared and added to one before the square root is taken again - we can simplify Equation 3.8 to

$$\theta = \arccos\left(\sqrt{\mathbf{p}} \cdot \sqrt{\mathbf{q}}\right). \tag{3.9}$$

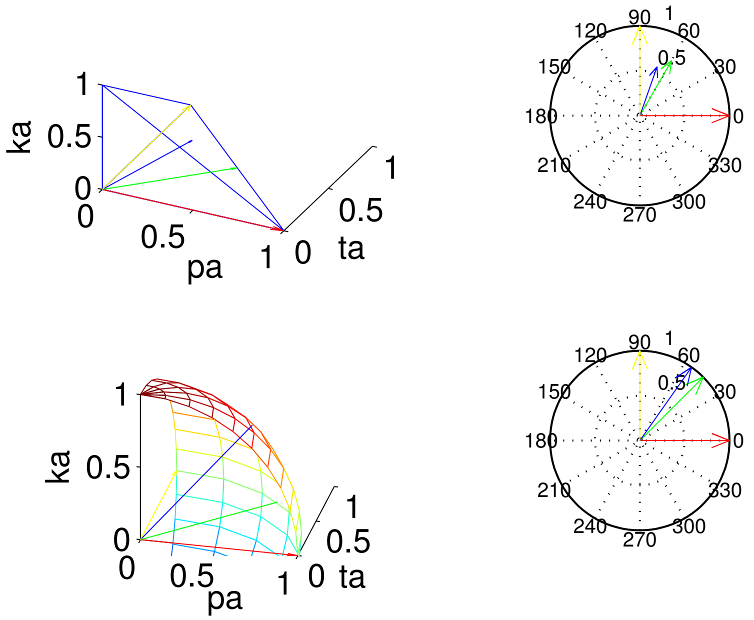


Figure 3.3: Example for Hellinger distance calculation. If the probabilities are taken without a $\sqrt{}$ transformation, they naturally lie on a simplex (an $N-1$ dimensional hyperplane). If the inner product between probability p and q is taken, the length is always 1, thus it lies on the sphere.

Table 3.5: Extract from a confusion matrix for /ba/ split for the female /ba/$_f$ and male /ba/$_m$ token. The table does not contain data from a real subject; it is a contrived example.

	/ʃɑ/	/ʒɑ/	/bɑ/	/dɑ/	/fɑ/	/gɑ/	/kɑ/	/mɑ/	/nɑ/	/pɑ/	/sɑ/	/tɑ/	/vɑ/	/zɑ/
/bɑ/$_f$ ($\mathbf{p}$)	0	0	**0.6**	0	0	0	0	0	0	0.1	0	0.3	0	0
/bɑ/$_m$ ($\mathbf{q}$)	0	0	**0.6**	0	0	0.3	0	0	0	0	0	0.1	0	0

For the example from Table 3.5 the following angle can be obtained:

$$\begin{aligned} \theta &= arccos\left(\sum_{n=1}^{N}\sqrt{p_n}\sqrt{q_n}\right) \\ &= arccos\left(\sqrt{0}\sqrt{0}+\ldots+\sqrt{0.6}\sqrt{0.6}+\sqrt{0}\sqrt{0.3}+\sqrt{0.1}\sqrt{0}+\sqrt{0.3}\sqrt{0.1}\right) \\ &= arccos(0.7732) = 0.6869 \rightarrow 39.4^{\circ}. \end{aligned} \tag{3.10}$$

The $\surd$-transformation is illustrated in a contrived example in Figure 3.3. It can be seen that the square root transformation constrains the vectors to lie on a sphere; without the transformation they lie on a simplex. For the case of two orthonormal basis vectors $\mathbf{b_1} = [100]$ and $\mathbf{b_2} = [010]$, $b_1 \cdot b_2 = 0$; thus the *arccos* is 90 °.

In many cases the vector for a specific token will consist of many zero components (see Table 3.5). This means that the listener only picks a few out of the N choices. In general the confusion group size, measured by the entropy, grows inversely proportionally with the SNR [dB]. As described earlier, the entropy is a measure of how random the answers for a given consonant or token are. It therefore also indicates how many of the N dimensions a vector occupies. Taken together, the entropy and direction cosine (or equivalently the angle) give detailed information about the effects of NAL-R.

These methods are powerful tools for comparing listeners, token, experiments, or even ears of the same listener. They allow specific questions to be asked that take into account all the information a confusion matrix has to offer. As examples, the following questions are offered:

- Which tokens for each listener are impacted the most by NAL-R? Namely, for which tokens are the angles between the FG and NAL-R experiments the largest?

- Which listener is impacted the most by NAL-R for a given token?
- What is the average angle between a particular response and the correct response?
- What is the mean angle between the subjects' responses and the correct answer and how does it change from one condition to the other?
- What is the variance of the angles between the subjects' responses and the correct answer? How does it change from one condition to the other?

3.3.6 K-means

Once a proper distance metric is defined, the vectors in this space may be clustered. For each token, there are 2x4x14=112 (2 conditions, 4 SNRs and 14 ears) vectors (i.e., data points) in the fourteen dimensional space. K–means is one of the traditional methods for doing a cluster analysis. The aim of a cluster analysis is to partition m samples into K–clusters. The motivation is very intuitive: the samples that are close to each other should share the same cluster indicators. The K-means algorithm therefore gives the cluster index of each sample by the nearest cluster center and gives the cluster center by the centroid (i.e., mean) of its members. The major drawback of K-means is that it is very sensitive to the initializations and prone to local minima when finding the solution (i.e., the solution is typical not unique). One practical way to solve that issue is to run the K-means algorithm multiple times with different initial conditions and calculate the error.

Formally, K-means is formulated as the minimization of a cost function of sum of squares: min $J_K = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - m_k\|^2$. $x_i, i = 1, 2, m$ are data samples and $X = (x_1, x_2, \ldots, x_m)$ are the samples combined in one matrix, $m_k = \sum_{i \in C_k} x_i / n_k$ is the centroid of cluster C_k with n_k samples.

Application of Clustering to NAL-R In this text the *k-means* algorithm is used to group the data points into $K = 4$ clusters, with each cluster represented by its cluster centroid $\mathbf{c}_k$, $k = 1, \ldots, K$. The centroids are then sorted according to their entropy. By comparing the centroid ($\mathbf{c}_k$) assignments of two points of a subject at a given SNR - representing the two different gain conditions - it is possible to investigate the impact of NAL-R. For all tokens, $\mathbf{c}_1$ (smallest entropy) represents the centroid of the points closest to the correct answer. Subjects that go from a higher entropy cluster ($\mathbf{c}_2$,$\mathbf{c}_4$,$\mathbf{c}_4$) to $\mathbf{c}_1$ at a given SNR because of NAL-R are considered cases where NAL-R is successful. These pairs are assigned to the category "Best" (B: $\mathbf{c}_x \rightarrow \mathbf{c}_1$, $x = 2, 3, 4$). Points that leave $\mathbf{c}_1$ because of NAL-R are cases where NAL-R failed, thus categorized as "Worst" (W: $\mathbf{c}_1 \rightarrow \mathbf{c}_y$, $y = 2, 3, 4$). Pairs of points that stay in the same cluster are classified as "Neutral" (N: $\mathbf{c}_z \rightarrow \mathbf{c}_z$, $z = 1, 2, 3, 4$). The points that change cluster but do not leave or go to $\mathbf{c}_1$ are either classified as "Improved" (I) or "Degraded" (D) depending on whether they changed to a lower or higher entropy cluster (I: $\mathbf{c}_x \rightarrow \mathbf{c}_y$ and D: $\mathbf{c}_y \rightarrow \mathbf{c}_x$, $x < y$). For the cluster analysis subject 02R and 02L were excluded, also 14R is excluded since the subject only participated in the FG experiment.

This analysis helps to answer 3 basic questions:

1. How many members do the clusters contain?

2. How do the clusters for the two tokens of the same CV differ?

3. How are the confusions grouped? What are the confusions for a given token?

4. What effect does NAL-R have on the confusions? Are there clusters that only exist because of NAL-R? Are there trends of NAL-R moving sounds to the low-error cluster or out of the low-error cluster?

A second way of analyzing the CM data is described in Appendix A, which discusses an alternative and more advanced clustering algorithm probabilistic latent semantic indexing (PLSI).

Chapter 4

Results

The results are divided into two parts: (a) audibility, and (b) the effects of NAL-R on CV perception. As mentioned in Chapter 3, several novel tools are introduced. The most fundamental tool is the probability of error. Second is the entropy measure. These are then combined to get a global view of the effect of the NAL-R "treatment." Furthermore more detailed results of the clustering are provided.

4.1 Audibility

Posner and Ventry (1977) found that subjects perform below their maximum speech discrimination abilities if tested under MCL conditions. Our data, however, proves that based on entropy measures almost all the tokens were fully audible to all the subjects under both conditions. This suggests that the entropy, in quiet, is a more meaningful audibility measure for CV identification experiments than LTASS and PTTs. Low entropy requires consistency, proving that the ear hears the signal, even when the error is large. Figure 4.1 shows the entropy as a function of the error for each token in quiet. Each subject ear is represented by a symbol, which is filled for the left ear and open for the right ear. Entropy reference conditions are indicated by lines (i.e., 1, 1.5, 2 bit curves) as noted in the legend.

Despite the uncommon approach of measuring CV confusions at MCL, we demonstrated, based on the low entropy in quiet, that audibility was not an issue. A token's audibility is not rigorously defined by the average speech spectrum and HL(f). These two incongruent measures have little to do with the audibility of the acoustic cues in

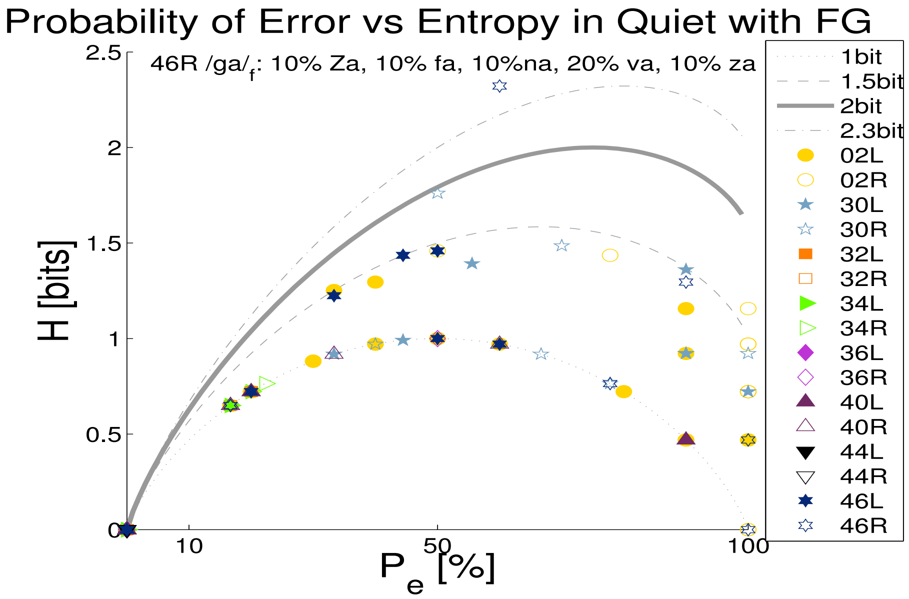


Figure 4.1: Token entropy $\mathcal{H}$ as a function of the token error P_e plot for all subjects and tokens in quiet (FG condition). Entropy is low, even though in many cases the error is high, thus audibility is not an issue. The 2-bit curve is a reasonable audibility threshold based on the Miller and Nicely (1955) confusion groups. According to this definition only the female token of /gɑ/ for subject 46R is not audible, indicated by the data point which lies well above 2-bit curve. For this point only, the confusions (i.e., /ʒ,f,n,v,z/ + /ɑ/) and their frequencies are displayed in a label. All the other tokens are audible for all subjects. Each subject symbol appears 28 times, i.e., once for each token.

speech (Régnier and Allen (2008)). Given the results of our CV recognition experiments, we propose the use of entropy to define audibility, as opposed to PTA and LTASS. The following reasons further support this proposal:

1. The LTASS is irrelevant when it comes to CV perception because CV cues are bursts or frequency edges (Régnier and Allen (2008); Li et al. (2012); Li (2010)), whereas the longtime speech spectrum is dominated by the vowels.

2. CV perception is binary: the acoustic speech cue can either be heard or not (Singh and Allen (2012)).

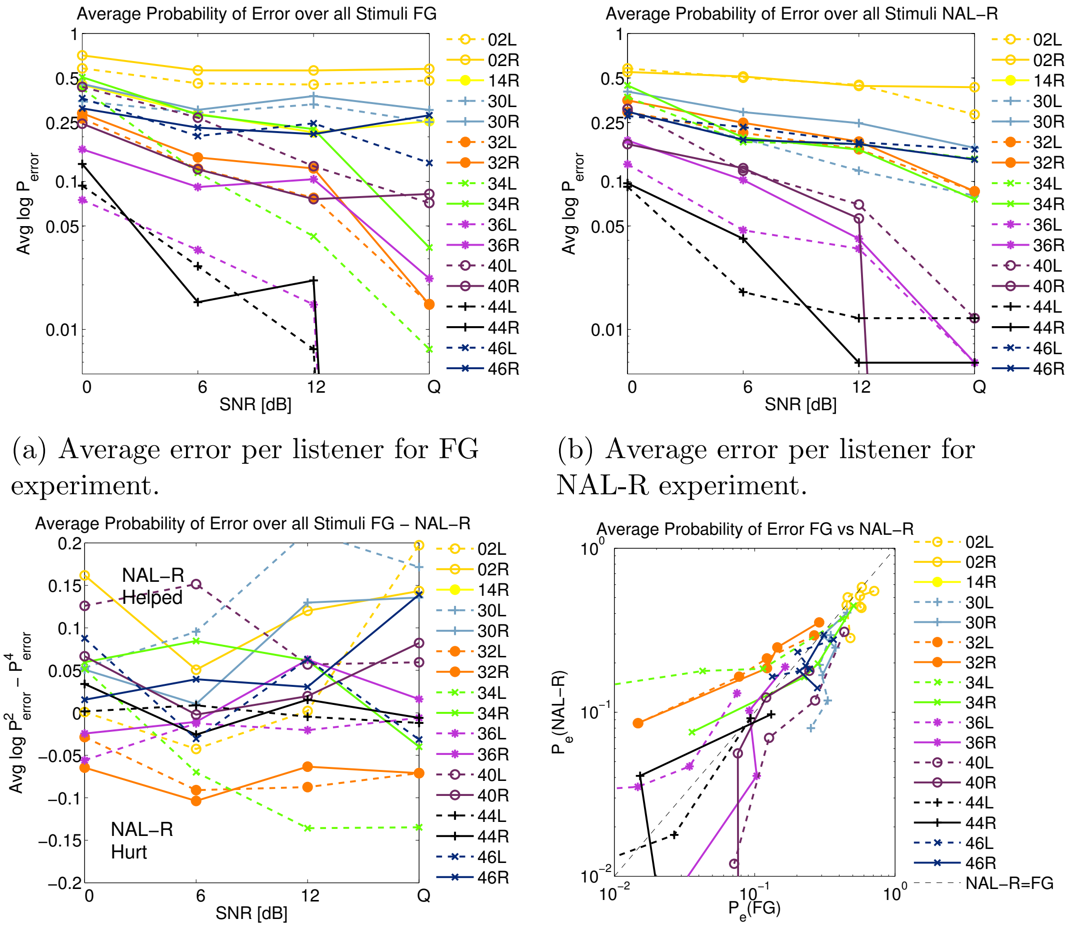
3. Natural wide-band speech is not tones. Thus PTTs do not characterize the audibility of acoustic speech cues as indicated by the 3DDS method (Li (2010)). Non-linear effects such as forward masking play an important role in the perception of speech cues (Wright (2004)).

From the reasoning stated above, it follows that a token with zero entropy, even with 100% error, must be audible. This is plausible since the ear must be listening to some signal properties, otherwise it could not be consistent. On the other hand, a listener who responds randomly across all 14 consonants has $P_e = 0.93$ (7% correct) and $\mathcal{H} = 3.8$bits, indicating the listener cannot hear the signal at all. The 7% correct represents chance performance in this example. The average size of the Miller and Nicely (1955) confusion groups (/p, t, k/; /b, d, g/; /f, θ, s, ʃ/; /v, ð, z, ʒ/; /m, n/) is three; therefore, a response with 3 equally likely responses (i.e. 2 confusions, resulting in $\mathcal{H} = 1.5$) is to be expected at the threshold of audibility. However, if a fourth response or even more also show, the token can be classified as inaudible. The subject is most likely guessing when confusions outside of a known confusion group appear. For example, if a 14-sided die always comes up on one of 3 sides, the die cannot be unbiased. That is, the token is audible if there is a significant bias. In Figure 4.1 (a) the 2-bit curve representing the proposed audibility threshold is plotted thicker.

According to the of audibility, of 2 bits of entropy, all tokens are audible for all subjects in the FG experiment, except /gɑ/$_f$ for 46R. The point belonging to this inaudible token is clearly above the audibility threshold (i.e., 2-bit curve) in Figure 4.1; it has an entropy of 2.32 [bit]. 46R responded 40% of the time correctly (i.e., /gɑ/). In addition, it confused /gɑ/$_f$ with /ʒɑ,fɑ,nɑ,vɑ,zɑ/ to various degrees (see Figure 4.1).

4.2 Top-down Analysis

A summary analysis of the data reveals what one would expect from NAL-R, namely that it decreases the average error. A repeated measure ANOVA (2 exps x 4 SNRs



(a) Average error per listener for FG experiment.

(b) Average error per listener for NAL-R experiment.

(c) Difference between the average error in the FG and NAL-R experiment.

(d) Average error of the FG and NAL-R experiment, plotted against each other.

Figure 4.2: Average error over all consonants: (a) Probability error averaged over all 14 consonants versus SNR for the Flat-Gain experiment. Note the log scale in (a) and (b) and the log-log scale in (d); also note that the quiet condition is plotted at the position of 18 [dB] SNR.

x 14 consonants) resulted in a significant ($p < 0.05$) difference in means between the two experiments ($F[1, 15] = 6.491$, $p = 0.023$) (Han (2011)).

Figure 4.2 compares the probability of error average over all consonants for each listener for the two experiments. The different ears vary in their error rates within an experiment (see Figure 4.2a and 4.2b for the flat-gain experiment and the NAL-R experiment respectively). In order to see the effect of NAL-R amplification, the error from the two experiments for each listener can be plotted against each other as shown in Figure 4.2d; there, P_{eFG} is the abscissa and P_{eNAL-R} the ordinate. It is clear from this analysis that NAL-R is not always lower in error. From this chart it is easy to identify which listeners benefit from NAL-R amplification. One may also see now that the error rates depend on the SNR for all listeners (monotonic function of SNR for all but 02R, 02L, 44R, 46R and 46L) in both Figure 4.2a and 4.2b. On the other hand, from the error difference plot (Figure 4.2c) we see that the differences (the impact of NAL-R) are relatively independent on the SNR. For example Ear 32R goes in quiet from 1.5% error (FG) to 8.6% (NAL-R) and is one of the four ears that on average does not benefit from the frequency dependent gain prescribed by NAL-R. Other ears that do not benefit from NAL-R are 32L, 36L and 34L. An ear that on average clearly benefits is 40L, which goes from an average error of 7.1% to 1.2% in quiet.

This findings are consistent with what was found by a first analysis of Han, who collected the data for the two experiments, in her thesis she states (Han, 2011, p. 61):

> Although NAL-R provides significant benefit on average, Exp. IV has uncovered many specific cases in which NAL-R fails, and in which adjustments in signal strength based on the CLP (Consonant-Loss Profile) would provide much greater benefit to HI patients.

In order to further investigate the effects of NAL-R, a categorization scheme taking into account both the error rates and the entropy is adopted (see Figure 4.3). As discussed in Chapter 3, both experiments have 4 SNRs, 14 CVs and 2 utterances per CV. Sixteen out of the 17 ears from the flat gain experiment returned for the

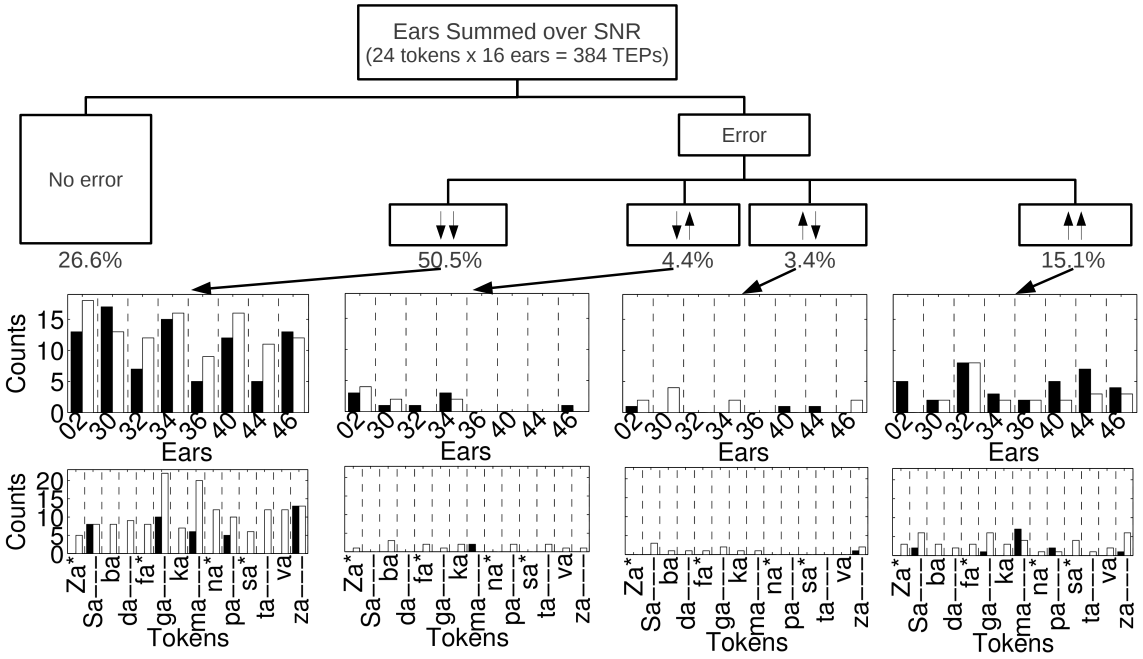


Figure 4.3: Categorization of the CV perception data for the 24 tokens and 16 listeners collapsed over four SNRs. For 102 (26.6%) of the 384 TEPs there was zero error in both conditions. The remaining 282 TEPs are grouped into one of 4 major categories; in the category labels the first arrow indicates what effect NALR had on the entropy, the second one indicates what happened to the error with NALR: (↓↓) (50.5%), (↓↑) and (↑↓) are small categories (4.4% and 3.4%)(↑↑) (15.1%). The histograms display the listener (top) and token (bottom) distributions. They show many of the TEPs in one category belong to a particular ear or token. The black bars represent the left ear and the male token, respectively, whereas the white bar represents the right ear and the female token, respectively. The * indicates the male token was excluded for the analysis (e.g., Za*).

NAL-R experiment. This gave us 3584 (2x16x14x2x4) different test conditions for which each was tested 5 to 10 times (see Table 3.3). For a comparison between FG and NAL-R, however, we must remove four tokens (/sɑ/, /ʒɑ/, /nɑ/, /fɑ/, since they were changed between the two experiments (see Table 3.4). This leaves 1536 test conditions that are the same in both experiments, each of which was again tested between 5 and 10 times.

In order to split up the test cases into categories, the data was collapsed across SNR, resulting in 384=1536/4 *token-ear-pairs* (TEPs). The classification of these TEPs is shown in Figure 4.3. In $\approx$ 27% of those cases the subjects did not make any errors, whether with flat-gain or with NAL-R. These $\approx$ 27% of the cases were classified as no-error sounds. They need no further analysis.

The remaining 73% can be further categorized. They are split up into four categories, dependent both on their token entropy $\mathcal{H}$ as well as token error. In Figure 4.3 in the category labels, the first arrow indicates what effect NALR had on the entropy of a TEP, the second one indicates what happened to the error with NALR. Most of the TEPs (50.5%) are the cases where NAL-R decreased both the entropy and error ($\downarrow\downarrow$). The second largest group is the one where NAL-R increased both the entropy and error (15.1%) ($\uparrow\uparrow$). The other two categories only contain the few remaining TEPs (i.e., 4.4% and 3.4%) (*cf* Figure 4.3). An additional analysis where the sounds are broken up into categories according to their clusters in both experiments can be found in Section 4.5.

As seen in Figure 4.2d, 32R, 32L, 36L and 34L are the subjects not benefiting from NAL-R. In order to see if those were the only subjects for which both entropy and error decreased with NAL-R "treatment" (i.e., $\downarrow\downarrow$), the histograms of the ear distributions are plotted on the bottom of Figure 4.3. It can be seen that all categories are approximately uniform across both listeners and tokens. This highlights the importance of not averaging. Even though most listeners benefit on average (Figure 4.2d), most of them seem to have specific problems with a few tokens.

One may wonder why the subjects 32R, 32L, 36L and 34L do not have significantly more TEPs in the $\uparrow\uparrow$ category but clearly show a bad impact of NAL-R on their average error in Figure 4.2d. The explanation is straightforward. The difference

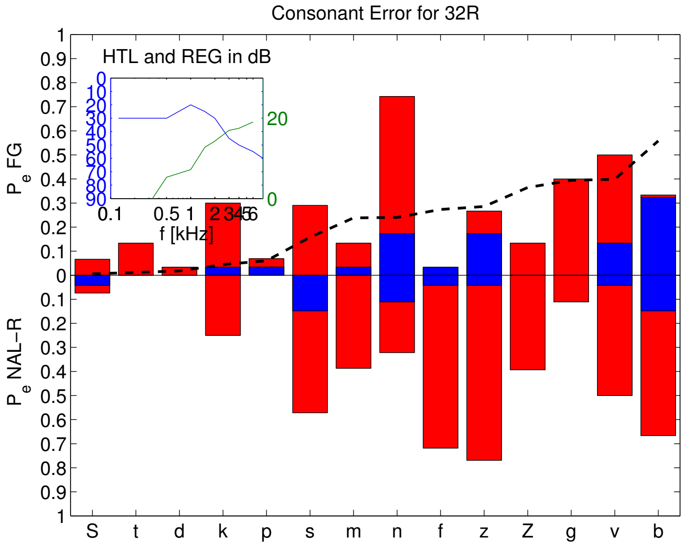


Figure 4.4: Difference between tokens for the two experiments for ear 32R. The audiogram and insertion gain for the ear are plotted in the upper left corner. The dashed black line indicates the average error across all subjects.

that NAL-R makes for the TEPs of those subjects in the ⇈ category is more severe. In order to verify the large change in error for a few tokens, error rates for both tokens and both experiments can be plotted (i.e., four error rates per consonant). The two tokens belonging to the same experiment are plotted on top of each other (Figure 4.4); the token with the higher error is always plotted in red and the one with lower error in blue. Since they are plotted on top of each other, the amount of red showing indicates the difference between the two tokens. The consonants on the abscissa are sorted according to the average error across all subjects. This allows one to identify sounds where the listener is performing differently than the group. The average error for all listeners for the flat-gain experiment is shown as a dashed black line. In the case of 32R shown in Figure 4.4, it can be seen that this ear has more difficulties than the average ear in identifying one of the /nɑ/ tokens in the flat-gain experiment. However, the NAL-R amplification fixes this problem and decreases the error on this token significantly. On the other hand, for the CV /zɑ/ ear 32R performs about average for both tokens in the flat-gain experiment. However, where one of the tokens gets better with the NAL-R amplification the other one increases significantly in error. The same is true to a lesser extent for /bɑ/, /sɑ/, and /fɑ/, which explains why the average performance as shown in Figure 4.2d gets worse. This illustrates that the problems are very specific. The same is true for subjects 32L, 36L and 34L, which is discussed in the results of the further analyses below.

4.3 $\mathcal{H}(P_e)$ Charts

Entropy does not take into account the specific confusions that were made, but only the number and distribution of confusions. The entropy with NAL-R goes down in many cases (see Figure 4.3). This means the responses with NAL-R generally show smaller confusion groups. To further investigate the effects of NAL-R, $\mathcal{H}(P_e)$ charts will be used. By looking at the change of $\mathcal{H}(P_e|SNR)$ for the two conditions NAL-R and FG, one can identify the effects of NAL-R. Ideally NAL-R should decrease the error while not increasing the entropy (i.e., $\downarrow\downarrow$ in Figure 4.3). A decrease in error combined with an increase in entropy (i.e., $\uparrow\downarrow$) is problematic since it means that the confusions become more random. A decrease in entropy with a constant error may be interpreted as an improvement, since it reduces the randomness of the answer.

In Figure 4.5 – 4.10 (in the right panel) where the error P_e is on the abscissa and the entropy $\mathcal{H}$ is on the ordinate, the listeners are encoded by the marker symbol (legend on the right). The SNR is indicated by the size of the symbol (legend in the plot). Furthermore, each symbol on the plot is accompanied by a number in a very small font, indicating how many confusion the listener made. The finely plotted lines on the graphs represent reference entropy. In the upper left corner of the plots, numbers indicate how many of the 68 or 64 points, for FG or NAL-R respectively, are closest to a specific reference line. By studying the distributions of these numbers, the effects of NAL-R on the entropy can be quantified. The labels next to the numbers in the upper left corner have the following meaning:

> **2bit curve:** Number of listeners that are closest to 2.25 bit or any higher bit curve.

2bit curve: Number of listeners that are closest to the third reference line — which stands for three equal likely confusion — a particular token right at a given SNR.

1.5bit curve: Number of listeners that are closest to the second reference line — which stands for two equal likely confusion — a particular token right at a given SNR.

1bit curve: Number of listeners that are closest to the first reference line — which stands for one confusion — a particular token right at a given SNR.

0% error: Number of listeners that recognized all the trails of a particular token right at a given SNR.

The $\mathcal{H}(P_e)$ charts will be accompanied by the so-called confusion bars (in the left panel in Figure 4.5 – 4.10). All the ears that participated in the experiments are on the abscissa. The probabilities of the possible 14 responses for the two experiments FG (on top) and NAL-R (on the bottom) are on the ordinate. Each listener was tested at four different SNRs; however, in order to simplify the plots the responses were collapsed across SNR. For each ear the total number of trials and the angle to the correct answer are indicated as a number in the confusion bar plots. If the right and the left ear of a subject participated in the experiments, the angle between the two ears of the same subject are displayed as a third number (second angle) above the confusion bar of the right ear.

For each token the following questions are discussed (the labels in parentheses will be used for each paragraph below to identify which question the particular paragraph answers):

Listeners How many listeners have sufficient errors to be taken into the analysis? What is the average error for the token? (Confusion bars)

Confusions Which are the main confusions for both experiments? Are they consistent across listeners? (Confusion bars)

Normal Hearing) Are the confusions that were made in the NAL-R experiment confusions that are expected (same Miller and Nicely confusion groups, expected from the normal hearing 3DDS data of the particular token)?

Entropy-Curves How many listeners are close to a certain entropy curve? How does the distribution of listeners change from one experiment to the other? ($\mathcal{H}(P_e)$ plot)

Ears For how many subjects are the two ears remarkably different when measured by the Hellinger distance between the two over SNR summed ears? (Confusion bars)

A summary of the results can be found in Table 4.1. The following pages will discuss the results for three consonants /k, b, s/ + /ɑ/; the rest of the consonants can be found in Section B.1.

4.3.1 f101ba

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /bɑ/ can be found in Figure 4.5.

Listeners 13 out of the 16 ears have sufficient errors to be taken into the analysis. The error on average is 37.9% and 35.9%, in the FG and NAL-R experiment respectively.

Confusions The main confusions for the female token of /ba/ in the FG experiment are /da/, /ga/ and /va/. /da/ occurs in all the responses of the listeners with error. Even though the degree of error varies across listeners, the confusions they are making are consistent. The same is true for the NAL-R experiment, the /da/ confusions get even more prominent and the entropy of the answers appears to be lower.

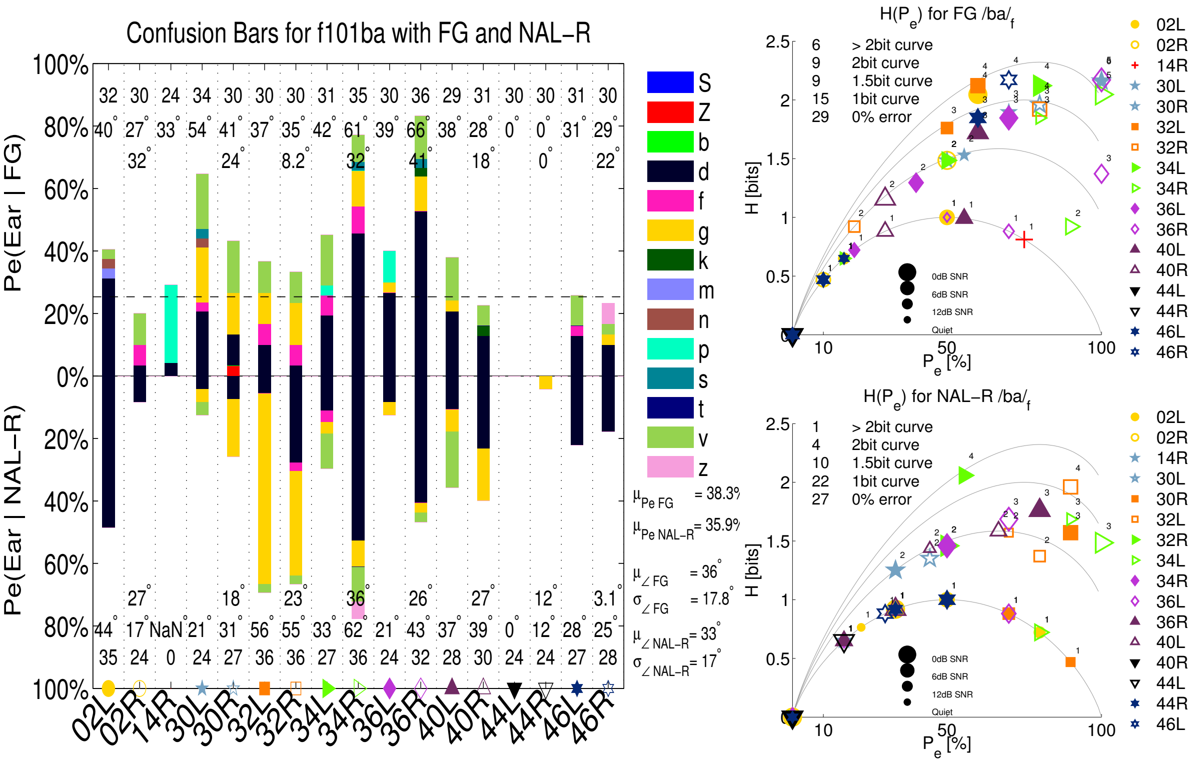


Figure 4.5: **Left: Confusion Bars** for the token f101ba. All the ears that participated in the experiments are on the abscissa. The probabilities of the possible 14 responses for the two experiments FG (on top) and NAL-R (on the bottom) are on the ordinate. Each listener was tested at four different SNRs; for this display the four SNRs are collapsed. For each ear there is a number indicating the number of trials. The other number displayed for only the right ear is the difference of the two ears of a subject measured by the cos direction, between the square roots of the two probability vectors. **Right:** $\mathcal{H}(P_e)$ **charts** show a symbol for every listener at four SNRs (64=16x4). SNR is coded by the size of the listener symbols. The charts also show four curves that the entropy follows if either 1,2,3, or 4 confusions are equally likely. The numbers in the top left corner indicate how many listeners are close to which of those entropy curves. The small number next to the symbols indicates the number of confusions present in the response. The top plot is always the plot for the FG condition, and the bottom plot is for the NAL-R condition. In the NAL-R condition points are expected to move down closer to the 1st entropy curve.

Normal Hearing The main confusions /da/ and /ga/ are expected; they are also plosives and their perceptual cues lay in the same time-region and only differ by frequency (/ga/ is a mid-frequency cue and /da/ is a high-frequency cue). Also the /va/ confusion makes sense, and was already observed by Li et al. (2010); they said:

> An especially interesting case is the confusions between /ba/ and /va/ (Fig 5)[1]. Traditionally these two consonants were attributed to two different confusion groups based on their articulatory and distinctive features. However, in our experiments, we find that consonants with similar events tend to form a confusion group. Thus /ba/ and /va/ are highly confusable with each other because they share a common F2 transition. This is strong evidence that events, not distinctive features, are the basic units for speech perception.

Entropy Curves The reduction in entropy is verified by the $\mathcal{H}(P_e)$ plots. The number of points closest to the third and above the third entropy curve reduce from 15 to 5. However, it should be noted that the NAL-R experiment has fewer points at zero error than the FG experiment.

Ears The two ears of the same listener are remarkably different ($\measuredangle > 30°$) for 3 listeners (02, 32, 36) in the FG experiment. In the NAL-R experiment only listener 34 shows big differences between the two ears. Listeners 02, 36, and 40 also show fairly large differences ($> 25°$).

4.3.2 m112ba

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /bɑ/ can be found in Figure 4.6.

[1] Fig 5 in Li et al. (2010) on p. 2606.

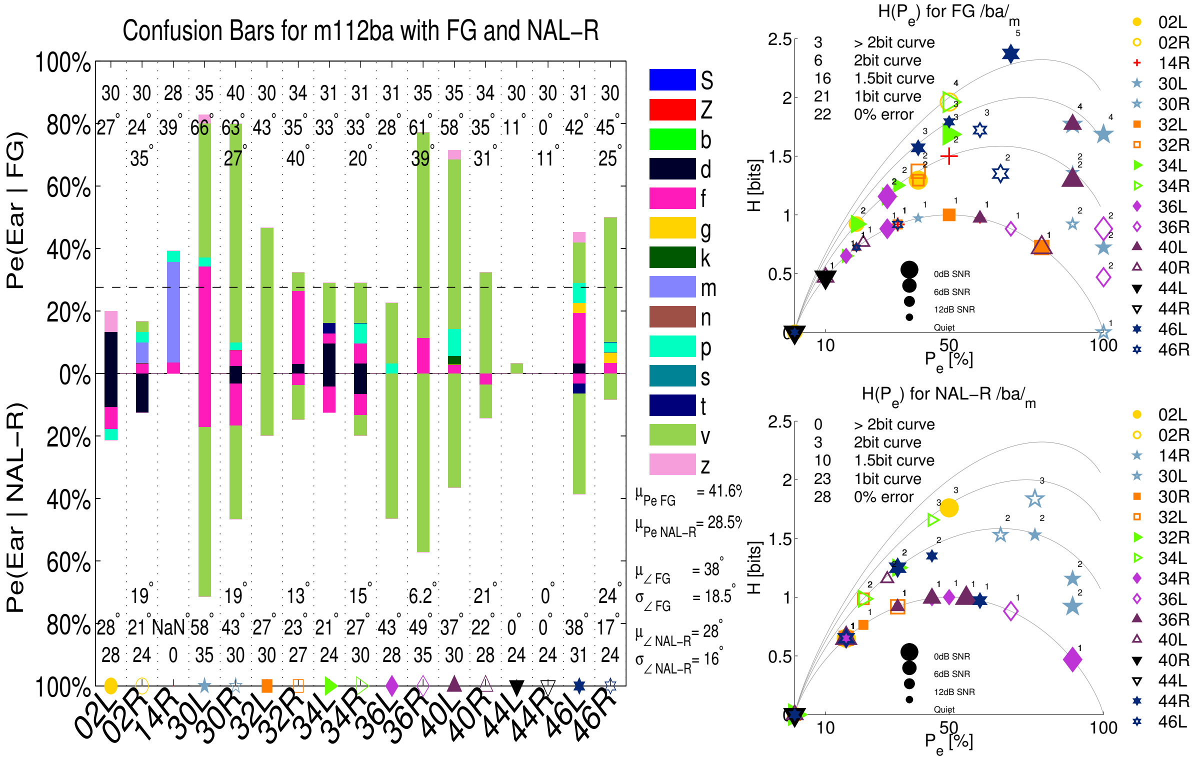


Figure 4.6: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m112/ba/. For a detailed description see Figure 4.5.

Listeners 13 out of the 16 errors have sufficient errors to be taken into the analysis. The error on average is 41.5% and 28.5% in the FG and NAL-R experiments, respectively. In comparison to the female token in the FG experiment the error is higher, but for the NAL-R experiment it is lower.

Confusions The main confusions for the male token of /bɑ/ in the FG experiment are /vɑ/, /fɑ/ and /pɑ/. /vɑ/ is by far the most likely one. Even though the degree of error varies across listeners, the /vɑ/ confusion is present in all the ears with errors except 02L. The same is true for the NAL-R experiment; however, many of the low–grade confusions do not show up anymore (entropy decreases), although the /vɑ/ confusions do get stronger. In the ears 36R, 36L and 40R the /vɑ/ confusions are the only confusions. Since the three ears all have relatively high errors ($> 40\%$), those three ears would probably benefit from training; the plasticity of the auditory system could help to overcome the problems for /bɑ/ for those listeners.

Normal Hearing The main confusion /vɑ/ as mentioned above is an interesting one, but one that is also observed in NH.

Entropy Curves The reduction in entropy is verified by the P_e vs. $\mathcal{H}$ plots. The number of points closest to the third and above the third entropy curve reduce to a third from 9 to 3. With NAL-R none of the points is above the third curve. However, it should be noted that the NAL-R experiment has fewer points at zero error than the FG experiment.

Ears The two ears of the same listener are remarkably different ($\measuredangle > 30°$) for three listeners (02, 32, 36) in the FG experiment. In the NAL-R experiment no listeners show remarkable differences between the two ears; for all the listeners the difference between the ears went down with NAL-R.

4.3.3 f103ka

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /kɑ/ can be found in Figure 4.7.

Listeners 9 out of the 16 ears have sufficient errors. The error on average is 26.1% and 27.6% in the FG and NAL-R experiments, respectively.

Confusions The main confusion is /tɑ/, and /pɑ/ also is a strong confusion. There are a few other confusions like /zɑ/, /fɑ/ and /gɑ/ but all to a negligible degree.

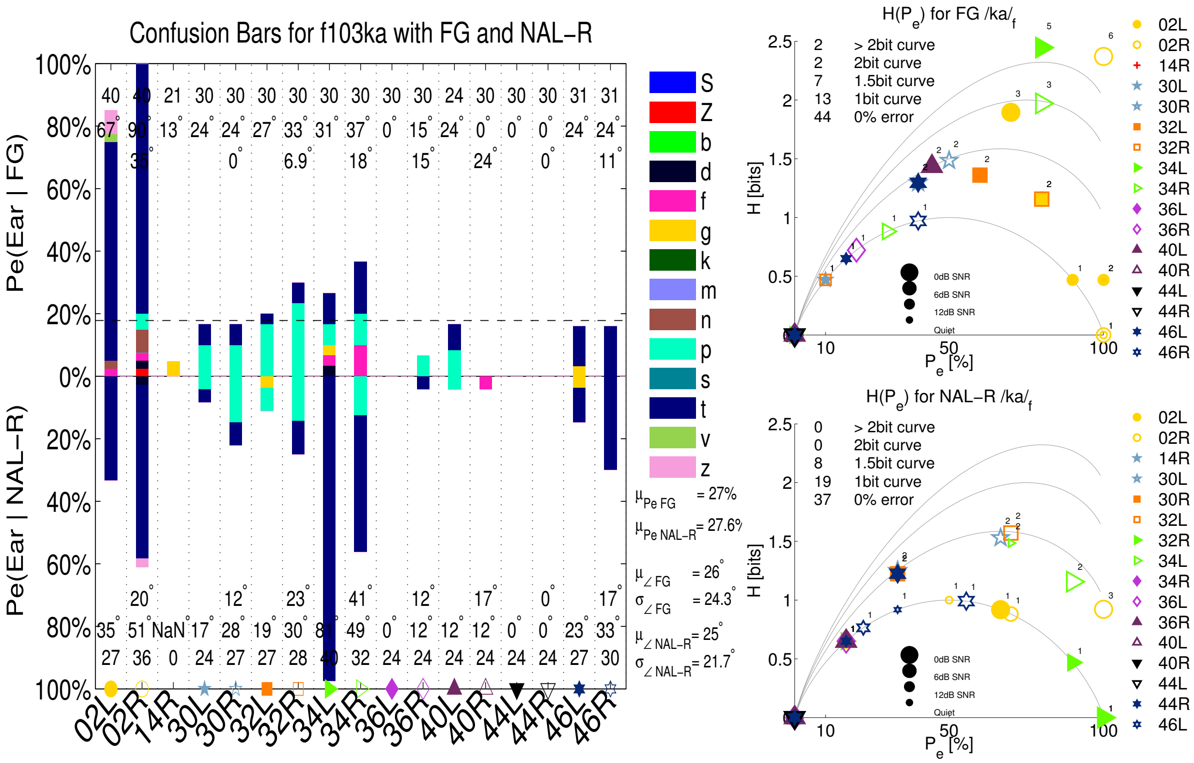


Figure 4.7: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f103/ka/. For a detailed description see Figure 4.5.

Normal Hearing /pɑ/, /kɑ/, and /tɑ/ form a common confusion group. Their cues lie in the same time region; they are distinguished by where the cue in frequency lies (/pɑ/ has a low frequency cue around 0.3-1 [kHz], /kɑ/ lives in the mid frequencies 1-2 [kHz], /tɑ/ in the higher frequencies 3-8 [kHz]). Confusions within this group are expected and also show up in the NH experiments.

Entropy Curves The distribution of the points on the curves shifts down. Many points move to the 1bit curve (number of points closest to the curve increase from 13 to 19). Also for 34L the entropy goes down, but the error goes up. 34L with NAL-R is fully convinced it hears /tɑ/ when f103ka is played. Also in the right ear of 34 the /tɑ/ confusion increases. Since both ears show the same phenomenon, plasticity may play a major role. This subject could greatly benefit from training. In general it can be said that the NAL-R confusions are all from the /pɑ/, /tɑ/, /kɑ/ group. More random confusions are eliminated. However, some of the /tɑ/ confusion might be introduced because of the high-frequency boost. For example for subject 46R the /tɑ/ confusion gets stronger with NAL-R.

Ears There is one subject for both experiments that shows remarkably different ears (02 for FG, 34 for NAL-R). 32R shows an interesting difference: the left ear has no error at all, whereas the right ear shows major confusions with /pɑ/ and /tɑ/. Interestingly, the degree of the confusions does not change significantly from the FG to the NAL-R experiment. The problem therefore seems to be in the outer periphery; however, the NAL-R compensation strategy does not seem to fix it.

4.3.4 m111ka

The confusion bars and the $\mathcal{H}(P_e)$ plots for m111/kɑ/ are shown in Figure 4.8.

Listeners This token is more robust compared to the female token, the error rates are smaller; 16.3% and 2.3% for the FG and NAL-R experiment respectively. Only 3 listeners have enough error to be taken into account for further analyses.

Confusions Even though the token is more robust, the confusions stay the same. The main confusion is /tɑ/. All the other confusions are negligible.

Normal Hearing This token in NH experiments shows confusions with /tɑ/ and also /pɑ/. The confusions seen in the HI data are therefore expected.

Entropy Curves The errors disappeared in most cases. Especially subject 02 shows large improvements. Again, as with the female token, NAL-R increases the /tɑ/ errors for 34L. Due to the larger /tɑ/ errors the entropy increases as well. Which can be seen in the NAL-R P_e vs $\mathcal{H}$ plot.

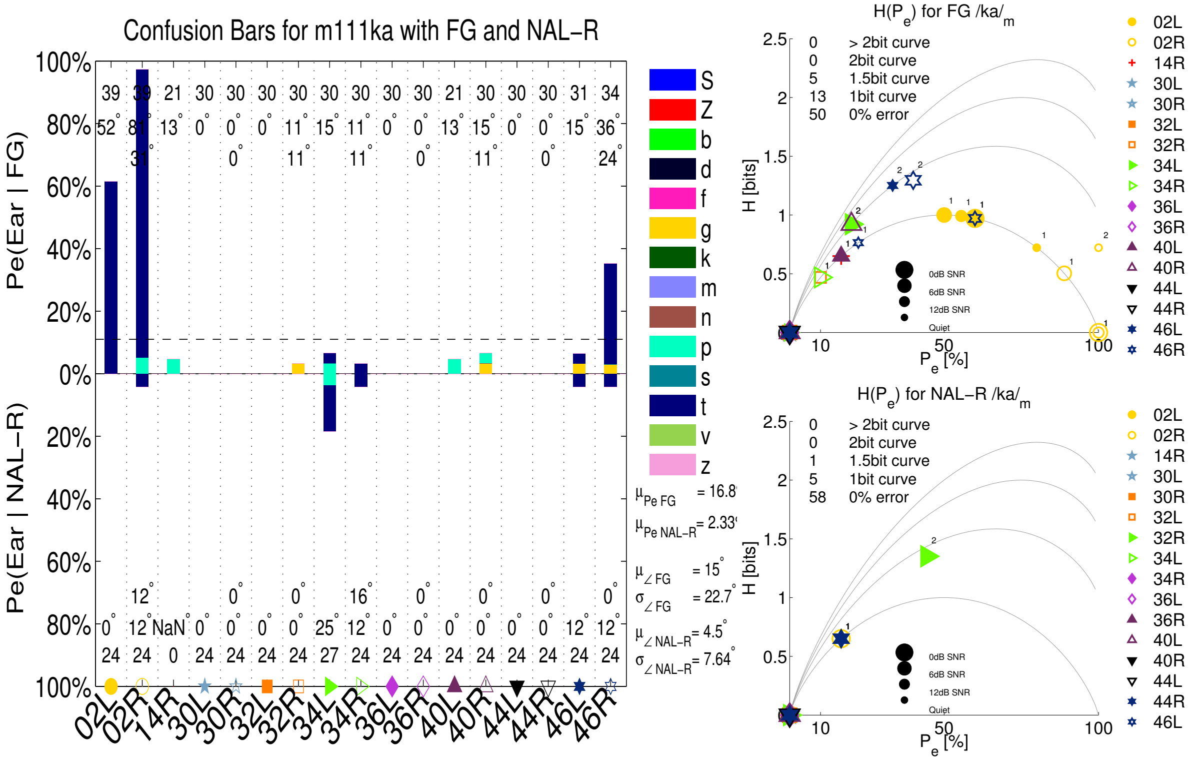


Figure 4.8: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m111/ka/. For a detailed description see Figure 4.5.

Ears For 46, the left has an advantage over the right ear in the FG experiment. NAL-R equalizes the two ears out. For 34 on the other hand NAL-R makes the two ears different by increasing the /ta/ confusion in the left ear.

4.3.5 f103sa

The confusion bars and the $\mathcal{H}(P_e)$ plots for f103/sɑ/ are shown in Figure 4.9.

Listeners 6 out of 16 listeners have enough errors for further analysis. The average error is 19.9% and 12.2% for the FG and NAL-R experiments respectively.

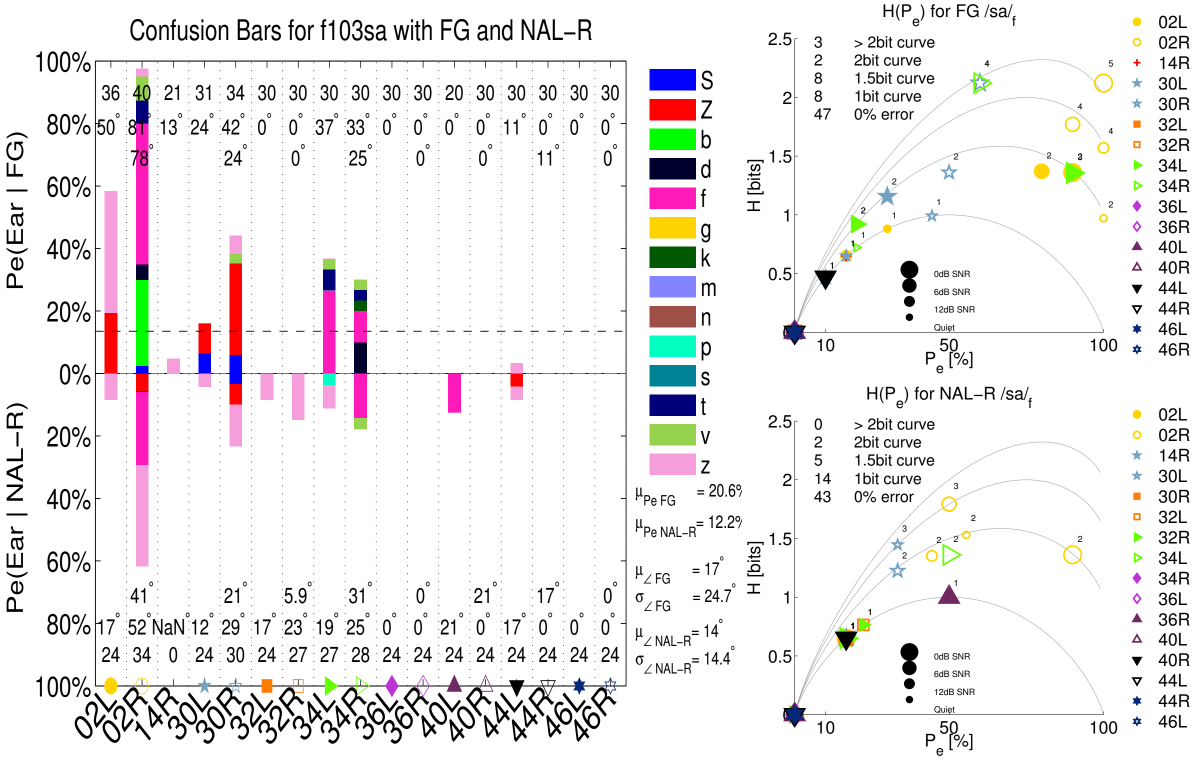


Figure 4.9: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f103/sa/. For a detailed description see Figure 4.5.

Confusions The main confusions are /fɑ/, /zɑ/ and /ʒɑ/. There are also /bɑ/, /ʃɑ/, /tɑ/, /vɑ/ and /dɑ/ confusions.

Normal Hearing What are the expected confusions for /sɑ/? Beren: /ʒɑ/,/fɑ/,/zɑ/ The confusions of the hearing impaired listener agree with the normal hearing data to a large extent; however, the hearing impaired add additional confusion such as /bɑ/, /ʃɑ/, /tɑ/.

Entropy Curves The error goes down, the unexpected confusions disappear. However, the expected confusions stay and compete, which leaves the entropy high.

Ears For the FG experiment the ears of 02 show different confusions, but also different error rates. They do not seem to share a single response. For the NAL-R experiment, 02 shows differences with the right ear having high error and the left ear low error; further differences can be seen for the subject 34, where the confusions despite NAL-R are different.

4.3.6 m120sa

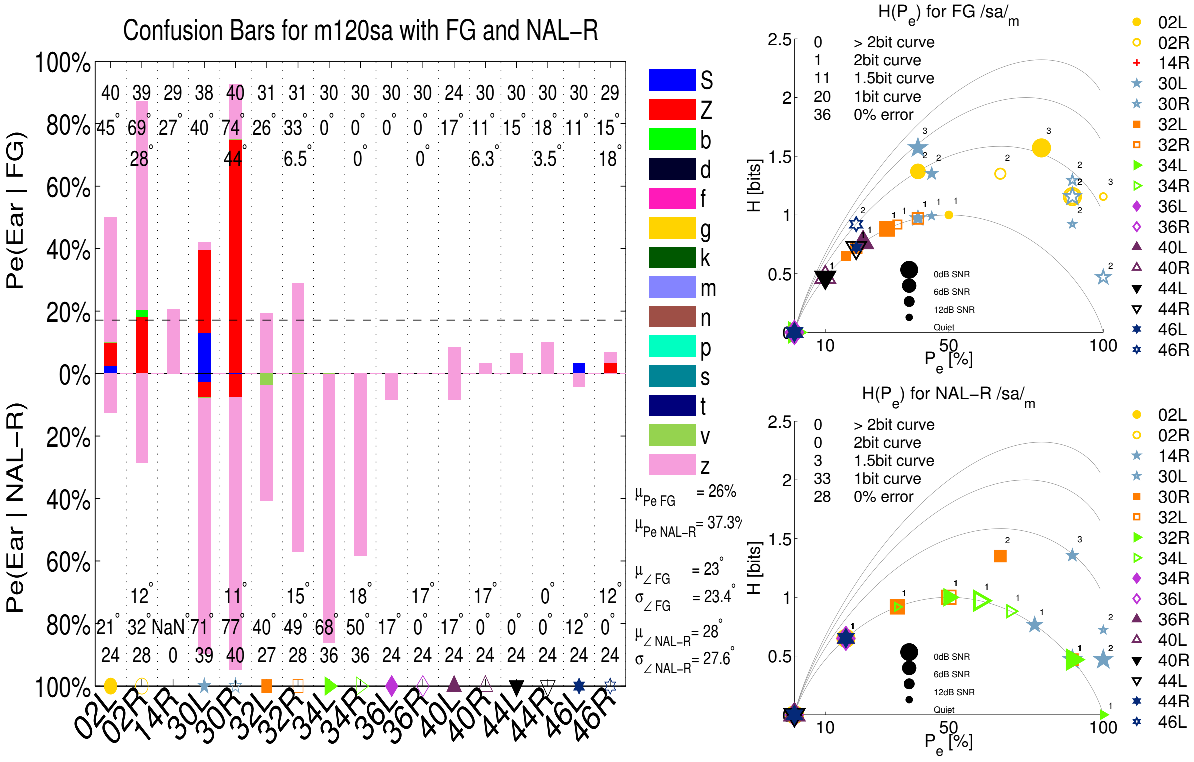


Figure 4.10: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m120/sa/. For a detailed description see Figure 4.5.

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /sɑ/ can be found in Figure 4.10.

Listeners 5 out of 16 listeners have enough errors for further analyses. The average error is 25.7% and 37.3% for the FG and NAL-R experiments respectively.

Confusions The main confusions are /zɑ/ and /ʒɑ/. There are also /ʃa/, and very few /bɑ/ confusions. Interestingly in the NAL-R experiment /za/ dominates and all other confusions are gone.

Normal Hearing /zɑ/ is expected.

Entropy Curves The entropy goes down; however, the error goes up in quite a few ears. Ears seem to be prone to respond with /zɑ/, even ears like 34L and 34R even though they did not show any confusions in the FG experiment start to make high /zɑ/ confusions.

Ears 02 and 30 are different in their error rates in the FG experiment. In the NAL-R ears become more similar.

Table 4.1: The *List* column shows how many of the 16 ears have enough error to be taken into account for further analysis. The *NALR* column shows what happened to the entropy: ↓ down, ↑ up. The symbol ˆ indicates that NALR reduced the entropy, yet it still remained high; “=” indicates no significant change. The *Ears* column shows how many out of the 8 listeners have ears that perform differently. $\bar{P}_e$ shows the average error.

Token	List	Conf (+ /ɑ/)	NALR	Ears /8		$\bar{P}_e$ (%)	
	/16			FG	NALR	FG	NALR
f109gɑ	14	/d, v, b, f/	↓ ˆ	0	0	46.9	36.1
m112bɑ	13	/v, f, p/	↓	3	0	41.5	28.5
f101bɑ	13	/d, g, v/	↓	3	1	37.9	35.9
f103mɑ	12	/v, n/	↑↓	2	2	26.7	18.8
f106zɑ	10	/Z, v/	=	3	3	34.4	28
f109fɑ	10	/s/	↓	1	1	31.4	18.9
m118zɑ	10	/Z, s/	↓	1	2	30.6	11.7
f101nɑ	10	/m, v/	=	1	1	17.3	5.8
f103kɑ	9	/t/	↓	2	2	26.1	27.6
f103ʃɑ	9	/s, z/	↓	0	0	9	9.5
f105ʒɑ	8	/z, S, g/	↑	2	1	40.1	32.6
f103pɑ	8	/t, k/	↓	1	0	23	20.8
f101vɑ	7	/m, f/	↓	2	1	27.4	20.3
m111gɑ	7	/d/	↓	0	0	21.5	21.4
f103sɑ	6	/f, Z/	↓ ˆ	1	3	19.9	12.2
m118pɑ	6	/t/	↓ ˆ	3	1	10.9	4.1
m120sɑ	5	/z/	↓	2	0	25.7	37.3
f105dɑ	4	/t/	↓	1	1	9.8	2.3
m111kɑ	3	/t/	↓	1	1	16.3	2.3
m118mɑ	3	/n/	=	0	0	9.2	2.6
f108tɑ	3	none	↓	3	1	8.7	1.6
m112tɑ	2	none	↓	2	1	5	1.3
m118ʃɑ	2	/Z, z/	↓	1	0	4.5	1
m118dɑ	1	/t/	↓	0	0	6.3	0.8

4.4 Impact of NAL-R on Average Hellinger Angles

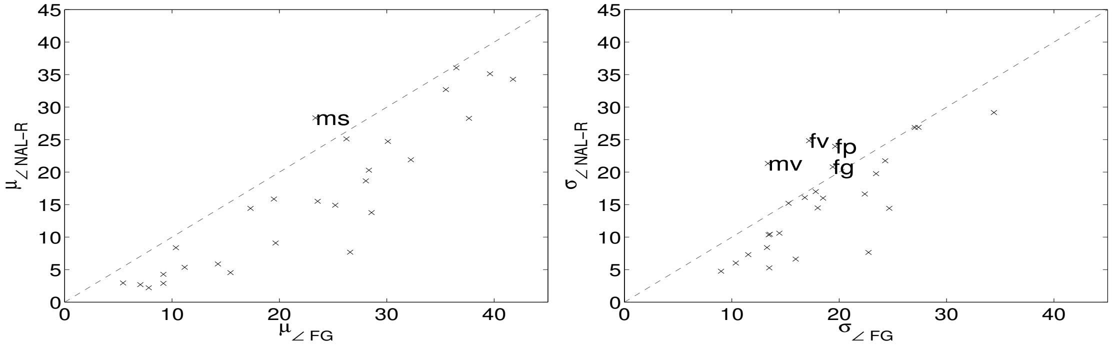


Figure 4.11: NAL-R decreases the mean angle between the correct response and the listeners' responses for all tokens except m120/sɑ/. It also decreases the standard deviation, and therefore makes the listeners' responses more consistent for all tokens but m120sɑ, f101vɑ, f103pɑ, f109gɑ.

Above, the main confusions are identified by hand; a more rigorous way to take confusion into the analysis is by calculating the cos-direction between the square root of the probabilities and the correct response. When people get more consistent in their responses, those angles should become more similar for all the listeners; therefore, their standard deviation $\sigma_{\angle}$ should go down. Also, if the responses become "better" the mean angle ($\mu_{\angle}$) of all listeners should get smaller. The scatter plots for the mean angle ($\mu_{\angle}$) and the standard deviation ($\sigma_{\angle}$) of the angle can be seen in Figure 4.11.

From the scatter plots it can be seen that the mean angle ($\mu_{\angle}$) goes down with NAL-R for 23 tokens; only for m120sɑ is the angle in the NAL-R experiment ($\mu_{\angle NALR}$ =28°) bigger than in the FG experiment ($\mu_{\angle FG}$ =23°). The variance of the angles ($\sigma_{\angle}$) also goes down in all but four cases: m120sɑ, f101vɑ, f103pɑ, f109gɑ.

4.5 K-means

Once a distance metric is defined in a vector space, the angle clusters may be properly defined.

For each token, there are 2x4x14=112 (2 experiments, 4 SNRs and 14 subjects) data points in the fourteen dimensional /p, t, k, f, g, d, b, s, z, ʃ, ʒ, m, n, v/ + /a/ space. The k-means algorithm is used to group the data points into $K = 4$ clusters. The choice of four is a choice made by the author. In some cases it turned out to be the wrong choice and will be pointed out in the text below. However, $K = 4$ often led to one cluster with low errors, two clusters grouped according to the two main confusions and a fourth one with a collection of outliers. The clusters are identified and displayed as stacked bar graphs. The $K = 4$ cluster means represent the centroids of the clusters and are also displayed as bar graphs. The color scheme is adapted according to the specific case, with white always representing the proportion of the correct answer. The confusions are plotted on a gray-scale gradient; the confusion that occurs the most often is black, and the less frequently a confusion occurs the lighter its gray-scale value becomes. If all the members of a cluster are plotted as stacked bar graphs, the color-scheme leads to the first cluster being almost entirely white. Given the white bars it is easy to visually underestimate the size of this low error cluster. Therefore the number of cluster members split up into the two conditions is displayed in the upper left corner. By comparing vector pairs for the same subject and SNR across the two conditions, it is possible to study the impact of NAL-R (see Section 3.3.6).

In the following three CVs (i.e., 6 tokens) are shown and commented in detail. For all examples, first the cluster means are displayed. The cluster means for the two tokens of a CV are displayed next to each other so differences in the grouping of the confusions are readily visible. The display of the centroids is followed by the display of the four clusters and their members; the caption of each cluster figure indicates how many members the cluster contains. These four plots are shown separately for both tokens. The text accompanying the figures answers the four questions asked in Section 3.3.6. The analysis should prove that clustering of confusions leads to meaningful results; it should furthermore provide insight into effects of NAL-R. For f103/sa/, for example, NAL-R makes a group of listeners perceive voicing where there is none. For each displayed case the numbers of members that fall into the category "Best" (B: $\mathbf{c}_x \rightarrow \mathbf{c}_1$, $x = 2, 3, 4$), (W: $\mathbf{c}_1 \rightarrow \mathbf{c}_y$, $y = 2, 3, 4$), "Neutral" (N: $\mathbf{c}_z \rightarrow \mathbf{c}_z$, $z = 1, 2, 3, 4$) or "Improved" (I) and "Degraded" (D) (I: $\mathbf{c}_x \rightarrow \mathbf{c}_y$ and D: $\mathbf{c}_y \rightarrow \mathbf{c}_x$, $x < y$) are displayed.

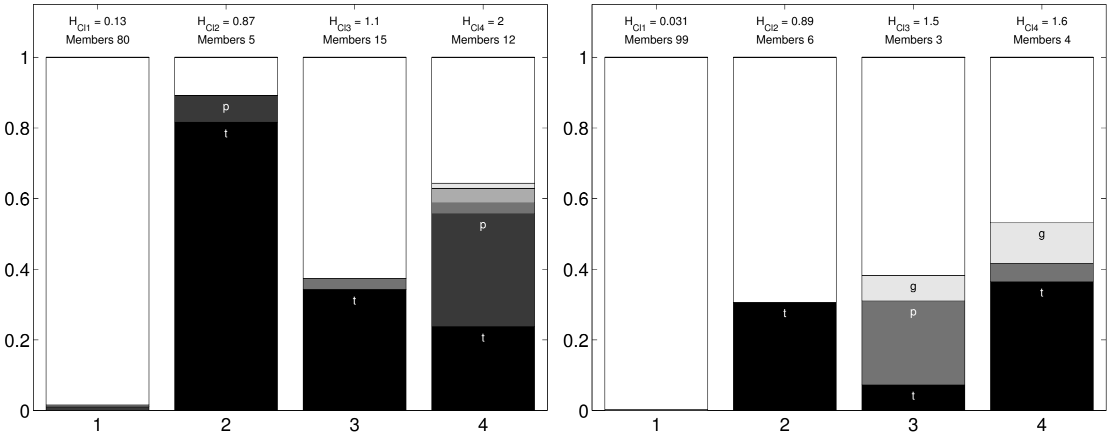
If the responses of the same listener at the same SNR in the two experiments differ only a little, they will be grouped into the same cluster and insignificant changes are thus eliminated. Therefore, this analysis can be seen as a statistically more relevant result of the top-down analysis in Figure 4.3. When examining all 1568 cases (4x14x24=1344), one can see that 191 cases (14.2%) fall into the "B" category and that in 76 cases (5.68%), NALR failed ("W" category). The "N" category contains 68.7% of the cases, "I" 9.2% and "D" 2.3%. This clearly shows that the large category ⇈ mostly consists of small improvements that do not differ much from the responses in the FG condition. Also the "W" category corresponding to ⇊ becomes smaller. The biggest category now is the "N" category in which both condition fall into the same cluster.

4.5.1 K-means Clustering: /k/

/kɑ/ (Figure 4.12) shows a difference between the two different tokens, but also shows the effect of NAL-R.

f103/kɑ/

The clusters show strong /tɑ/ and /pɑ/ confusions (Figure 4.12a). The second cluster is entirely made up by data points of the right and left ear of subject 34 under the NAL-R condition (Figure 4.13b). As can be seen in Figure 4.13b all the data points fall into the "D" (degraded) or "W" (worst) category. Three of the 5 members fall in the Worse (W) category; remember that means without NAL-R the subject performed well enough to be in the low-error cluster (cluster 1), but with



(a) Average confusion vector for all 4 clusters for f103/kɑ/. B: 4, W: 8, N: 42, D: 2, U: 0

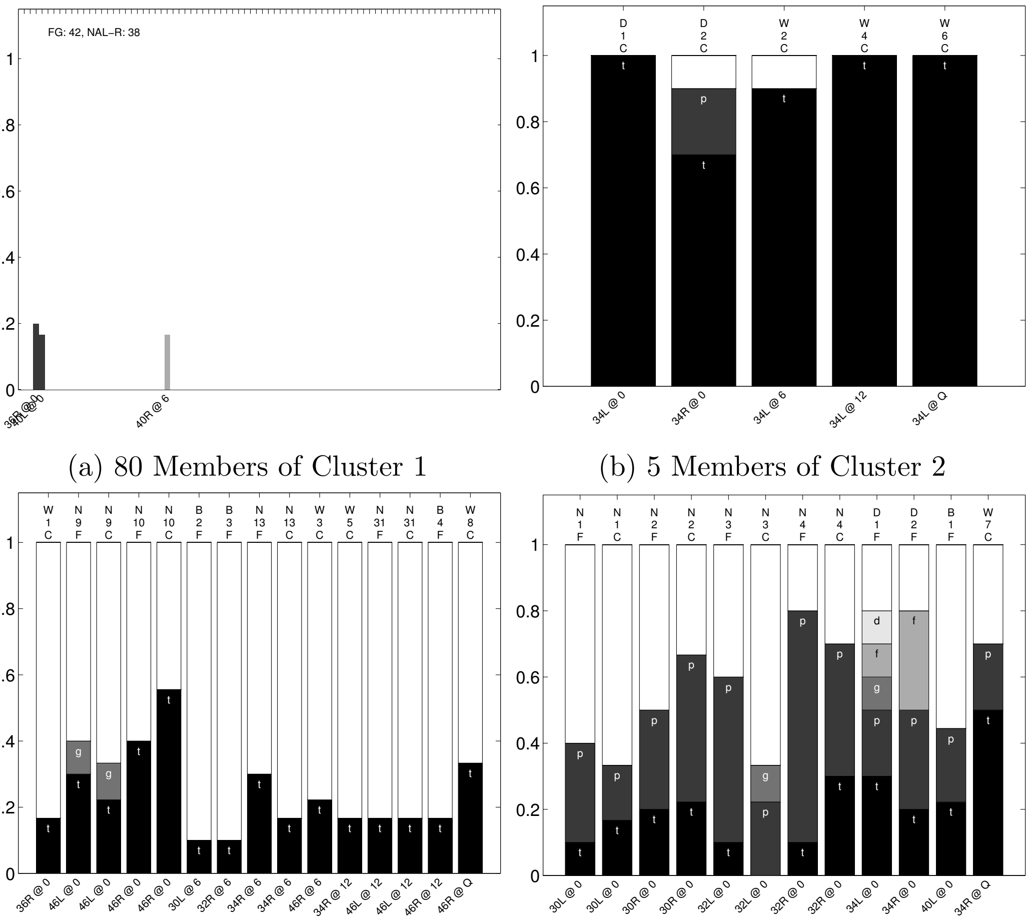
(b) Average confusion vector for all 4 clusters for m111/kɑ/. B: 6, W: 3, N: 46, D: 0, U: 1

Figure 4.12: The centroids resulting from the K-means cluster analysis for the two tokens of /kɑ/.

NAL-R the subject moved out of the low-error cluster. The error rates in cluster 2 (Figure 4.13b) are all greater than 90% and represent the greatest errors in the analysis. That means the immense /ta/ confusions in cluster 2 are all solely due to NAL-R. The two data points in the second cluster that are not in the W category come from the highest entropy cluster (Figure 4.13d), which means their entropy decreased but their error went up.

m111/kɑ/

The average error of this token is smaller and the /pɑ/ confusions are much less frequent; they only occur for four members (34L @ 0, 40R @ 0, 40L @ 6, 34L @ 0). As can be seen in Figure 4.12 (b) /tɑ/ is still the main confusion; it is the confusion causing data points to move out of the low-error cluster (W cases) (Figure 4.14b).



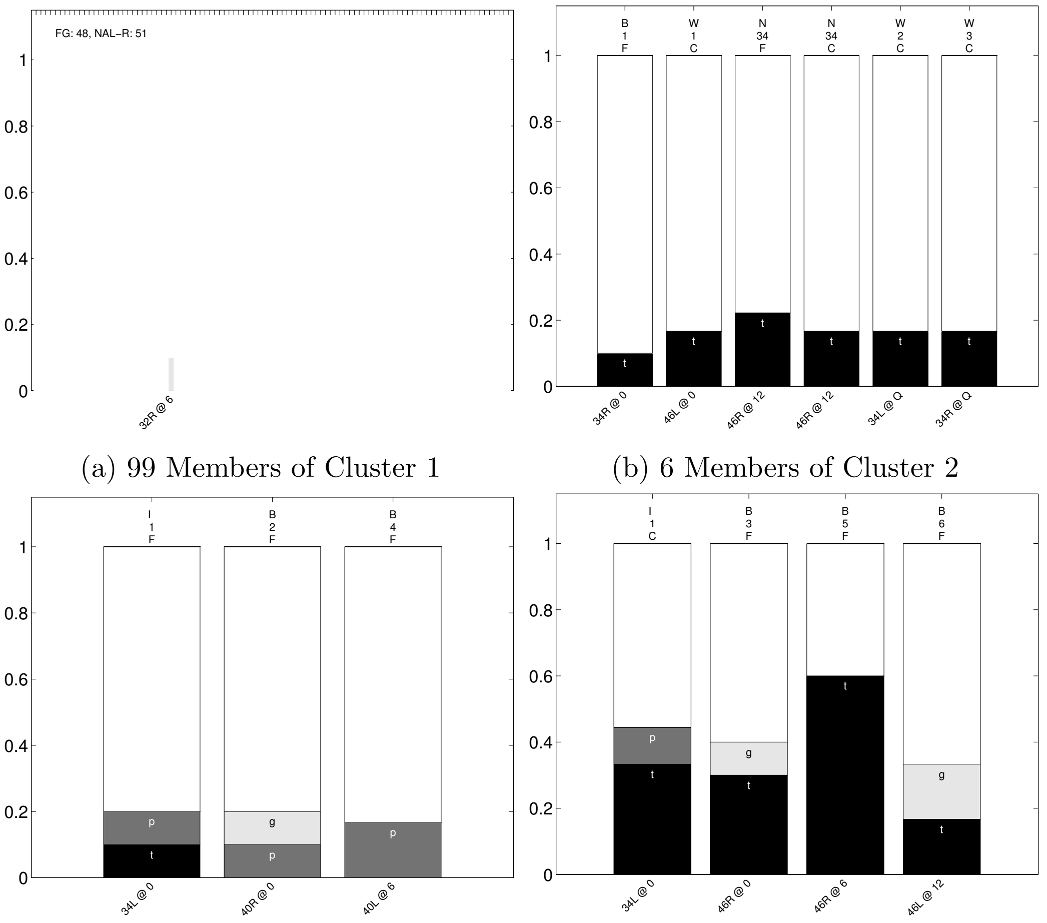
(a) 80 Members of Cluster 1

(b) 5 Members of Cluster 2

(c) 15 Members of Cluster 3

(d) 12 Members of Cluster 4

Figure 4.13: K-means clusters for f103/kɑ.



(a) 99 Members of Cluster 1

(b) 6 Members of Cluster 2

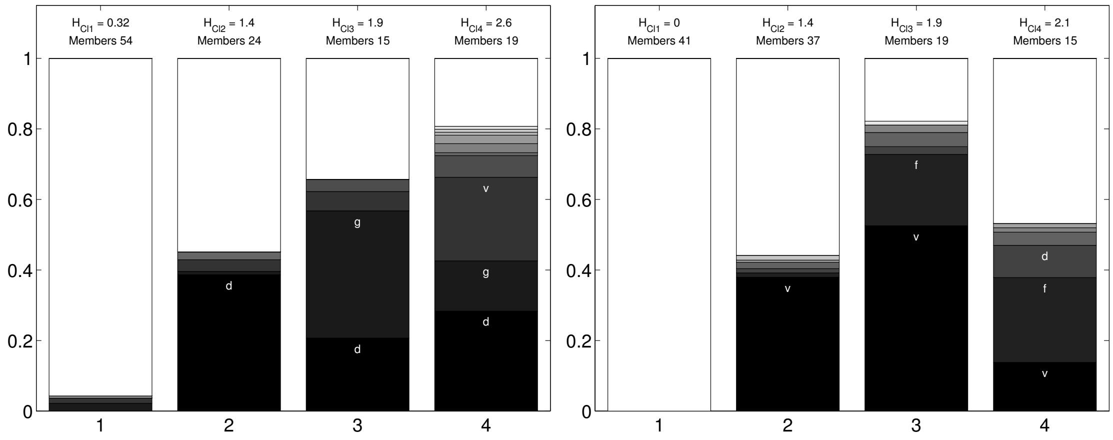
(c) 3 Members of Cluster 3

(d) 4 Members of Cluster 4

Figure 4.14: K-means clusters for m111/ka.

4.5.2 K-means Clustering: /bɑ/

/bɑ/ in general is a high-error token (i.e., the low-error clusters only have 54 and 41 members for the female and the male token respectively), which makes it an interesting case to examine in detail (Figure 4.15).



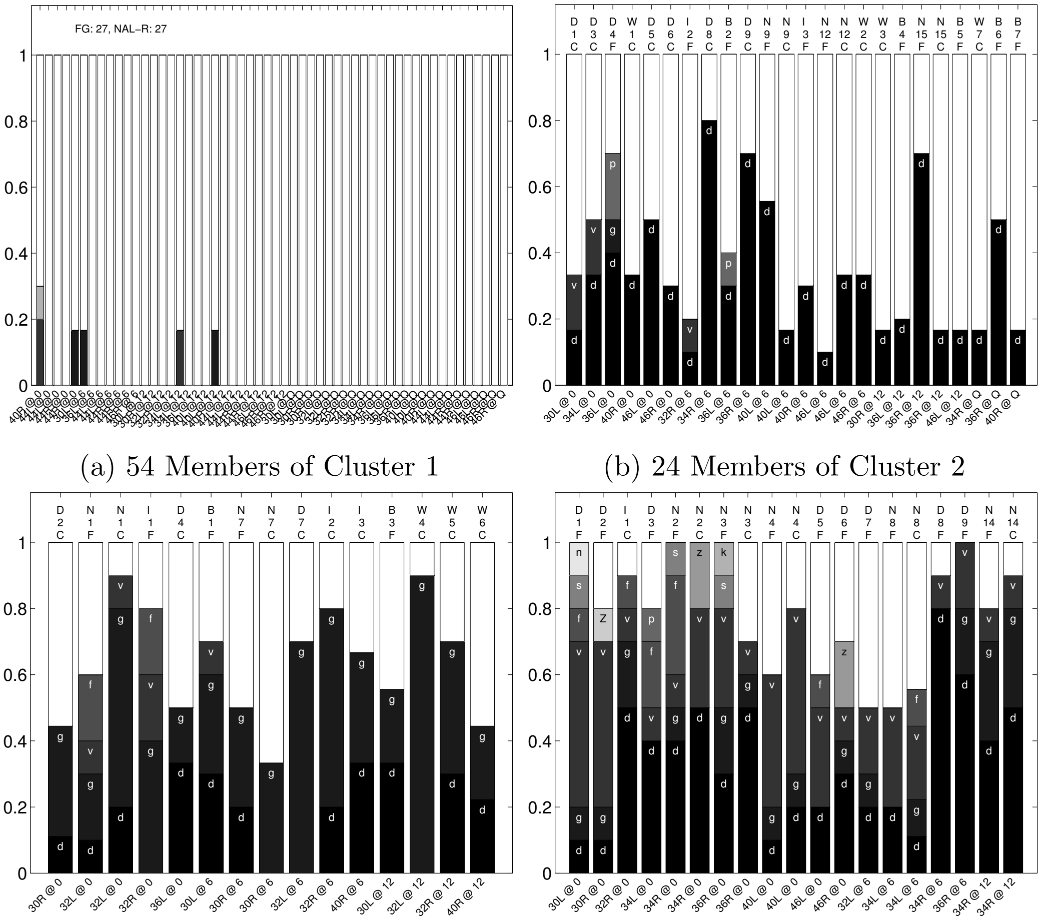
(a) Average confusion vector for all 4 clusters for f101ba. B: 7, W: 7, N: 30, D: 9, U: 3

(b) Average confusion vector for all 4 clusters for m112ba. B: 10, W: 1, N: 31, D: 11, U: 3

Figure 4.15: The centroids resulting from the K-means cluster analysis for the two tokens of /bɑ/.

f101/bɑ/

/dɑ/ confusions are most prominent, occurring to a significant degree in all three error clusters (Figure 4.15a). The first error cluster consists of responses with only /da/ confusions. The third cluster (Figure 4.16c) consists of subjects confusing /bɑ/ with /dɑ/ and /gɑ/; 10 out of the 15 members belong to the NAL-R experiment. Cluster 4 (Figure 4.16d) is a mix of cluster 2 and 3, containing subjects with confusions of /da/, /ga/ and /va/.



(a) 54 Members of Cluster 1

(b) 24 Members of Cluster 2

(c) 15 Members of Cluster 3

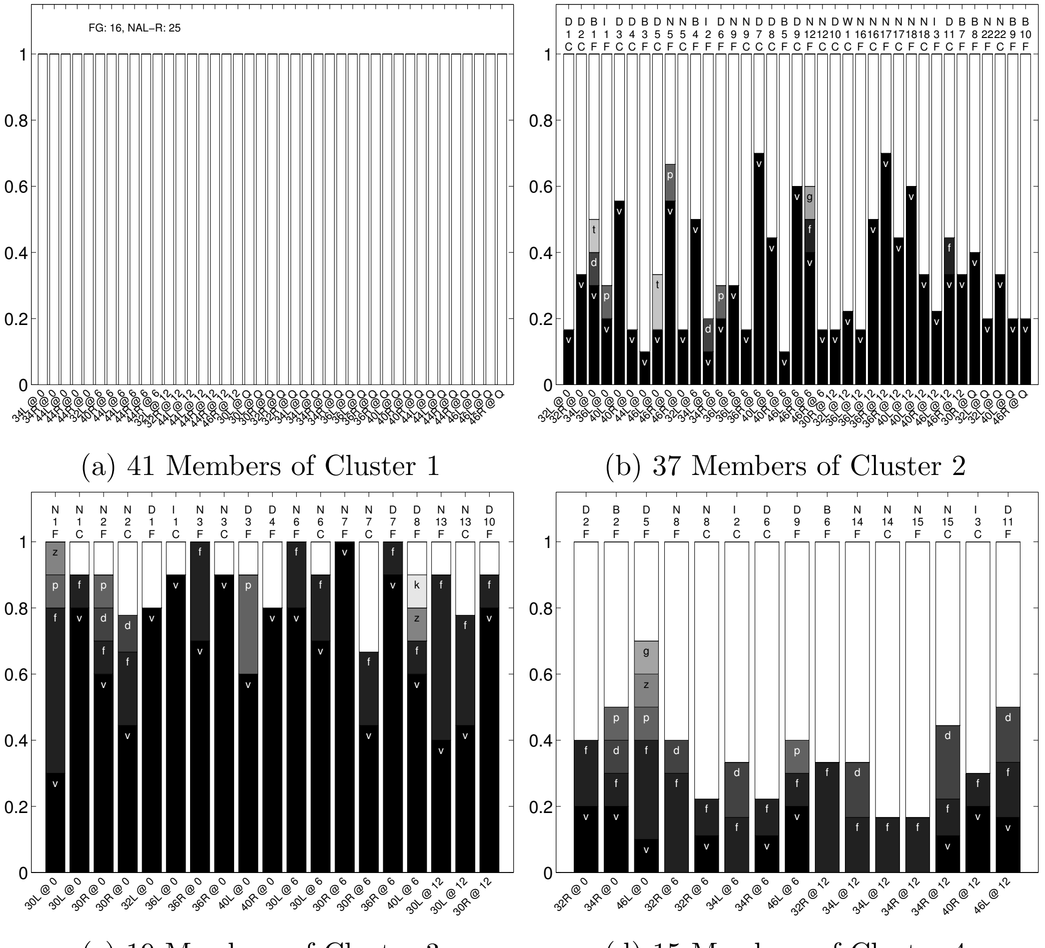
(d) 19 Members of Cluster 4

Figure 4.16: f101ba

m112/bɑ/

Whereas /dɑ/ and /gɑ/ were the main confusions for the female /bɑ/ token, /vɑ/ and /fɑ/ are the main confusions for the male token (Figure 4.15b). In both cluster 2 and 3 (Figure 4.17b and (c)), the confusions are almost exclusively with /va/; the difference between the clusters is the error rate. Cluster 2 has low error (< 50%) whereas cluster 3 (Figure 4.17c) has high errors (>80%). For all cases where NAL-R decreases the error for a FG member in cluster 3, the subject changes to cluster 2 and therefore stays consistent with the confusion but gets lower error. For one case (36L

@ 0 dB), the subject starts in cluster 2 with a 20% /va/ and 10% /pa/ confusion. NAL-R gets rid of the /pa/ confusion, but increases /va/ by 70%, certainly an undesirable effect of NAL-R. Cluster 4 (Figure 4.17d) contains cases with fairly low error (<50%); beside the /va/ confusion that is present in all clusters, cluster 3 has a strong tendency to /fa/.



(a) 41 Members of Cluster 1

(b) 37 Members of Cluster 2

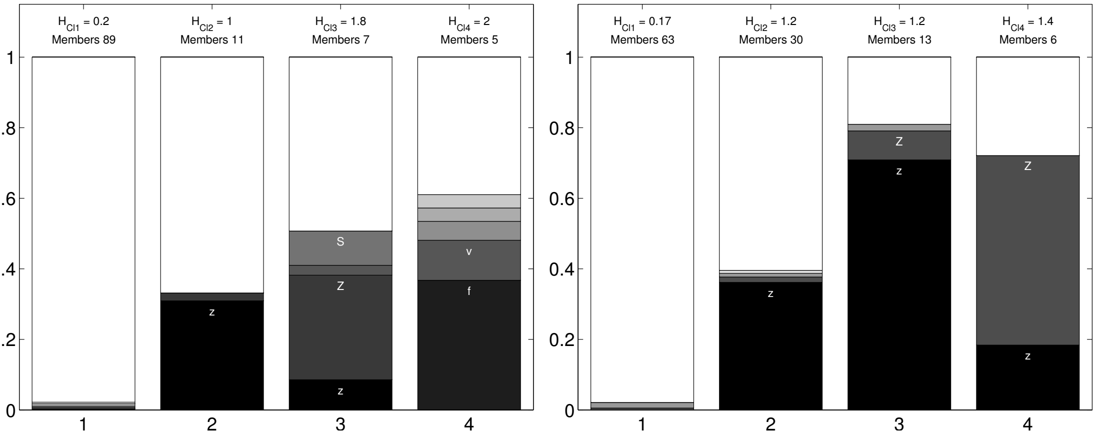
(c) 19 Members of Cluster 3

(d) 15 Members of Cluster 4

Figure 4.17: m112ba

4.5.3 K-means Clustering: /sɑ/

Both tokens have high /zɑ/ errors (Figure 4.18). Subject 30 with m120/sɑ/ also has high /ʒɑ/ errors. Subjects 30 and 34 seem to have more problems with /sɑ/ than the other subjects. It is an illustrative case of NAL-R making /zɑ/ confusion worse in 30 and 34.



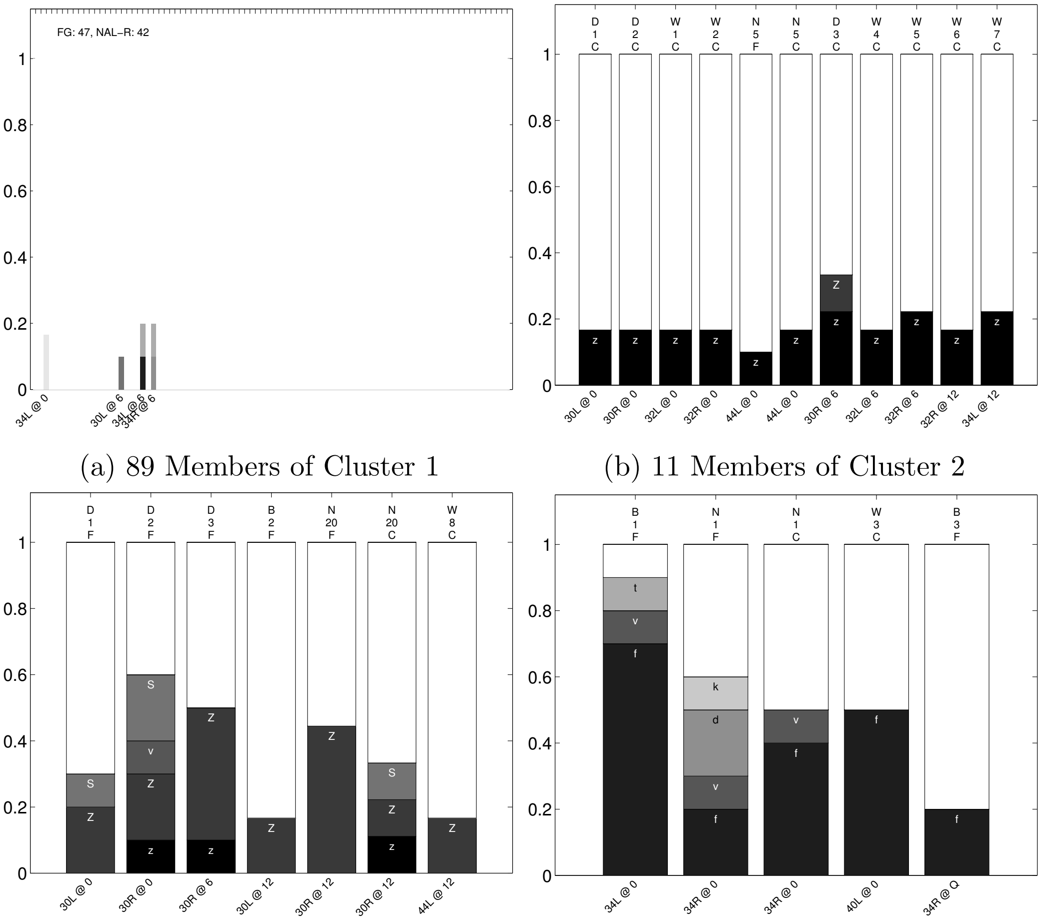
(a) Average confusion vector for all 4 clusters for f105Za. B: 7, W: 16, N: 25, D: 3, U: 5

(b) Average confusion vector for all 4 clusters for f105Za. B: 4, W: 1, N: 50, D: 1, U: 0

Figure 4.18: The centroids resulting from the K-means cluster analysis for the two tokens of /sɑ/.

f103/sɑ/

Both tokens show common confusions of /zɑ/ and /ʒɑ/ (Figure 4.18). The female token f103/sɑ/ also has a prominent /fɑ/ confusion (Figure 4.18a). The /fɑ/ confusion interestingly exists in both the FG and NAL-R experiment and only for two subjects (40 and 34) (Figure 4.19d). This confusion seems to be subject/token related and not due to NAL-R.



(a) 89 Members of Cluster 1

(b) 11 Members of Cluster 2

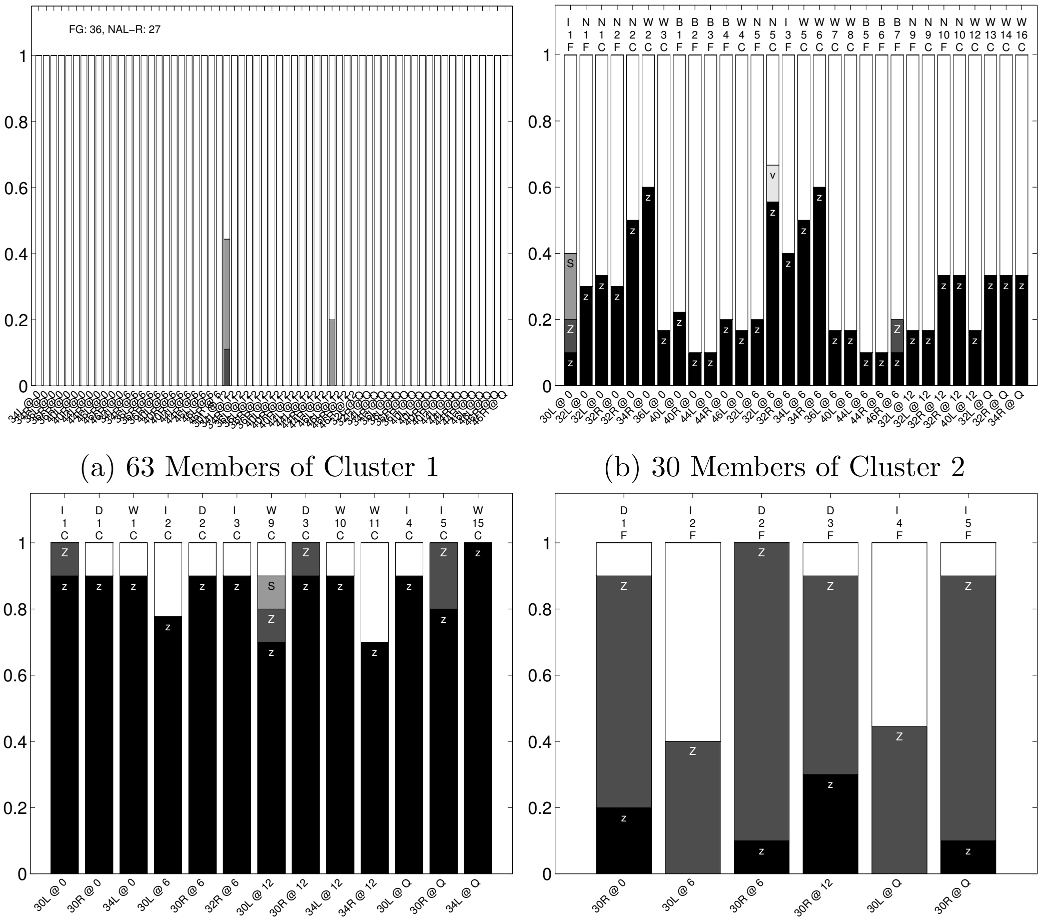
(c) 7 Members of Cluster 3

(d) 5 Members of Cluster 4

Figure 4.19: f103sa

m120/sɑ/

The male token of /sɑ/ shows strong /za/ confusions. Cluster 4 (Figure 4.20d) is only populated by subject 30 under the FG condition. The degree of /ʒɑ/ confusion is unique to subject 30. On the other hand, the strong /zɑ/ confusions in subjects 30, 34 and in one case in 32 are all in the NAL-R experiment (Figure 4.20). For those subjects NAL-R increases the /zɑ/ confusions. All the members from the highest entropy cluster (Figure 4.20d) are from the FG condition; they all change to the high-error /za/ cluster (Figure 4.20c) with NAL-R. The male token of /sɑ/ is another case where training could help, since /sɑ/ is constantly mislabeled as /zɑ/.



(a) 63 Members of Cluster 1

(b) 30 Members of Cluster 2

(c) 13 Members of Cluster 3

(d) 6 Members of Cluster 4

Figure 4.20: m120sa

Chapter 5

Discussion and Conclusions

With conventional (i.e., ANOVA, probability error P_e, entropy $\mathcal{H}$) and newly introduced methods (i.e., Hellinger angles, clustering analysis), it has been possible to investigate the impact of NAL-R on HI CV perception. In this chapter the findings are summarized and further discussed. Five main points of this document are:

1. Audibility is illogically defined in the existing literature: pure tone thresholds cannot predict consonant vowel (CV) recognition. Only CV recognition scores can tell us if speech is audible. Our proposed definition is to use an entropy measure to fix this shortcoming.

2. Based on our proposed entropy measure, testing the subjects at their most comfortable loudness (MCL) provided audible CVs.

3. NALR lowers the token entropy. According to our proposed definition, this means that NAL-R amplification at MCL makes the audible tokens more audible.

4. 15.1% of the time, NAL-R makes CV recognition (i.e. audibility) worse. Those cases are widely distributed over all tested tokens and ears, which means they represent very specific problems of individual ears to specific tokens. These cases are of particular interest, as they represent cases where NAL-R overamplified the speech. These tokens are cases where there is a need for improvement.

5. The Hellinger distance (HD) is a powerful new tool for the analysis of confusion

matrices (CMs). It allows one to characterize distances between responses of listeners, allowing one to cluster and visualize CM data.

5.1 MCL Testing

This study is unique in the way that it allows the subjects to adjust their listening level to MCL during the experiment. This is consistent with real-world conditions of hearing aid users, who may change the volume setting at any time. As can be seen in Table 3.2, the subjects generally chose a higher intensity for the NAL-R experiment, despite the spectral compensation.

Audibility is an improperly defined term in the literature in terms of the average speech spectrum and *hearing level* (HL). Such intensity measures are based on the assumption that the audibility of tones is equivalent to the audibility of speech cues, an assumption that has proved to be false. The audibility can only be measured via token speech scores. These have proven to be in disagreement with the HL(f) predictions. A common understanding is that speech audibility may be defined as that proportion of a long-term average speech spectrum (LTASS) which is above a subject's pure tone thresholds. It is likely that this misunderstanding is a result of the poor understanding of the prediction of the articulation index, which was designed to predict the average score in normal hearing listeners. It has been unclear how to modify the articulation index for HI ears. Hearing aids are fitted so that the *real-ear gain* (REG) of the hearing aid compensates for the differences of threshold and average speech spectrum. However, even if audibility is restored with an appropriate REG, audibility of individual sounds may not be guaranteed by this method.

We proposed a more rigorous definition which is based on entropy rather than on PTTs and LTASS. We believe that this provides a more meaningful definition, since it is also well defined in CV perception experiments, where the approach based on LTASS fails because it is mainly determined by vowel energy. As shown by the 3DDS method, for consonant perception, short bursts and onset times matter; these cues are not taken into account by the LTASS.

From the presented audibility definition based on entropy, it follows that a sound with 100% error ($\mathcal{H} = 0$ [bit]) must be audible. This consistency indicates audibility; while the listener does not hear the token correctly the response is consistent. Given the consistency, it can be assumed that audibility is not the main problem. On the other hand, a listener who responds randomly across all 14 consonants (e.g., $P_e = 0.93$ and $\mathcal{H} = 3.8$ [bits]) cannot hear the signal. Consistency is a proof of audibility. The average size of the Miller and Nicely (1955) confusion groups (/p, t, k/; /b, d, g/; /f, θ, s, ʃ/; /v, ð, z, ʒ/; /m, n/) is three. Based on this, a response with more than three confusions can be considered as inaudible. In a $\mathcal{H}$ vs. P_e plot an audibility reference line can be plotted by assuming equally likely confusions. In Figure 4.1 (a) the 2-bit curve representing the assumed audibility threshold is plotted thicker.

5.1.1 Impact of NAL-R

Generally speaking, NAL-R decreases the entropy (see *NAL-R* column in Table 4.1, Figure 4.3 and also, the k-means result). This effect means that, on average, NAL-R reduces the token confusion groups, which means that the ears become more consistent in their responses. This is consistent with the decrease in the standard deviation $\sigma_\angle$ (see Figure 4.11 on p. 58). That is, the token angles in the 14-dimensional space between the responses and the correct answer decrease for all ears. A third consistent measure is that the mean angle ($\mu_\angle$) of all ears per token decreases with NAL-R. In summary, NAL-R not only decreases the randomness of the answers, but also causes ears to agree more on a token.

This observation might provide new insight on how to train subjects with specific problems, since the confusion group gets smaller, which means they agree more on the signal they hear. Their hearing loss might amplify conflicting cues, such that the hearing impaired listener may hear the wrong cues.

We have demonstrated the effectiveness of NAL-R via a speech test instead of pure tone tests. This proves that speech can be used as a diagnostic tool, if not averaged across tokens. From Table 4.1, we know all the listeners for which CV tokens cause problems; thus we can get detailed information about their speech loss, which is poorly correlated with HL(f). Well characterized CVs must be used to find such specific problems in HI subjects (Singh and Allen (2012); Trevino and Allen (2013a,b)).

Our results show that in 15.1% of the cases NAL-R increases the entropy and error, denoted by $\upuparrows$ in Figure 4.3 on p. 41. Investigating such cases should help clarify how to improve prescriptive procedures (e.g., NAL-R). These cases are likely part of the reason why people are unsatisfied with their hearing aids as NAL-R makes 15% of the token-ear pairs (TEPs) worse. For patients with these complaints, an additional CV token speech test could provide valuable insight into the nature of their problems. Individual differences are always important in any diagnosis.

The Hellinger angle between the correct response and the subject's token response is an objective measure of their confusions. The mean angle ($\mu_\angle$) and the standard deviation ($\sigma_\angle$) of the angles for each token are expected to decrease once the ear becomes more accurate and more consistent in its response.

NAL-R has a significant impact on the standard deviation: a *paired t-test* results in $\alpha = 0.05 > p = 0.013$; in addition, the means of the two conditions are significantly different ($p = 2.0 \times 10^{-7}$). From the scatter plot in Figure 4.11, one can see that the variance of the angles ($\sigma_\angle$) decreases systematically with NAL-R in all but three cases: /va/$_f$, /pa/$_f$ and /ga/$_f$. The mean angle ($\mu_\angle$) decreases with NAL-R for all 24 tokens.

5.1.2 K-Means Algorithm

The impact of NAL-R is further supported by the k-means analysis, where listeners' responses are not collapsed over SNR, but rather are grouped according to their proximity in the Hellinger space. Having $K = 4$ clusters helps to come to a more meaningful result. If the responses of the same listener at the same SNR in the two experiments differ by a small angle, they will be grouped into the same cluster. In this way, insignificant changes may be eliminated.

When examining all 1568 cases (4x14x28=1568), one can see that 222 cases (14.2%)

fall into the “Best” category (i.e., NAL-R moved the listener into the best (low error) cluster), while in 89 cases (5.68%) NAL-R fails (“Worst” category, meaning NAL-R moved the listener into the worst cluster). The “Neutral” category contains most cases (68.7%). “Improved” (9.2%) means NAL-R moved the listener into a better cluster and “Degraded” (2.3%) means NAL-R moved the listener into a worse cluster. These numbers show that for a majority of the cases NAL-R failed to have a significant impact on the responses. For 14.2% of the responses NAL-R improved the scores. In only a small fraction of cases (5.68%) did NAL-R significantly decrease the performance. These are the most interesting cases which need to be further studied, since they represent cases where hearing aid fitting made the scores worse.

5.2 Speech as Diagnosis Tool

With the knowledge of the consonant cues (Figure 2.5a and 2.5b) and a test that is focused on natural speech without context, a detailed diagnosis of an individual’s speech hearing loss is possible. Thus we have demonstrated the effectiveness of NALR using a speech test to replace pure tone tests. This suggests that a carefully constructed speech test can be used as a diagnostic tool: With the results listed in Table 4.1, and the more detailed information that the confusion bars and entropy plots provide, we know exactly which CV tokens cause problems in which listeners. This gives us detailed information about their hearing loss. Thus carefully characterized CVs can be used to find specific problems in HI subjects, that PTTs cannot.

Two arguments for pure tones are often used: (i) pure tone testing is efficient in a clinical setting and (ii) it is universally applicable irrespective of *first language* (L1).

Argument (i) is easily addressed. With modern computer based testing, testing time is irrelevant; subjects can self–manage. Time invested during the accurate fitting of an expensive hearing aid is time well invested.

Argument (ii) is a valid argument. It is not argued that pure tone tests should be eliminated; rather, they have utility, as they easily identify problems in a basic way. They are simply not diagnostic.

Appendix A

Further Ways to Analyze CMs

In this appendix further ways to analyze CM data are described. Many different approaches have been tried by the author during the time working on this thesis. In this appendix, the probabilistic latent semantic indexing (PLSI) algorithm and its use to analyze CM data is described.

A.1 Probabilistic Latent Semantic Indexing (PLSI)

Probabilistic Latent Semantic Indexing (PLSI) is a clustering algorithm that optimizes a K-L divergence[1] objective function, just like other clustering algorithms such as NMF. NMF has been proven to often lead to better results than k-means due to the flexibility of NMF which has more parameters. NMF could also be used to cluster confusion matrices, but we will focus on PLSI here which can be seen as a probabilistic version of NMF, since it has the additional constraints to be probabilistic (the matrices are constraint to sum to one $\sum_{i,j} Xij = 1$). PLSI is one of the state-of-the-art unsupervised learning models in data mining, and has been widely used in many applications, such as text clustering, information retrieval and collaborative filtering Zhang (2012).

Latent variable model Probabilistic Latent Semantic Indexing (PLSI, Hofmann (1999)) has its origins in Latent Semantic Analysis (LSA, Deerwester et al. (1990)).

[1] During the work with the algorithm the idea developed to use the Hellinger distance instead of the K-L divergence. The Hellinger distance is a real distance and would therefore have an advantage over the K-L divergence. This idea, however, was not further pursued because of time constraints.

Given a probabilistic matrix X (i.e., $\sum_{i,j} X_{ij} = 1$), PLSI aims to obtain three non-negative matrices C, diagonal S and H such that CSH^T is the approximation of X. It therefore can be compared to a singular value decomposition (SVD), which is used to find the eigenvectors and values of a matrix. As with many other models PLSI model parameters are obtained by the Expectation Maximization (EM) algorithm. The algorithm iteratively increases the likelihood function until some convergence condition is reached; in the PLSI model the algorithm is constrained to maintain C, S and H probabilistic ($\sum_i C_{ik} = 1$, $\sum_k S_{kk} = 1$, $\sum_j H_{jk} = 1$).

For a simplified explanation, a document analysis task can be taken as an example. This was the original purpose of the algorithm. For the example, we assume the data can be expressed in 3 sets of variables:

- Documents: $d \in \mathcal{D} = \{d_1, \ldots, d_N\}$ - observed variables. N refers to the total count of documents in the analyzed collection.
- Words: $w \in \mathcal{W} = \{w_1, \ldots, w_M\}$ - observed variables. M is the number of distinct words in the collection of N documents.
- Topics: $z \in \mathcal{Z} = z_1, \ldots, z_K$ - hidden (or latent) variables. K has to be chosen in advance.

A set of documents can be represented as a matrix $X_{M \times N}$; each of the M words in the set of documents (i.e., the vocabulary) represents a column in the X matrix and each document a row; thus each element X_{ij} describes how often a word w_i occurs in a specific document d_j (X is sometimes also referred to as a bag of words). PLSI now tries to find three matrices C,H and S, such that X can be approximated by CSH^T. C_{ik} is the probability of $P(w_i|z_k)$ (z_k means the k-th hidden (latent) topic), H_{jk} is the probability of $P(d_j|z_k)$ and S is diagonal matrix with diagonal element $S_{kk} = P(z_k)$. The model is summarized in Figure A.1.

The EM algorithm maximizes the log-likelihood function of the PLSI model $L = \sum_{i,j} n(i,j) log P(wi, dj)$, where $n(i,j)$ is the co-occurrence number of word i and document j, and $P(wi, dj) = \sum_k C_{ik} S_{kk} H_{jk}$. X can be normalized to satisfy the likelihood function. With X we can re-write the log-likelihood function as

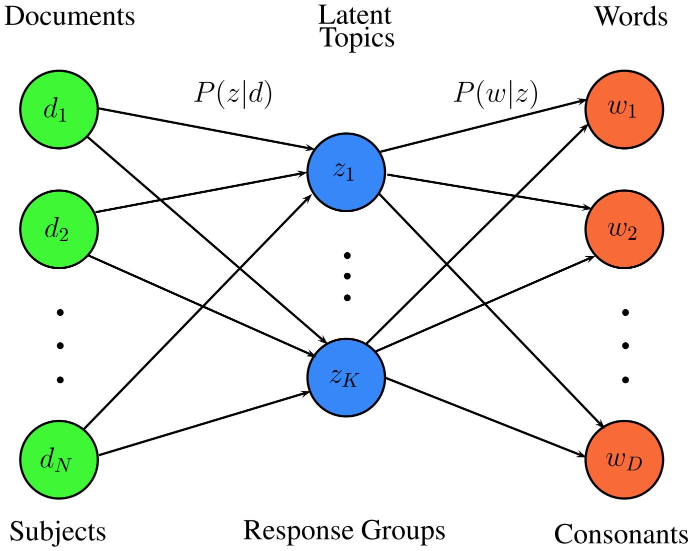


Figure A.1: The PLSI model and its generic structure. The layer of hidden (latent) topics are in the middle linking the documents on the left and the words on the right. This framework allows one to express a document as a mixture of topics expressed by the probabilities $P(z|d)$. The probabilities $P(w|z)$ indicate the frequency of expected occurrence for a word given a topic.

$$L = \sum_{i,j} X_{ij} log P(w_i, d_j). \tag{A.1}$$

The parameters C, S and H are estimated in every iteration of the EM algorithm. Some initial values for C, H, S are passed to the EM algorithm and it iteratively updates them according to the following formulas:

$$C_{ik} := \frac{\sum_j X_{ij} P_{ij}^k}{\sum_{i,j} X_{ij} P_{ij}^k}; \qquad S_{kk} := \sum_{ij} X_{ij} P_{ij}^k; \qquad H_{jk} := \frac{\sum_i X_{ij} P_{ij}^k}{\sum_{i,j} X_{ij} P_{ij}^k} \tag{A.2}$$

where P_{ij}^k is the probability of

$$P(z_k|w_i, d_j) = \frac{S_{kk} C_{ik} H_{jk}}{\sum_k S_{kk} C_{ik} H_{jk}} \tag{A.3}$$

By combining Equation A.2 and Equation A.3, the update equations for C, H and S are obtained Zhang (2012).

$$C_{ik} \ := \frac{\sum_j X_{ij} \frac{S_{kk} C_{ik} H_{jk}}{\sum_k S_{kk} C_{ik} H_{jk}}}{\sum_{i,j} X_{ij} \frac{S_{kk} C_{ik} H_{jk}}{\sum_k S_{kk} C_{ik} H_{jk}}} = C_{ik} \frac{\left(\frac{X}{CSH^T} H\right)_{ik}}{\left(C^T \frac{X}{CSH^T} H\right)_{kk}} \tag{A.4}$$

$$H_{jk} \ := \frac{\sum_i X_{ij} \frac{S_{kk} C_{ik} H_{jk}}{\sum_k S_{kk} C_{ik} H_{jk}}}{\sum_{i,j} X_{ij} \frac{S_{kk} C_{ik} H_{jk}}{\sum_k S_{kk} C_{ik} H_{jk}}} = H_{jk} \frac{\left(\frac{X}{CSH^T} C\right)_{jk}}{\left(C^T \frac{X}{CSH^T} H\right)_{kk}} \tag{A.5}$$

$$S_{kk} \ := \frac{\sum_{ij} X_{ij} C_{ik} H_{jk}}{\sum_k S_{kk} C_{ik} H_{jk}} = S_{kk} (C^T \frac{X}{CSH^T} H)_{kk} \tag{A.6}$$

Matlab has no built-in PLSI function. The code written for the analysis done in this thesis can be found in Section D.2.

Matrix factorization As indicated earlier with the comparison of PLSI to SVD, PLSI can be interpreted as a kind of matrix factorization. Our X (document-word) matrix is a large and sparse matrix. X has a column for each word in the collection of N documents and N rows. X is sparse because only a small number the whole vocabulary are used in each document. There are some common words that appear in every document (e.g., that, the, is); other words are topic specific (e.g. entropy) and only appear in a few documents. Given the sparsity of the matrix it is reasonable to assume that the dimensionality can somehow be reduced without losing much information. The same sparsity is also observed in confusion matrices. The matrices are sparse since a certain token produces confusions of only a small subgroup (confusion group) of the possible answers. X can be approximated ($\hat{X}$) by two low rank matrices L and R.

$$X \approx \hat{X} = L \cdot R \tag{A.7}$$

By looking at the size of the resulting matrices we can see the dimensionality reduction. L has a size of $N \times K$ and R is $K \times M$ ($K \ll M, N$). The inequality $N \cdot M \gg N \cdot K + K \cdot M$ holds true, and therefore the dimensionality is reduced. In addition if L and R are chosen correctly, they may reveal important information about the hidden structure of the data in X.

By using the EM algorithm, as described above, we get the following decomposition:

$$X = L \cdot U \cdot R \tag{A.8}$$

where the components have the following relations (Figure A.2):

- L consists of the probabilities for a document given a specific topic $P(d|z)$.
- U is a diagonal matrix. It contains the probabilities of the topics $P(z)$, namely how often does a topic occur in the set of documents.
- R contains the word probabilities $P(w|z)$.

All the three matrices are non-negative and normalized, since they represent probability distributions.

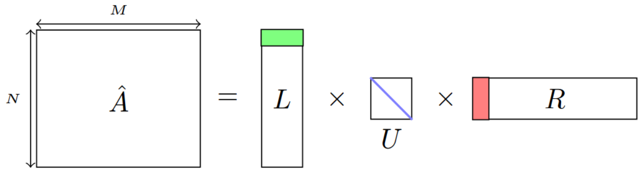


Figure A.2: The PLSI algorithm can also be understood as a matrix decomposition. In the figure X denotes the sparse matrix with documents as rows and words as columns. Each row in the L matrix (for example the green row shown in the picture) represents the probabilities of a specific document belonging to one of the latent topics $P(d|z)$. The blue diagonal contains the prior probabilities for the topics $P(z)$ (i.e. how likely is a topic to occur). The columns of R (for example the red one) consist of the probabilities of a word occurring given a certain topic $P(w|z)$. Applied to a CM the model links subjects, response groups and consonants.

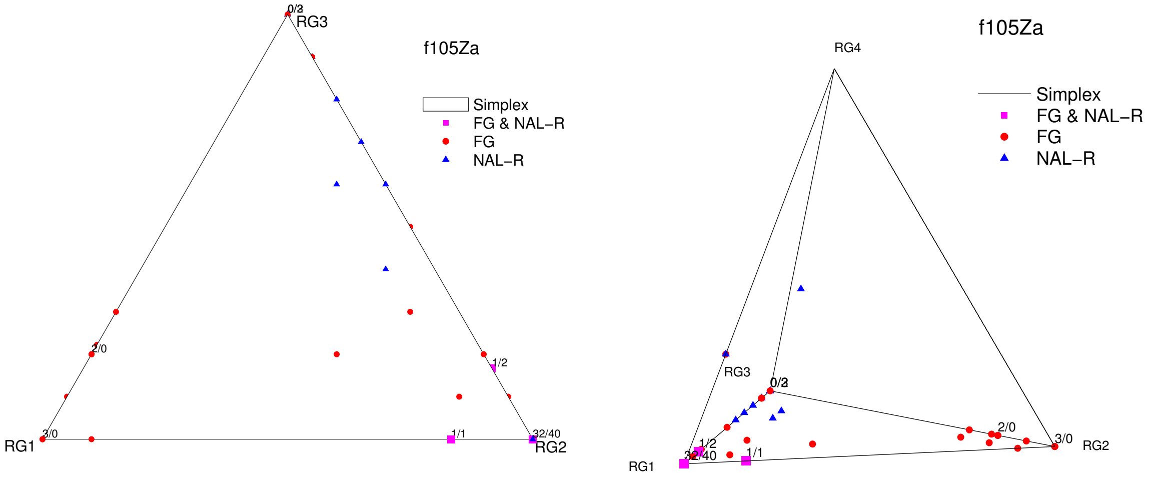
PLSI applied to CMs For a specific token, the document-word matrix X can be replaced by a subject-consonant matrix A. Each row of A consists of a subject's responses to a specific token. Each column of A is one of the possible answers (consonants). As in the document-word example above, A is going to be sparse. Only certain confusions will appear for a given token and subject. The matrix factorization will give us

$$\hat{A} = S \cdot C \cdot R \tag{A.9}$$

where the components have the following relations (Figure A.2):

- S consists of the subject probabilities $P(s|g)$. $P(s|g)$ is the probability of a subject s behaving according to a certain response group g.
- C is a diagonal matrix containing the prior probabilities of the confusion groups $P(g)$. It tells us which response group is the most important in explaining the data.
- R contains the components of the response groups. $P(c|g)$ is the probability of a consonant belonging to the response group g.

Since the probabilities summed over all response groups $G = \{g_1, \ldots, g_N\}$ are equal to 1 ($\sum_G P(s|g) = 1$), they lie on a simplex. If N is chosen to be smaller than or equal to 4, the simplex can be displayed and each listener can be displayed as a point on the simplex, as shown for $N = 3$ in Figure A.3a and $N = 4$ in Figure A.3b.



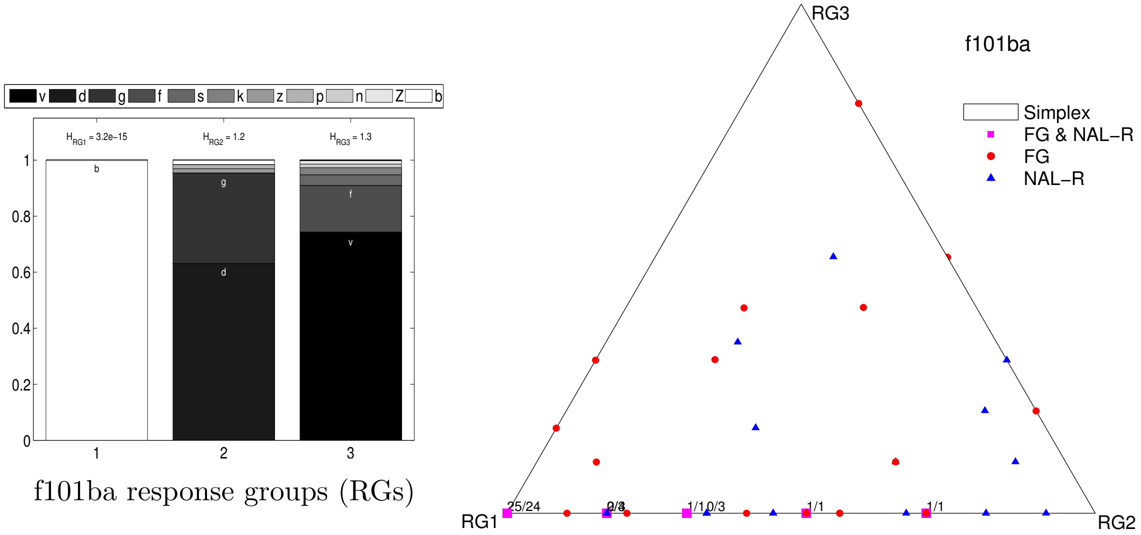
(a) 3 confusion groups

(b) 4 response groups

Figure A.3: The decomposition of the subject-consonant matrix in response groups allows one to plot high-dimensional data. Examples are shown for /ʒa/. In the decomposition in three response groups it can be seen that response group 1 (RG1) is only populated by listeners from the FG experiment, whereas RG2 is populated by both listeners from the NAL-R experiment and the FG experiment. With an additional degree of freedom (i.e. an additional response group), as shown in (b), the points are further spread out and better separated. RG4 is important in describing the NAL-R confusions: Since the RGs are sorted according their entropy, RG4 is the response group with the highest entropy. Thus NAL-R seems to increase the entropy for this token.

A.1.1 PLSI

The PLSI algorithm, as described in Section A.1, is a powerful tool that allows one to display the high-dimensional confusion matrix data and find response groups (RG) in the data without any *a priori* information. The number of RGs needs to be chosen by the user. If this number is smaller than 4, a visualization is possible. For the result section the number of RGs is kept to 3, since it allows one to visualize the data in an easily interpretable 2D plot. However, it should be noted that the number of response groups should be chosen according to the data and not the visualization. Some tokens build a small number of RGs, whereas others generate answers with a higher entropy and therefore need more RGs. The power of the algorithm can be demonstrated by comparing two examples, /bɑ/ and /sɑ/.

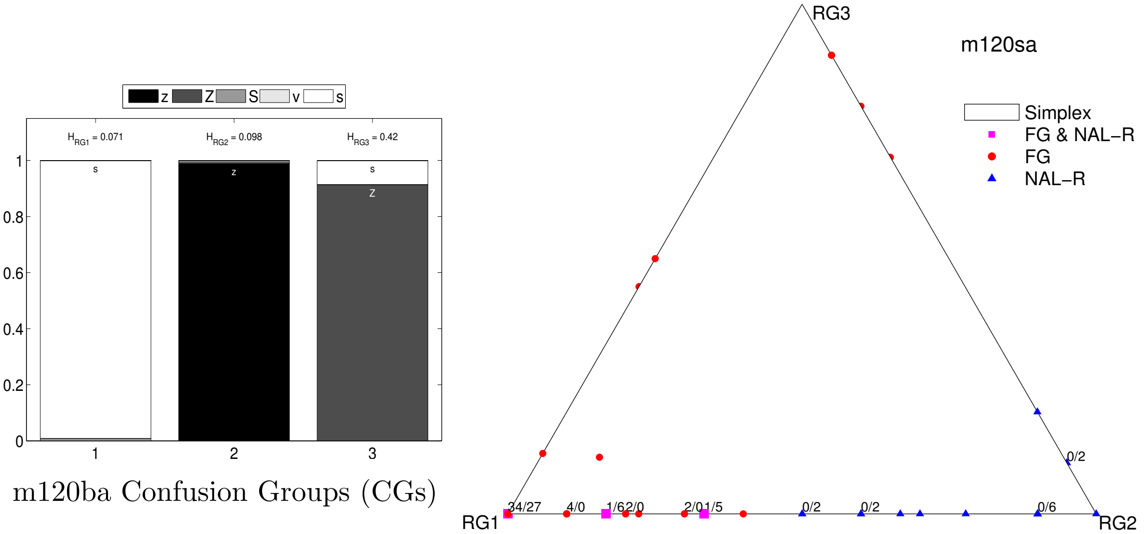


f101ba response groups (RGs)

Figure A.4: Results of a PLSI clustering for $/\text{ba}/_f$. (a) shows the three response groups found by the algorithm. (b) displays all the data points for the token as a mixture of the response groups of (a).

For f101ba the algorithm identifies the correct answer /ba/ as its own group (see Figure A.4 (a)). 25 out of the 56 points in the FG experiment fall right on this vertex of the simplex and can therefore be explained by only this group (the same is true for 24 out of the 56 points in the NAL-R experiment). Many of the remaining points are on the edge between RG1 and RG2. Therefore their answers consist of mostly the right answer /ba/ and the two main confusions /ga,da/. The third response group is dominated by /va/, but contains many small confusions as well (and therefore has the highest entropy). Most of the data only uses a little bit of the high-entropy component in their mixture. The display shows that the data for the two experiments are intertwined. According to the confusions no clear distinction between the two experiments can be made.

Another example where the data is better separable and where the effects of NAL-R are more readily seen is m120sa as shown in Figure A.5.



m120ba Confusion Groups (CGs)

Figure A.5: The PLSI algorithm with $N = 3$ groups the CM for m120sa into three meaningful response groups; one contains the correct answer, the second contains the main confusion /zɑ/, and the third group mostly containing the other major confusion /ʒɑ/.

The NAL-R points cluster around vertex RG2. Confusion group RG2 is dominated almost entirely by the /za/ confusion. Points on the edge from RG1 to RG2 therefore either get it right or make a /za/ confusion. The farther away they are from RG1, the more in error they are. Further, it is noticeable that the points that use RG3 in their mixture are all but three from the FG experiment. That means the entropy in the FH experiment is higher. Most points are dominated by two RGs, which means that the RGs chosen by the algorithm actually categorize the responses well.

Building groups and averaging in the case of m120sa is feasible, whereas averaging in the scattered case of f101ba should be avoided.

It should be noted at this point that, since the RG3 is always the RG with the highest entropy, the plot not only displays the distribution of the responses but also indicates how random the answers are (i.e. what their entropy is). The closer a point is to RG3, the higher its entropy. The same is true for points in the middle of the simplex. Having equal probabilities for all the three RG means that all possible confusions need to be used to explain the answers of a certain listener at a particular SNR.

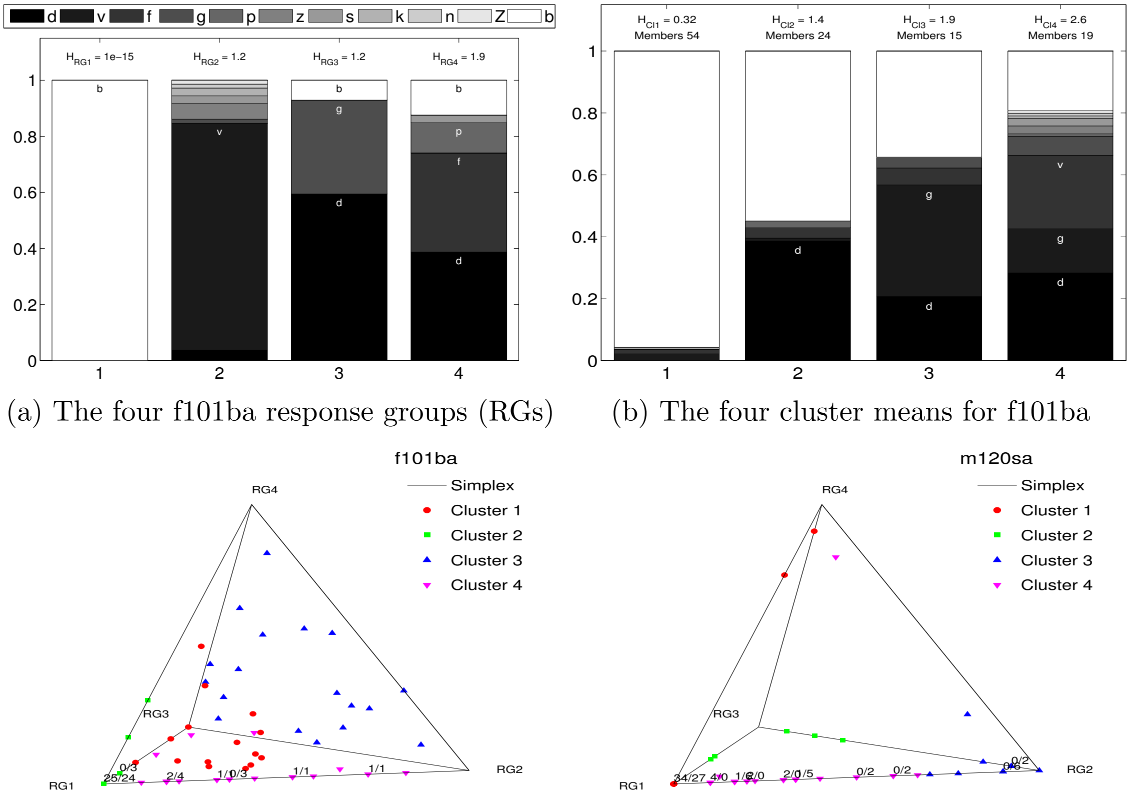
A.2 Comparison to k-means

The above described f101ba and m120sa are both analyzed with the k-means and with the PLSI algorithm. It is interesting to compare the two results (see Figure A.6). Since the k for the k-means algorithm was chosen to be 4, it is compared to the PLSI algorithm with four RGs. The clusters found by the k-means algorithm can be displayed in the simplex constructed by the RGs found with the PLSI algorithm.

It can be seen that the clusters found by k-means are in general also well grouped in the PLSI visualization. However, in Figure A.6c it can be seen that the blue cluster (cluster 3) spreads over a wide area. It can be concluded that either more clusters are needed to meaningfully describe the data by cluster means, or that the blue cluster should be viewed as an outlier cluster and should not be represented by its mean.

A.3 Conclusions on the Use of PLSI

To the author's knowledge, this document is the first to apply the PLSI algorithm to confusion matrix data (see Section A.1). The algorithm allows one to visualize the high-dimensional data without losing any information. Compared to previously



(a) The four f101ba response groups (RGs)

(b) The four cluster means for f101ba

(c) The k-means cluster data for f101ba displayed in the 3D simplex.

(d) The k-means cluster data for m120sa displayed in the 3D simplex.

Figure A.6: (a) shows the RGs of the token f101ba found by the PLSI algorithm for $N = 4$. (b) shows the cluster centroids found by the K-means algorithm for $K = 4$ for the same token. (c) and (d) display the clusters found by K-means colorcoded in the representation of response groups. This display allows to analyze how well the K-means algorithm worked.

used methods, where all rarely occurring confusions were concatenated onto one axis, this algorithm finds optimal confusion groups and estimates the probability for each listener to belong to these confusion groups. Compared to a k-means algorithm it does not make any hard assignments to groups and is therefore less[2] dependent on the choice of k. As seen in the k-means example in Section 4.5, the algorithm is often forced to use one of the k-groups as a collecting container for all the responses with high entropy. Those responses can be very different in nature (i.e., the Hellinger distance between the members of the k^{th} group can be large). However, because of the choice of k, they had to be grouped into one group (no more degrees of freedom were allowed). Therefore the k^{th}-mean averages responses that should be treated independently. By analyzing confusion matrix data with the PLSI algorithm one can decide which listeners can reasonably be grouped together and which should not. It is also an automated way to extract meaningful confusion groups without any a priori knowledge.

[2] The results of PLSI still are dependent on the number of confusion groups. However, the error does not necessarily decrease with the number of confusion groups.

Appendix B

Figure Appendix

B.1 Confusion Bars and $\mathcal{H}(P_e)$ Charts

In the following, all of the confusion bar plots and $\mathcal{H}(P_e)$ charts are discussed. For a detailed description of the plots see Figure 4.5 on p. 47 and surrounding text.

B.1.1 f103/ʃɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /ʃɑ/ can be found in Figure B.1.

Listeners In general f103/ʃɑ/ is a low error sound; only 9 out of the 16 ears have errors, and only 4 have error rates high enough to be considered for a further analysis (Figure B.1 left). The average error is 9.0% and 9.5%, in the FG and NAL-R experiment respectively. The error on average increases with NAL-R. However, this is due to only one ear, 02L.

Confusions The main confusions for the four ears are /sɑ/ and /zɑ/. To a smaller degree /ʒɑ/ is also part of the confusion group for the female token. In the NAL-R experiment the /zɑ/ confusions disappear completely.

Normal Hearing From the normal hearing data, confusions with /sɑ/ would be predicted, especially with the NAL-R high frequency boost. That explains why in

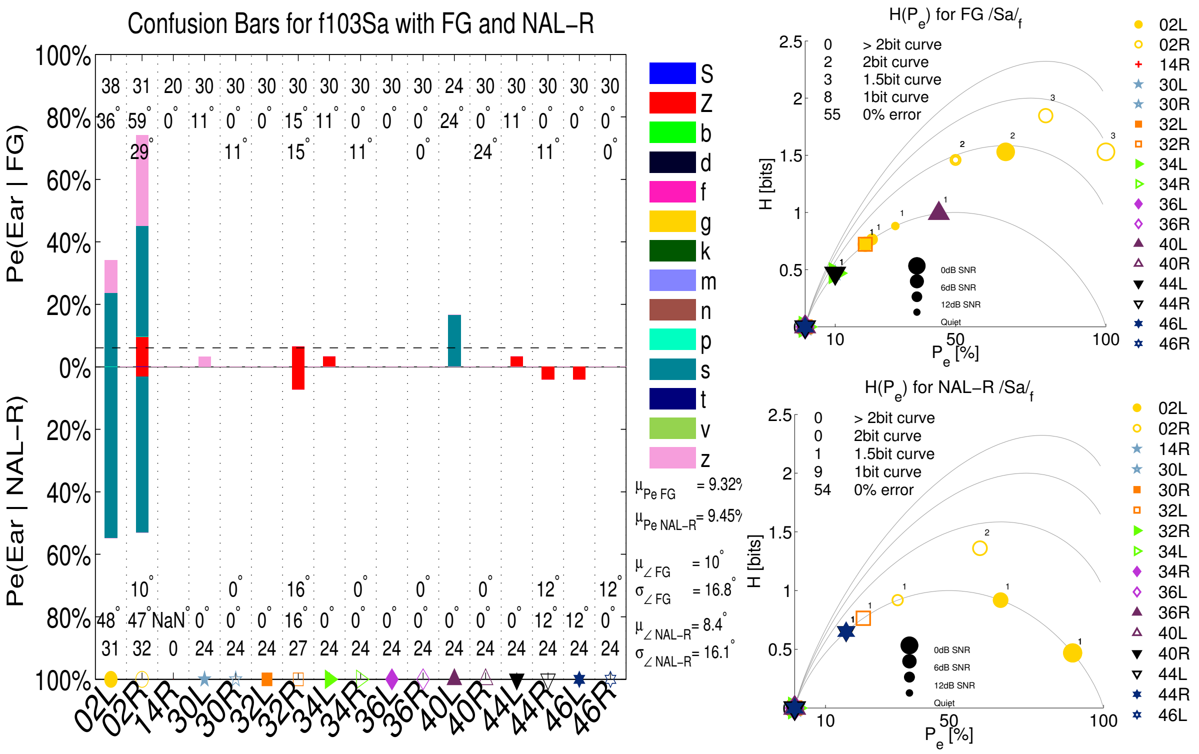


Figure B.1: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f103/ʃɑ/. For a detailed description see Figure 4.5.

the NAL-R experiment the probability for /sɑ/ confusions gets higher than in the FG experiment.

Entropy Curves Since the /zɑ/ confusions go away the entropy should be lower, especially for the ears 02L and 02R. The entropy curves confirm this observation; the yellow circles move down to the 1st and 2nd entropy curve.

Ears Out of the four ears with sufficient error, none is significantly different from the other ear. The largest difference can be observed in subject 02 in the FG experiment.

B.1.2 m118/ʃɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /ʃɑ/ can be found in Figure B.2.

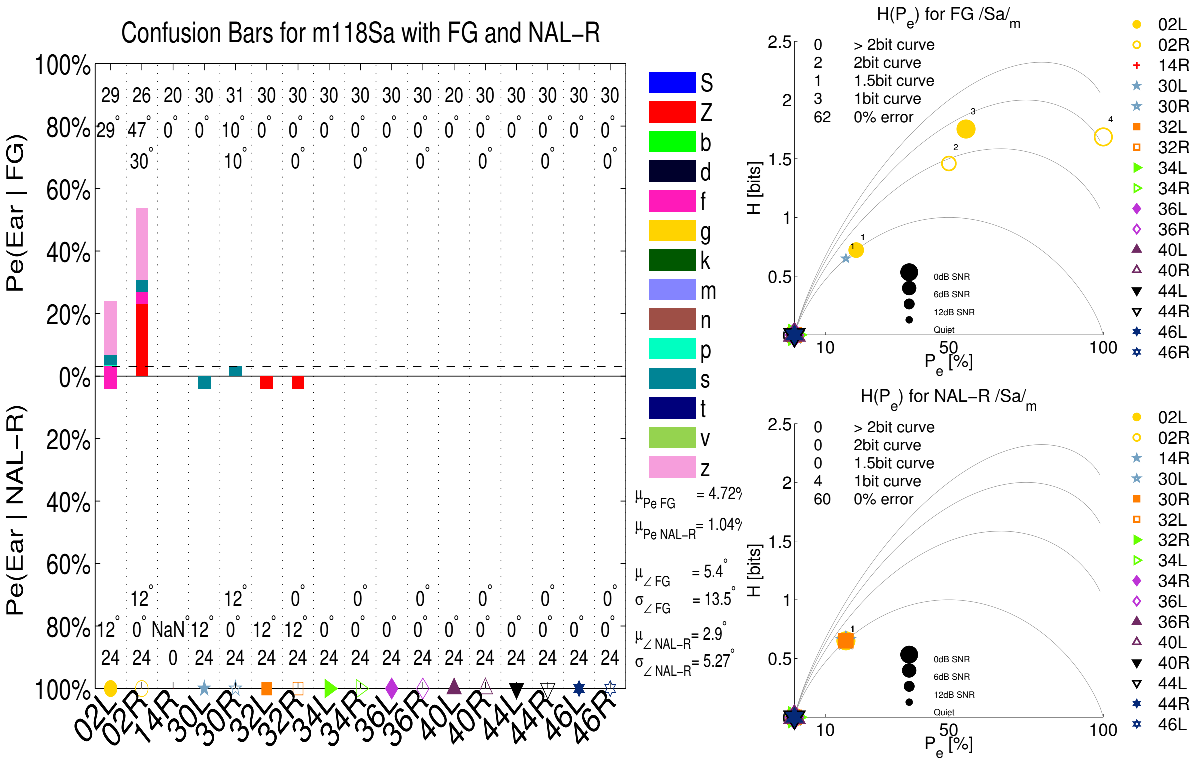


Figure B.2: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m118/ʃɑ/. For a detailed description see Figure 4.5.

Listeners m118/ʃɑ/ has even less error than its female counterpart. Only 2 out of the 16 ears have errors worth being considered for a further analysis. The average error is 4.5% and 1.04%, in the FG and NAL-R experiment respectively. NAL-R reduces the error to almost zero.

Confusions The main confusions are /sɑ/, /ʒɑ/ and /zɑ/; however, they are only made by subject 02, so their significance is questionable.

Normal Hearing According to the 3DDs data of this token /sɑ/ and /zɑ/ confusions are expected. The /zɑ/ can be observed in the data and would probably get stronger if the noise would be further increased.

Entropy Curves The entropy and error go down for the two ears with high enough error.

Ears In the FG experiment the two ears of 02 are different. Mainly due to the difference in /ʒɑ/ confusions, which only exists in the right ear.

B.1.3 f105/ʒɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /ʒɑ/ can be found in Figure B.3.

Listeners For f105/ʒɑ/ there are 8 out 16 ears with enough error (Figure B.3 left). The average error is 40.1% in the FG experiment and 32.6% in the NAL-R experiment.

Confusions The main confusion is /zɑ/, to a smaller degree /ʃɑ/ and /gɑ/ also appear in the responses. With NAL-R the /ʃɑ/ and /gɑ/ confusions get more likely (Figure B.3 left).

Normal Hearing The main confusion /zɑ/ is produced in exactly the same way as /ʒɑ/; the only difference is the friction noise which is not present in /zɑ/. This friction energy is easily masked by the noise. The confusion therefore makes sense.

Entropy Curves The entropy increases in many of the ears with NAL-R. However, the number of listeners per curve does not change (Figure B.3 right). The number of high error responses, however, decreases.

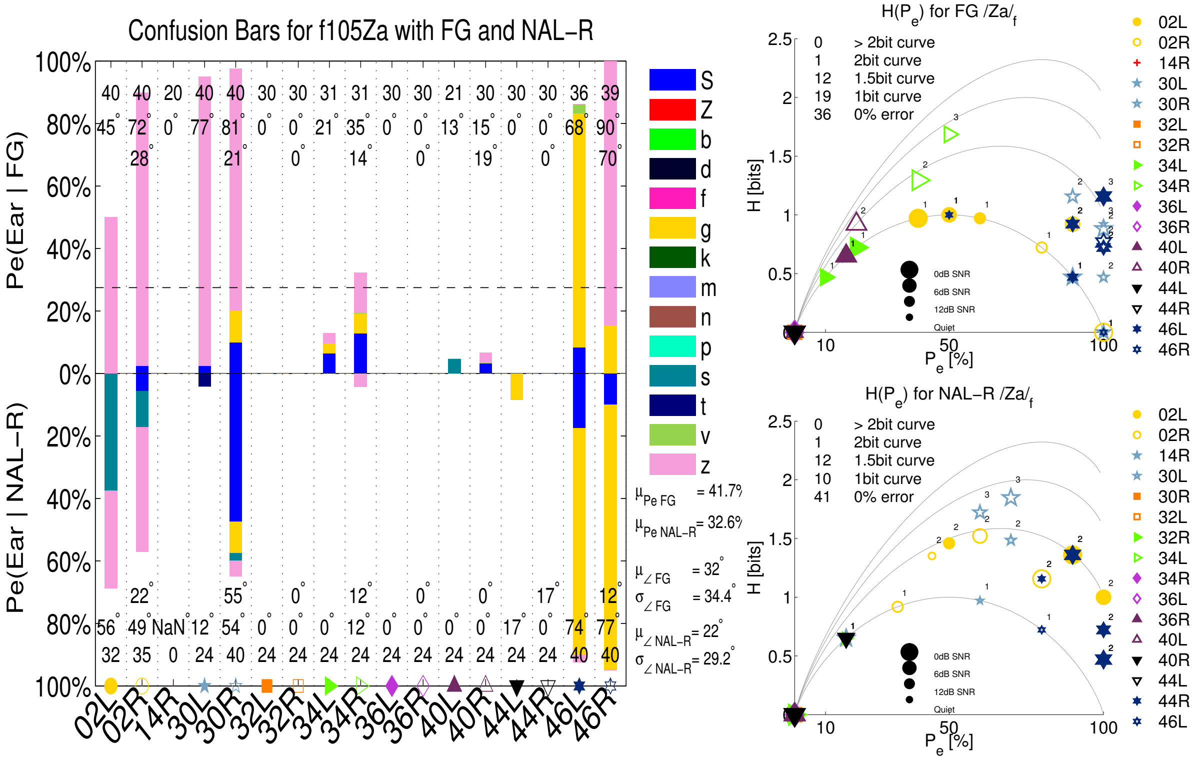


Figure B.3: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f105/ʒɑ/. For a detailed description see Figure 4.5.

Ears The ears of 46 are different for the FG experiment, with NAL-R they get more similar. 30 shows an interesting behavior for the two ears; they are different in both experiments, but NAL-R increases the difference. The left ear in the FG experiment shows high error but almost exclusively confuses the /ʒɑ/ with a /zɑ/; NAL-R eliminates this confusion totally. In the right ear, however, this confusion turns into a /ʃɑ/ (voicing is not perceived any more).

B.1.4 f105/dɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /dɑ/ can be found in Figure B.4.

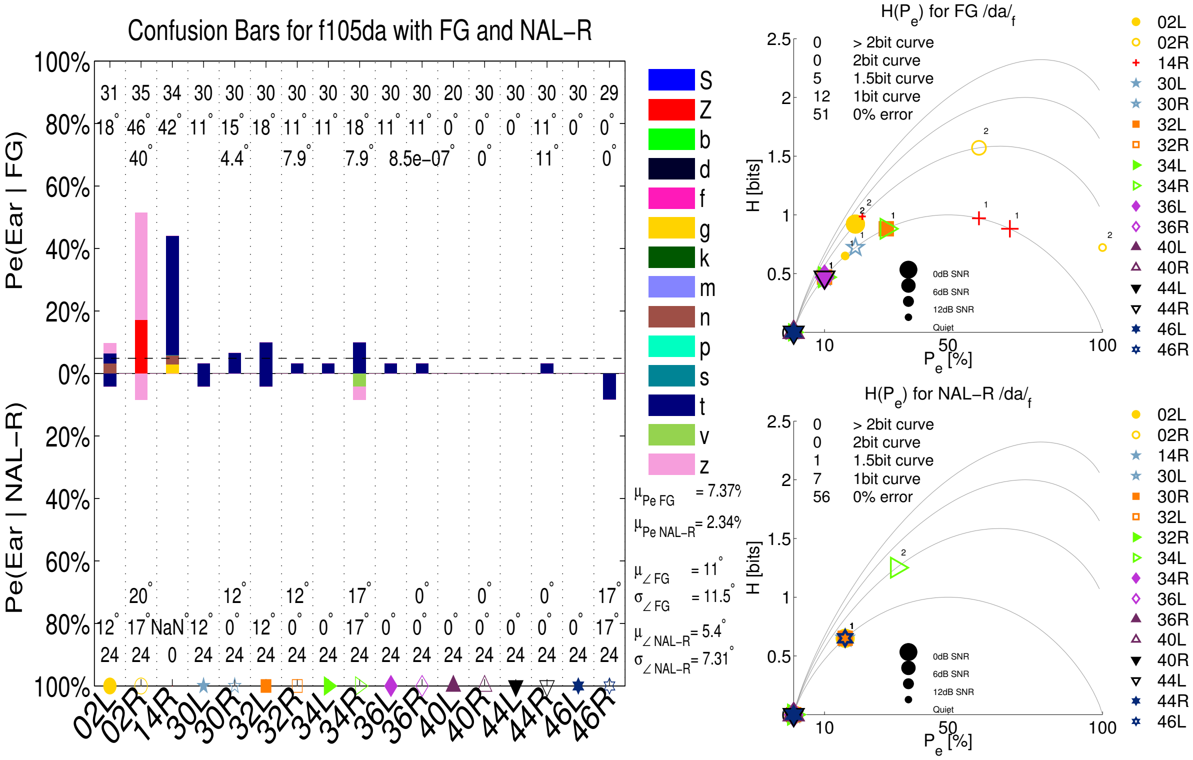


Figure B.4: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f105/dɑ/. For a detailed description see Figure 4.5.

Listeners For f105da there are only 3 out of the 16 ears with enough error. The average error is 9.8% in the FG experiment and 2.3% in the NAL-R experiment (Figure B.4 left).

Confusions The main confusion is /tɑ/; it appears in all but one of the responses with errors. Two other confusions, which only occur in the ear 02R, are /zɑ/ and /ʒɑ/. For the NAL-R experiment no listener shows significant errors.

Normal Hearing The /tɑ/ confusion makes sense; it is the unvoiced version of /dɑ/. Failing to recognize voicing in noise is a plausible cause for this confusion.

Entropy Curves The entropy curves are “empty” (low error sound); however, a decrease in entropy is noticeable nevertheless. There are 5 points close to the second curve in the FG experiment, whereas in the NAL-R experiment only one is close, and all others are on the 1st curve (Figure B.4 right).

Ears The ears of 02 are different (> 30°) in the FG experiment.

B.1.5 m118/dɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /dɑ/ can be found in Figure B.5.

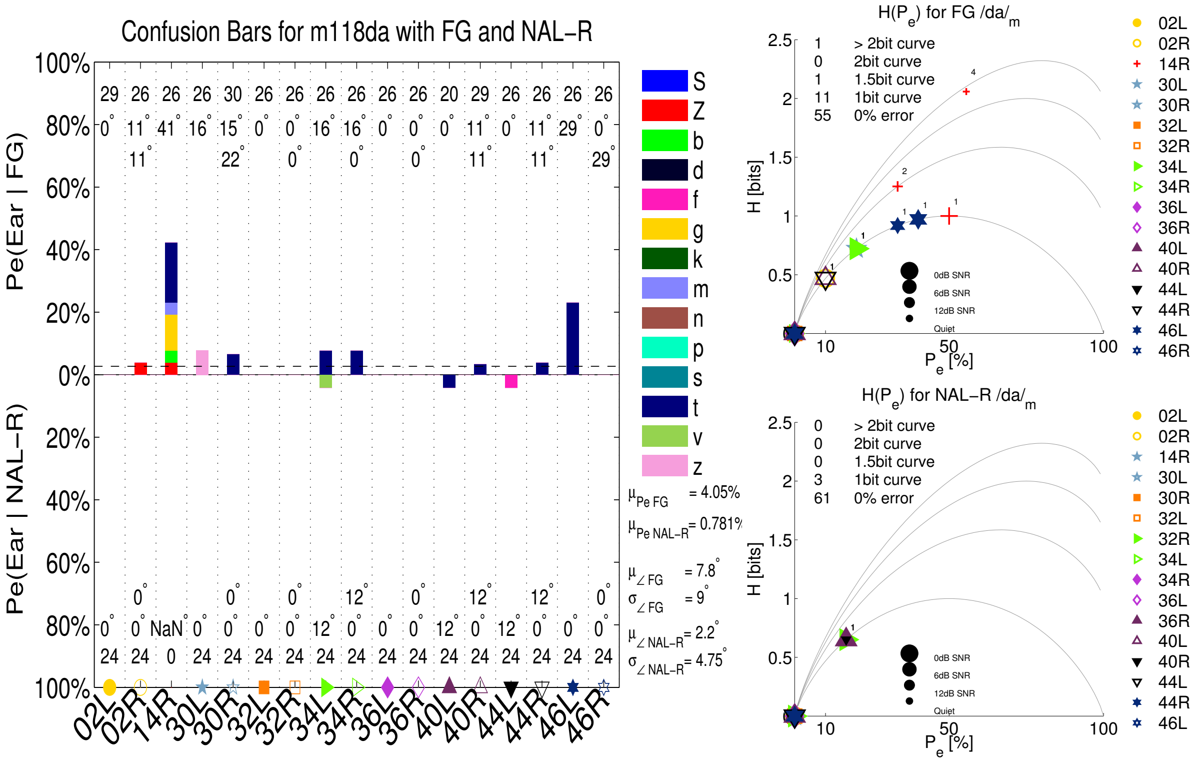


Figure B.5: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m118/dɑ/. For a detailed description see Figure 4.5.

Listeners The token m118/dɑ/ is a low error token; only 46L has enough errors to be taken into further analyses. Average error is 6.3% in the FG experiment in the NAL-R experiment the error reduces to almost zero 0.8%.

Confusions The main confusion is /tɑ/; it appears in all but two of the responses with errors and makes up 100% of the error for 46L. There are very few /zɑ/ and even fewer /ʒɑ/ errors (Figure B.5 left).

Normal Hearing The /tɑ/ confusion makes sense; it is the unvoiced version of /dɑ/. Failing to recognize voicing in noise is a plausible cause for the confusion.

Entropy Curves The error in the NAL-R experiment is very low. All the listeners respond with at most one confusion; they all lie on the 1bit curve (Figure B.5 right). The entropy therefore is also low.

Ears None of the ears is remarkably different form the other. The largest difference between the two ears is observed in subject 46.

B.1.6 f109/fɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /fɑ/ can be found in Figure B.6.

Listeners 10 out of the 16 ears have sufficient errors. The error on average is 31.4% and 18.9%, in the FG and NAL-R experiment respectively.

Confusions The main confusion is /sɑ/, it appears in all but one (02R) of the responses with sufficient errors in the FG experiment. Further confusions are /vɑ/, /ʃɑ/ and /zɑ/. In the NAL-R experiment all confusions except the main confusion /sɑ/ are greatly reduced; /sɑ/ the makes up most of the wrong answers (Figure B.6 left).

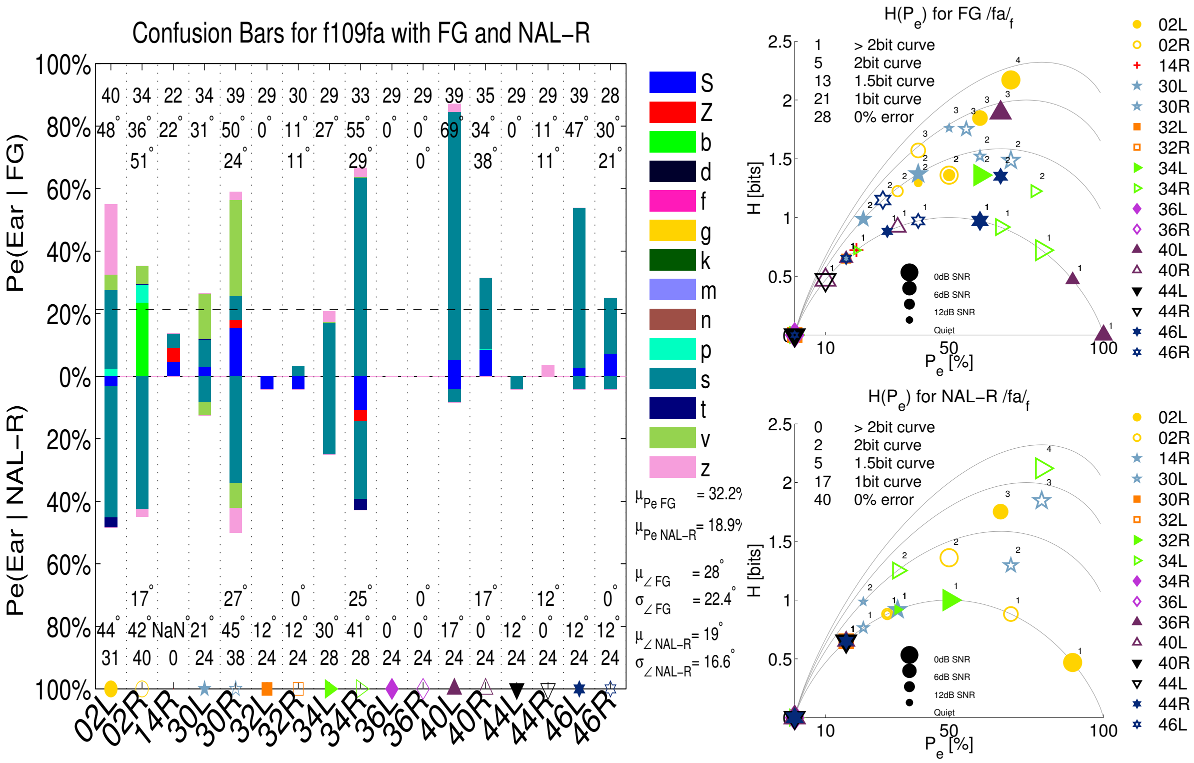


Figure B.6: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f109/fɑ/. For a detailed description see Figure 4.5.

Normal Hearing The confusion with /sɑ/ is hard to explain. The cue regions are in different places in time but also in frequency. The normal hearing experiments with this token show confusions with /tɑ/ and /bɑ/. Nevertheless the listeners are consistent in both experiments; therefore the signal needs to contain something triggering this response.

Entropy Curves The entropy clearly goes down. Six points are either close to the third curve or above the third curve. In the NAL-R none is above and only two are close to the 2bit-curve. Also the number of points close to the second reduces from 13 to 5 (Figure B.6 right).

Ears The ears of 02 are different in the FG experiment. In the NAL-R experiment the ears of 30 are the ears with the largest difference.

B.1.7 f109/gɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /gɑ/ can be found in Figure B.7.

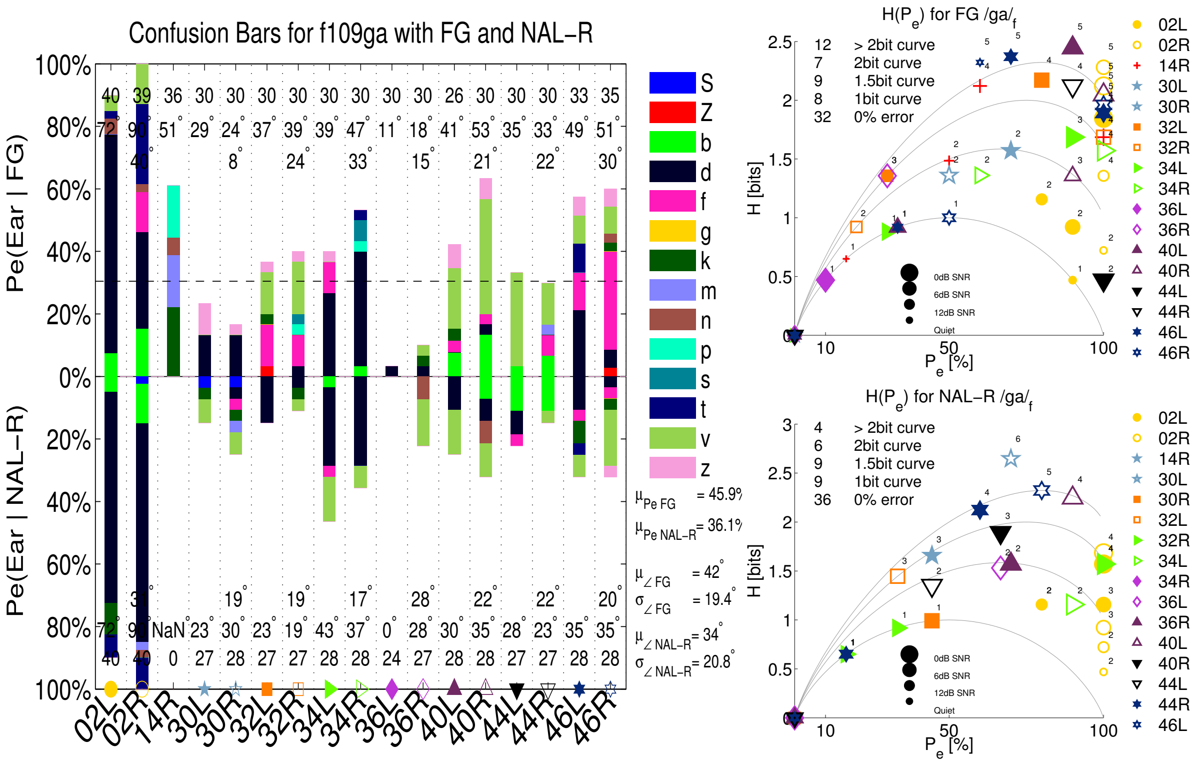


Figure B.7: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f109/gɑ/. For a detailed description see Figure 4.5.

Listeners 14 out of the 16 ears have sufficient errors. The error on average is 46.9% and 36.1%, in the FG and NAL-R experiment respectively.

Confusions A look at the confusion bars shows a colorful picture. There are many confusions for /gɑ/ and many listeners seem rather random in their responses (high entropy), it could be argued that /gɑ/ was not audible, especially for ear 46R. The major confusions are /dɑ/, /vɑ/, /bɑ/ and /fɑ/. Most confusions get decreased in their effects in the NAL-R experiment; only the /dɑ/ confusion remains at about the same rate (Figure B.7 left).

Normal Hearing The /dɑ/ and also /bɑ/ confusion are in the same Miller and Nicely confusion group and they show up in the NH experiments; in addition, /vɑ/ confusions are found in NH listeners. However the other confusions are hard to explain and are an indicator for guessing.

Entropy Curves The entropy goes down, but it stays surprisingly high. The female /gɑ/ token seems to be difficult for many ears even with NAL-R. The audibility of the token is questionable (Figure B.7 right top). It certainly increases with NAL-R.

Ears The two ears for 02, 34, 46 are different in the FG experiment. 02, 36 show the largest differences in the NAL-R experiment. These differences are another indicator for how random the answers are.

B.1.8 m111/gɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /gɑ/ can be found in Figure B.8.

Listeners Seven out of the 16 ears have sufficient errors. The error on average is 21.5% and 21.4%, in the FG and NAL-R experiment respectively.

Confusions The main confusion is /dɑ/. There are a few other confusions like /vɑ/ and /bɑ/ but all to a negligible degree. Compared to its female counterpart

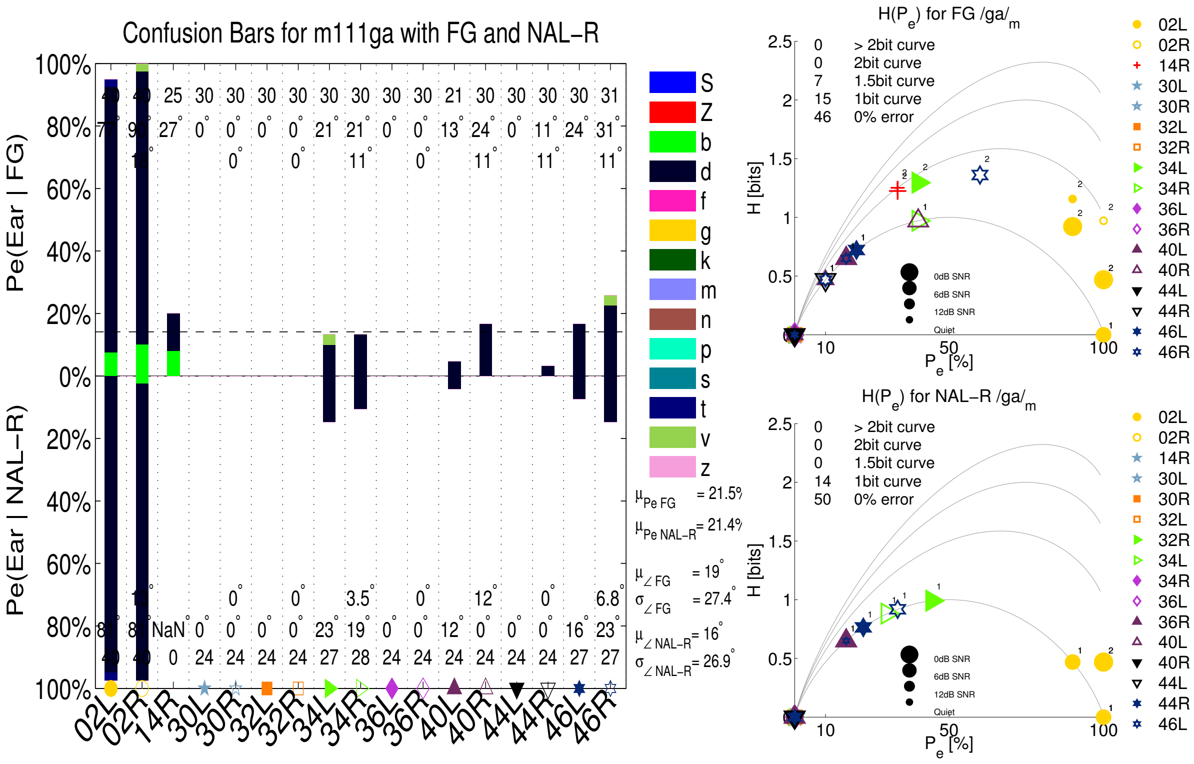


Figure B.8: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m111/gɑ/. For a detailed description see Figure 4.5.

the responses for this token are less random (Figure B.8 left).

Normal Hearing The /dɑ/ confusion has its cue in the same time region, its frequency is higher than the /gɑ/ cue. Confusions are expected, especially in the NAL-R experiment with high-frequency boost of NAL-R.

Entropy Curves The entropy goes down. All but one point with error only have /dɑ/ confusions and therefore they all lie on the 1-bit curve (Figure B.8 right); the point with two confusions (/dɑ, tɑ/) is also dominated by /dɑ/, the /tɑ/ confusion occurs with a probability of less than 5%.

Ears The only difference between ears is the amount of error. Confusion–wise they are all the same and therefore the angles between the ears are small.

B.1.9 f103/mɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /mɑ/ can be found in Figure B.9.

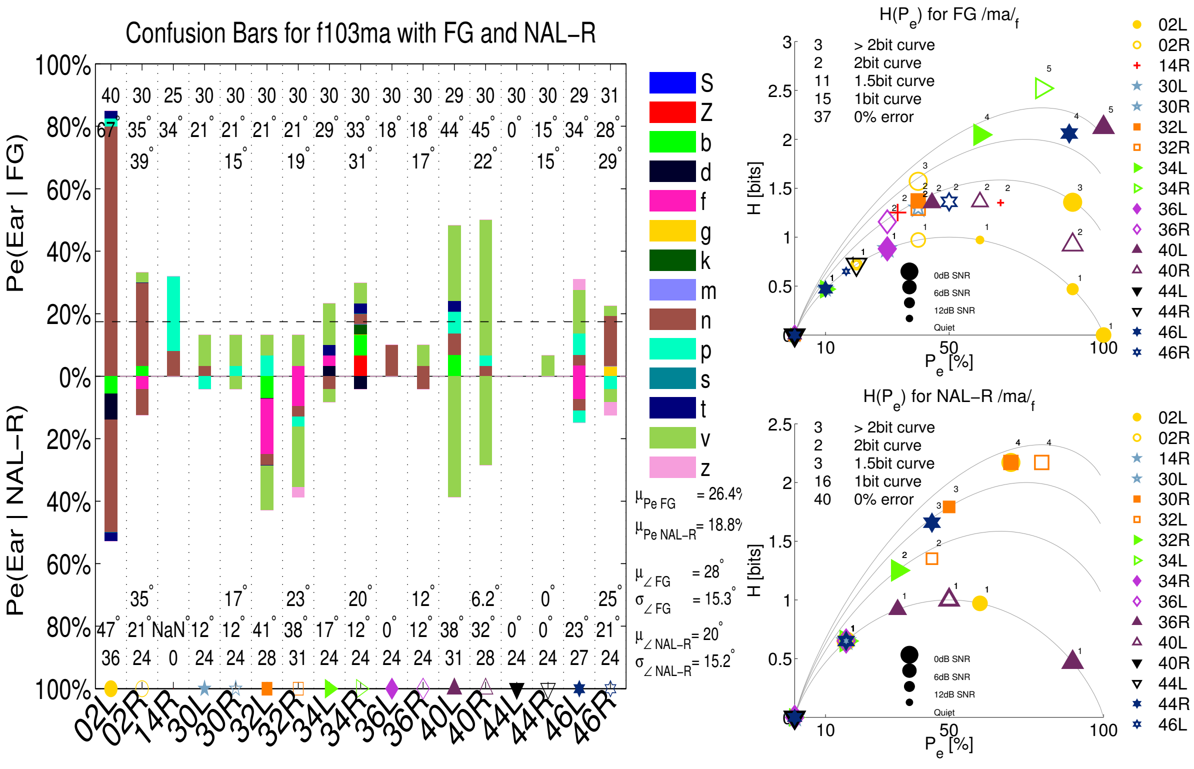


Figure B.9: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f103/mɑ/. For a detailed description see Figure 4.5.

Listeners 12 out of 16 listeners have enough errors for further analyses. The average error is 26.7% and 18.8% for the FG and NAL-R experiments respectively.

Confusions The main confusions are /vɑ/ and /nɑ/. Also /pɑ/ confusions show up in many listeners (Figure B.9 left).

Normal Hearing /nɑ/ confusions are expected. The two nasals /mɑ/ and /nɑ/ share the common feature of a nasal murmur and only differ from each other in the shape of F2 transition. /nɑ/ has a prominent downward F2 transition while /mɑ/ does not (Li and Allen (2011)).

Entropy Curves 40 is a perfect example for how NAL-R decreases the entropy. The high entropy answer of 40L becomes a low entropy one with NAL-R. Also 34 goes from a random response (audibility is questionable) to a low error response with one and two confusions. However, other subjects' performance (i.e. 32R) decreases with NAL-R. The entropy of the response greatly increases. Also the entropy for 02L increases while the error goes down. The number of points above the 2 bit curve and closet to the 2 bit curve stay the same, however the number closest to the 1.5 bit curve reduces from 11 to three (Figure B.9 right).

Ears Subject 02 has the largest differences between the two ears in both experiments. Subject 32 has two different ears. The left ear does not have any problems, while the right ear has a mild /vɑ/ confusion in the FG experiment and a high entropy response with NAL-R. Also the ears of 02 are different in both experiments. The left ear has a higher error in both experiments.

B.1.10 m118/mɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /mɑ/ can be found in Figure B.10.

Listeners 3 out of 16 listeners have enough errors for further analyses. The average error is 9.2% and 2.6% for the FG and NAL-R experiments respectively.

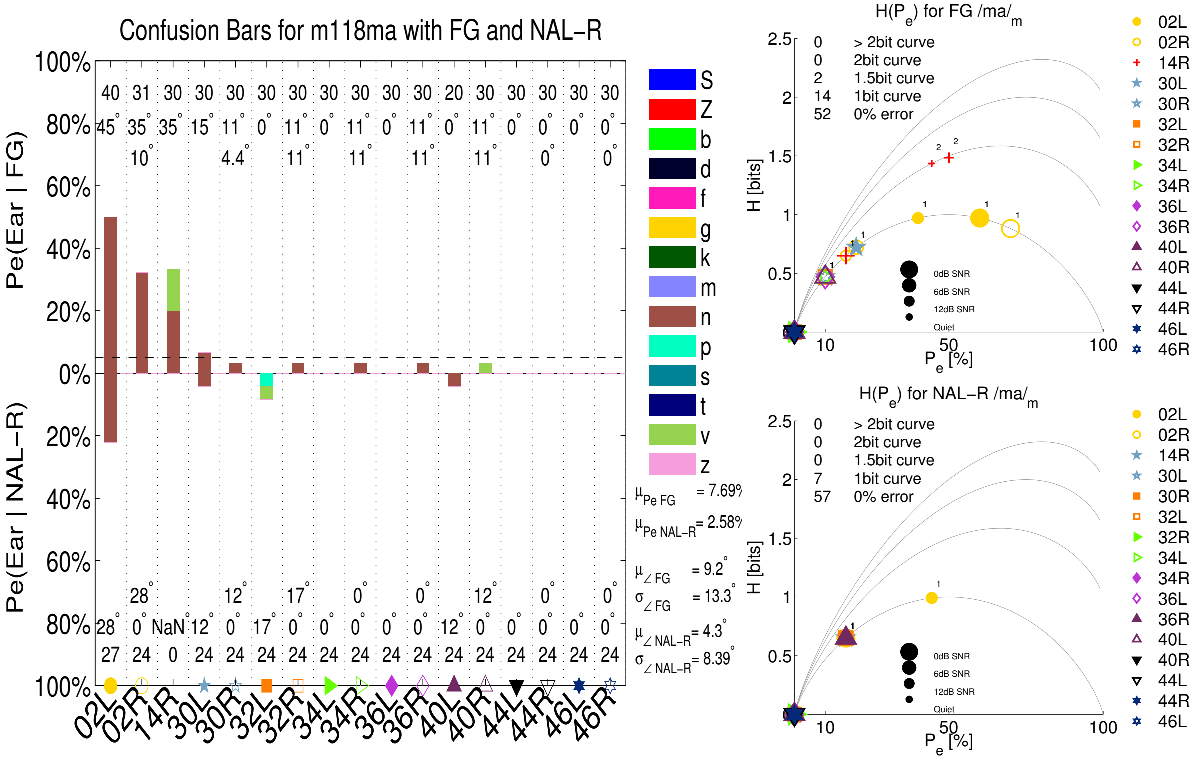


Figure B.10: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m118/mɑ/. For a detailed description see Figure 4.5.

Confusions The main confusions is /nɑ/.

Normal Hearing /nɑ/ confusions are expected. The two nasals /mɑ/ and /nɑ/ share the common feature of a nasal murmur and only differ from each other in the shape of F2 transition. /nɑ/ has a prominent downward F2 transition while /mɑ/ does not (Li and Allen (2011)).

Entropy Curves The error decreases only 02L still has significant errors. However, all the errors are with /bɑ/ and they are reduced greatly under the NAL-R condition. The entropy with NAL-R reduces, and all points are either at no error or on the 1bit curve (Figure B.10 right).

Ears There are no big differences between the ears.

B.1.11 f101/nɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /nɑ/ can be found in Figure B.11.

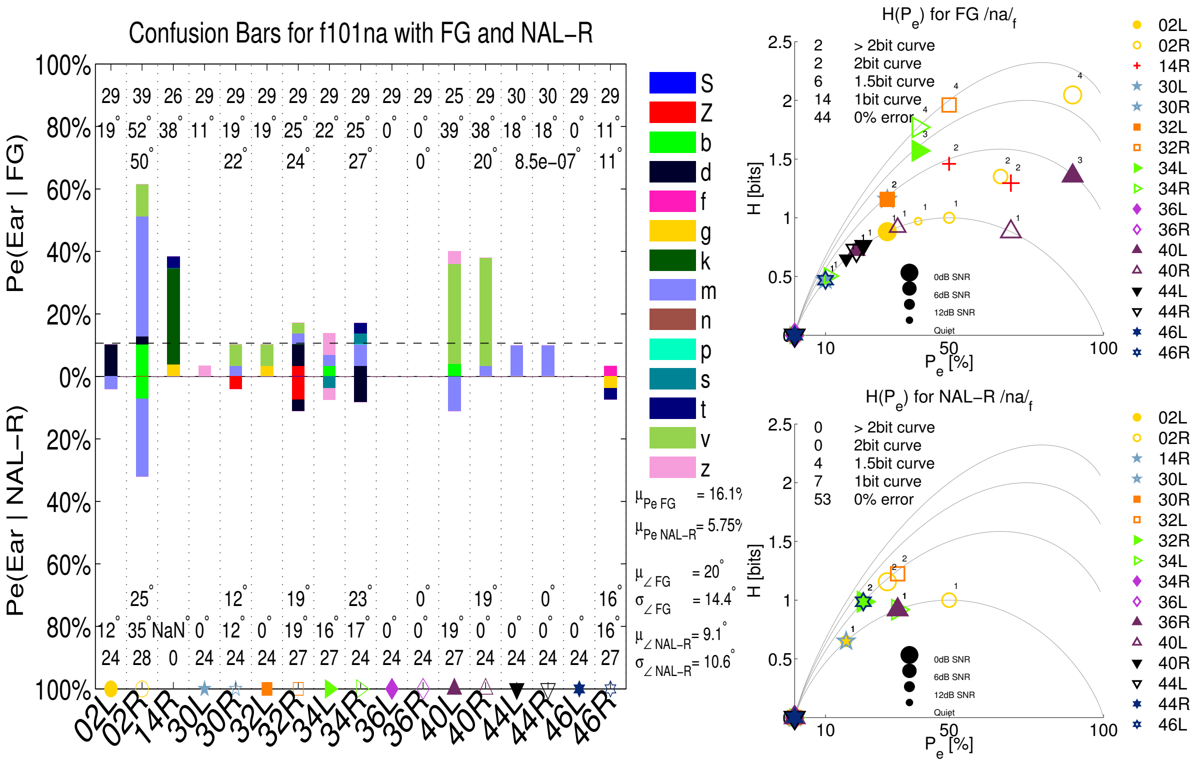


Figure B.11: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f101/nɑ/. For a detailed description see Figure 4.5.

Listeners Ten out of 16 listeners have enough errors for further analyses. The average error is 17.3% and 5.75% for the FG and NAL-R experiments respectively.

Confusions Even though the average error is small, the confusions vary widely across the listeners (Figure B.11 left). The main confusions are /mɑ/ and /vɑ/.

Normal Hearing The main confusions are explainable. It has been shown that a /nɑ/ can easily be converted into a /mɑ/ by removing the downward F2 transition (Li and Allen (2011)).

Entropy Curves The error decreases. Even though the errors are relatively small the entropy is still high and the confusions are still not consistent across ears. The high entropy of the answers suggest that the subjects were guessing at low SNRs (Figure B.11 right).

Ears Subject 02 shows a big angle between the two response vectors in the FG experiment. The ears of 34 have the same error but show completely different confusions. This is true for both experiments.

B.1.12 f103/pɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /pɑ/ can be found in Figure B.12.

Listeners 8 out of 16 listeners have enough errors for further analyses. The average error is 23% and 20.8% for the FG and NAL-R experiments respectively.

Confusions The main confusion is /tɑ/. The /kɑ/ confusions are small, but consistent across listeners in the FG experiment. In the NAL-R experiment the confusions are made almost exclusively with /tɑ/ (Figure B.12 left).

Normal Hearing The main confusion /tɑ/ has its cue in the same time region as the target sound /pɑ/. Also the /kɑ/ confusions can be explained. All the other confusions indicate an audibility issue, because the answers are random guesses.

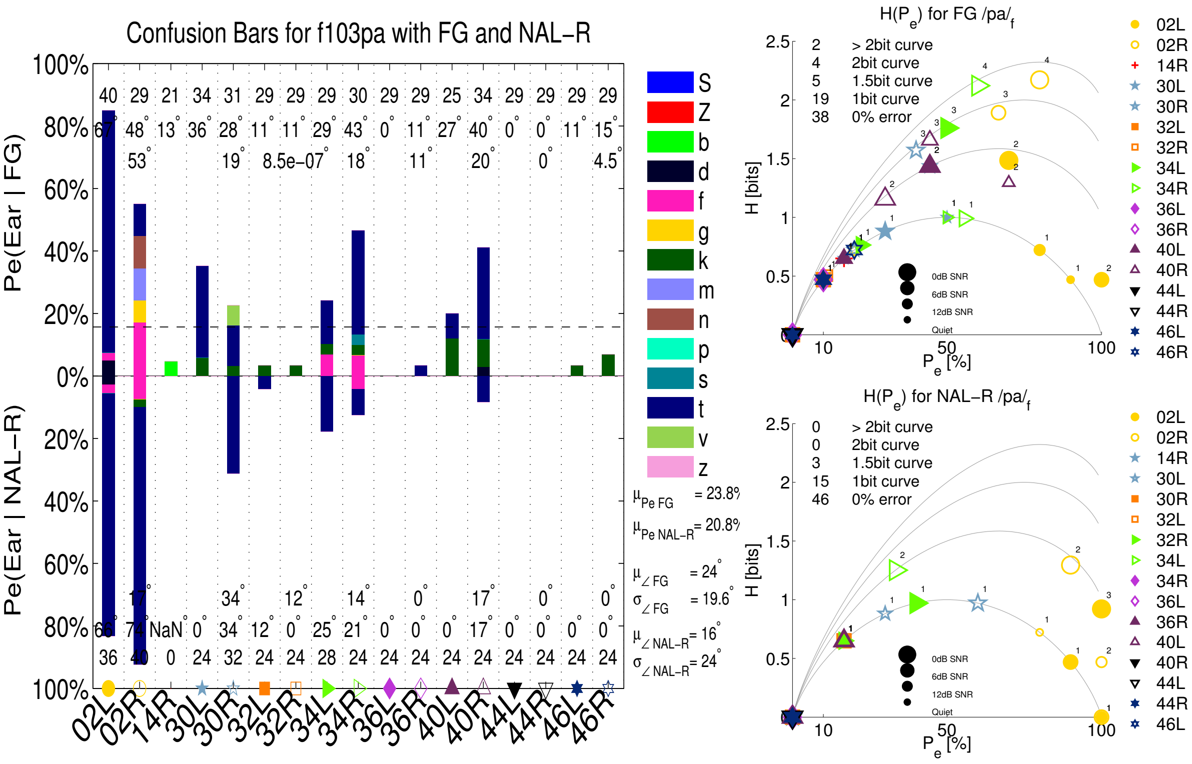


Figure B.12: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f103/pɑ/. For a detailed description see Figure 4.5.

Entropy Curves A very clear reduction in entropy is noticeable. 02R for example goes from a high-entropy answer, so high that audibility is questionable (Figure B.12 right), to a high-error answer with a low entropy. This ear, and in fact also the right ear of 02, would most likely benefit from training.

Ears The right ear of 02 in the FG experiment with its random responses differs greatly from the left ear, another indicator that audibility for the right ear was not achieved.

B.1.13 m118/pɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /pɑ/ can be found in Figure B.13.

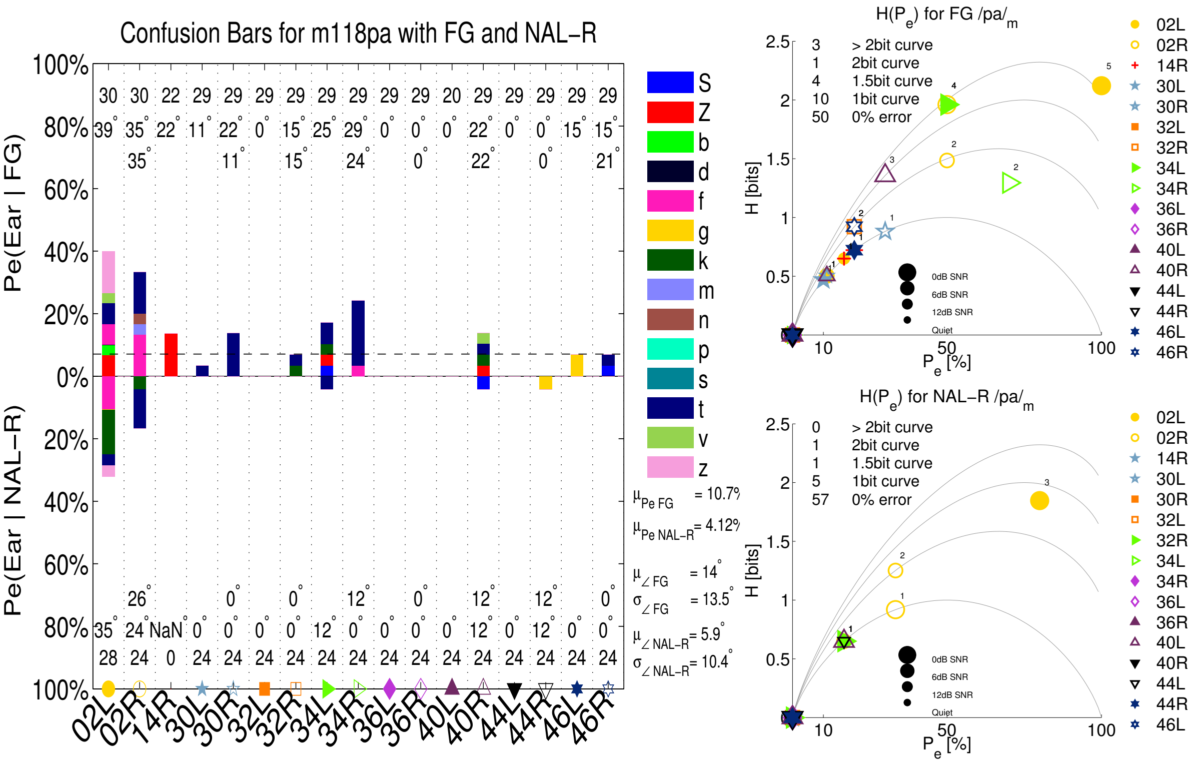


Figure B.13: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m118/pɑ/. For a detailed description see Figure 4.5.

Listeners 6 out of 16 listeners have enough errors for further analyses. The average error is 10.9% and 4.1% for the FG and NAL-R experiments respectively.

Confusions The main confusion is /tɑ/ (Figure B.13 left). Even though the error are low the entropy is high (Figure B.13 right).

Normal Hearing The main confusion /tɑ/ has its cue in the same time region as the target sound /pɑ/. The random other confusions are hard to explain.

Entropy Curves The error goes down for all ears except 44R who has one random error with /gɑ/. Noticeable form the P_e vs. $\mathcal{H}$ plots above is that, even though the error in the NAL-R experiment is low, the entropies of the responses are high (Figure B.13 right).

Ears The two ears of 02 are different, but both seem random. Also, the two ears of 34 are different: 34R has mostly /tɑ/ confusions, whereas the left ear shows /ʃɑ, ʒɑ, kɑ, tɑ/ confusions all to a similar degree. Also, 40R shows a few random confusions while 40L is 100% correct.

B.1.14 f108/tɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /tɑ/ can be found in Figure B.14.

Listeners Three out of 16 listeners have enough errors for further analyses. The average error is 8.7% and 1.6% for the FG and NAL-R experiments respectively.

Confusions It is hard to find main confusions. The confusions are different for all the listeners. The responses all have high entropies.

Normal Hearing /kɑ/ and /pɑ/ confusions would be expected. /pɑ/ shows up in some responses especially in 34R.

Entropy Curves The entropy and error goes down for all ears; many have perfect recognition with NAL-R. Since the entropies were high in the FG experiment, it could be that audibility was an issue (Figure B.14 right).

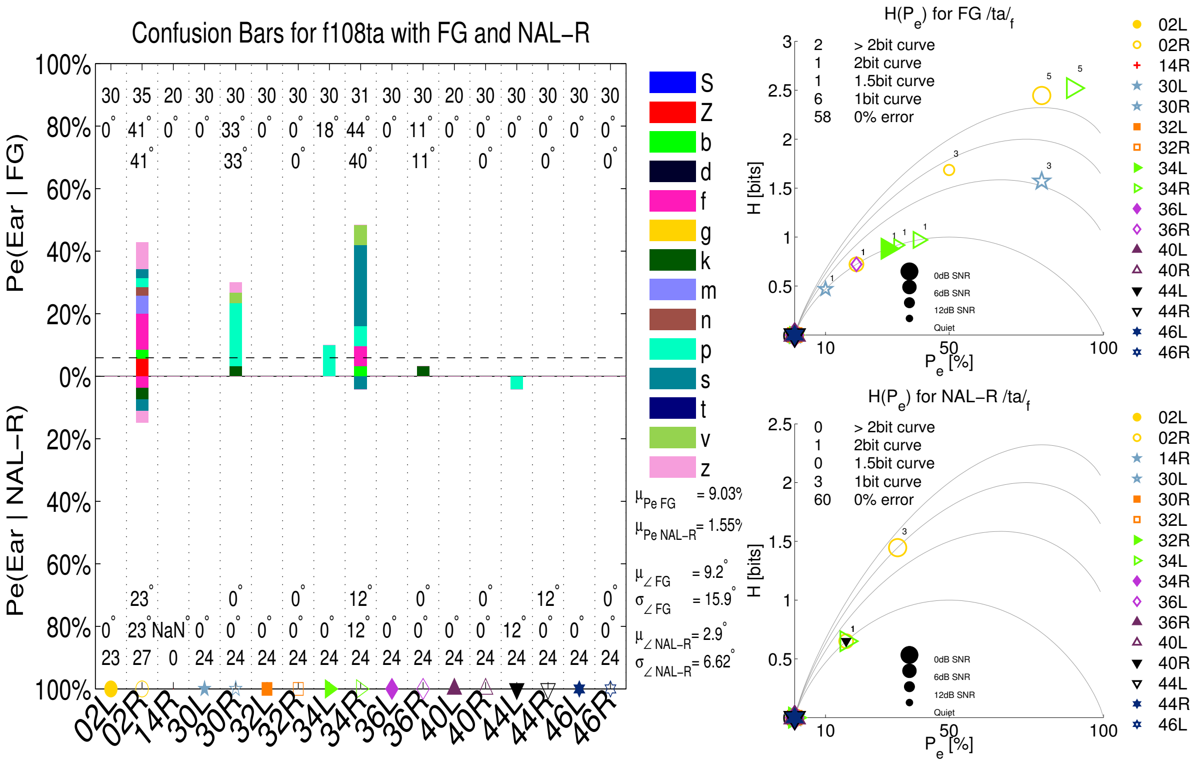


Figure B.14: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f108/tɑ/. For a detailed description see Figure 4.5.

Ears There are significant differences for the subjects 02, 30, 34. All of them have one ear with perfect recognition and another one with high error.

B.1.15 m112/tɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /tɑ/ can be found in Figure B.15.

Listeners 2 out of 16 listeners have enough errors for further analyses. The average error is 5.0% and 1.3% for the FG and NAL-R experiments respectively.

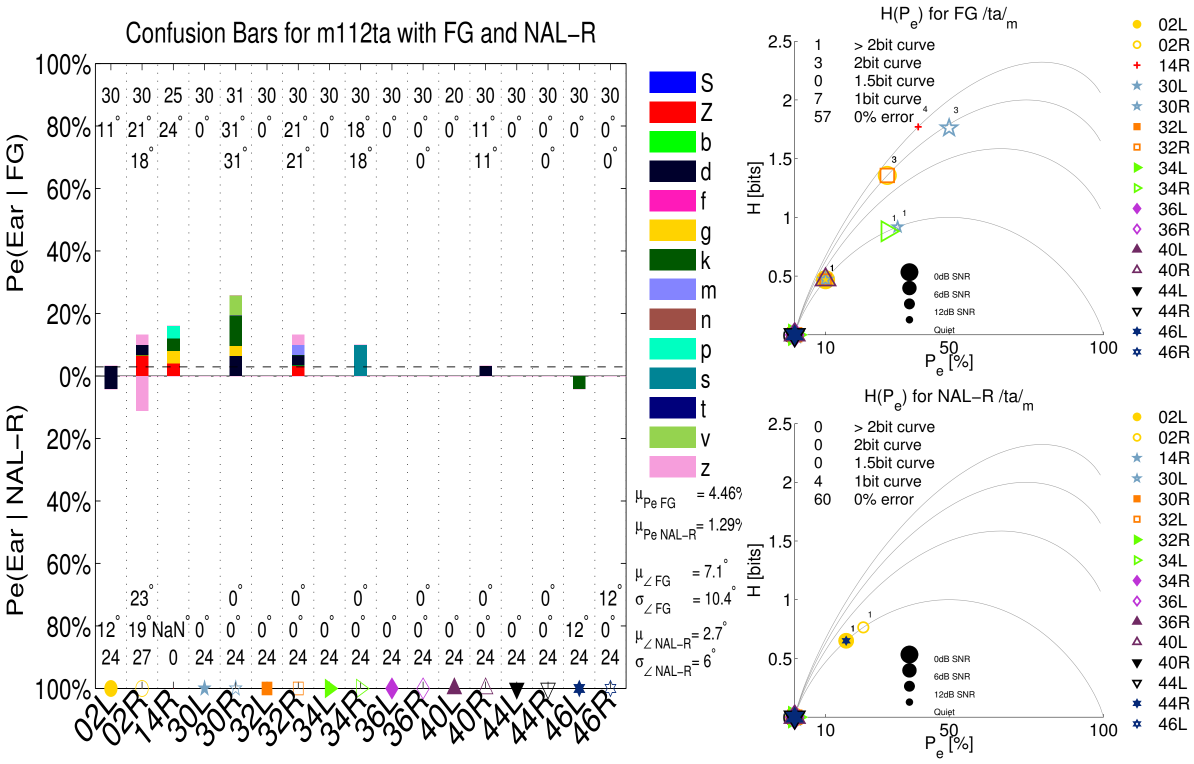


Figure B.15: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m112/tɑ/. For a detailed description see Figure 4.5.

Confusions As with the female token, there are many small, almost random, confusions instead of meaningful ones. In the NAL-R experiment there are only two confusions /zɑ/ and /dɑ/, none of which would be expected according to the normal hearing data.

Normal Hearing /kɑ/ and /pɑ/ confusions would be expected. /pɑ/ shows up in some responses especially in 34R.

Entropy Curves Error disappears (Figure B.15 right). Only 02R has significant errors.

Ears There are significant differences for the subjects 02, 30, 34. All of them have one ear with perfect recognition and another one with high error.

B.1.16 f101/vɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /vɑ/ can be found in Figure B.16.

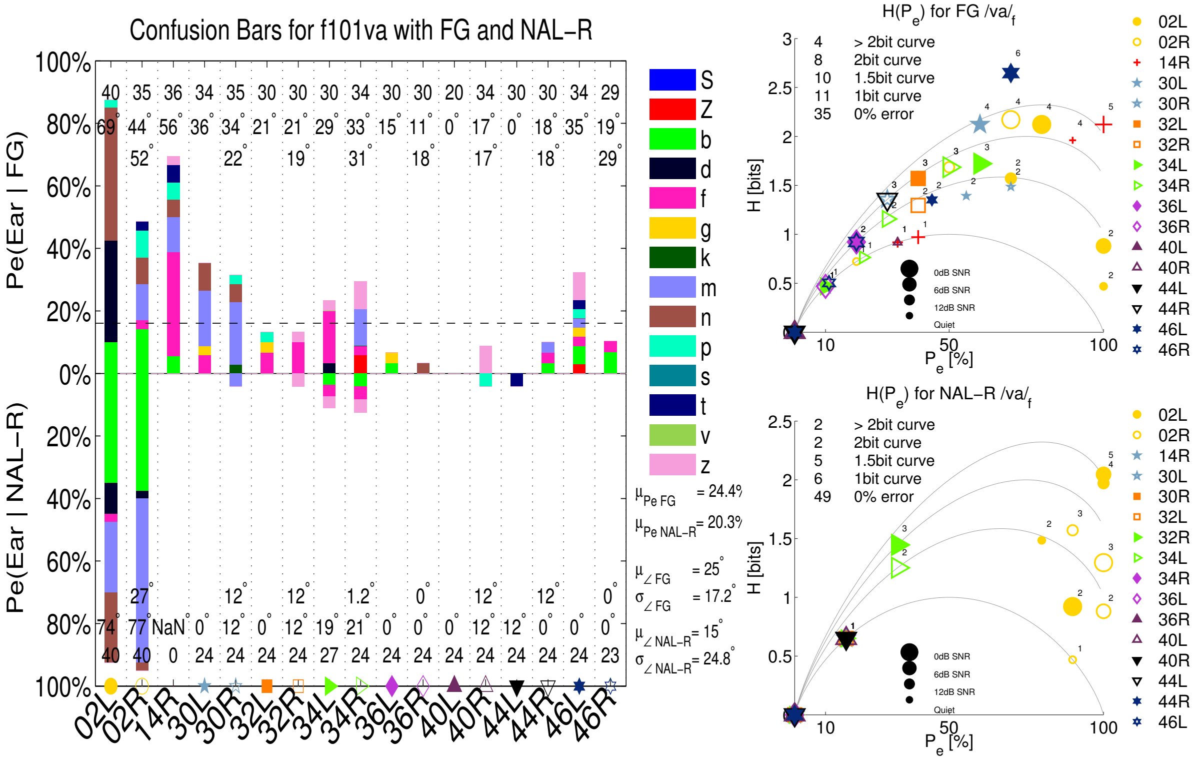


Figure B.16: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f101/vɑ/. For a detailed description see Figure 4.5.

Listeners Seven out of 16 listeners have enough errors for further analyses. The average error is 27.4% and 20.3% for the FG and NAL-R experiments respectively.

Confusions The main confusions are /mɑ/ and /fɑ/; however, there are also many other confusions in the responses (Figure B.16 left).

Normal Hearing NH people make /bɑ/, /fɑ/ and /nɑ/ confusions at SNRs lower than -10 [dB]. The observed /mɑ/ confusion is not surprising since /mɑ/ and /nɑ/ are very similar.

Entropy Curves Entropy and error decreased for all ears except 02. Subject 02 shows more error at higher entropy in the NAL-R experiment (Figure B.16 bottom right).

Ears The ears of 02 are different in both experiments. The ears of 46 are different in the FG experiment. One shows high error and high-entropy, whereas the other one has low error and its confusion group only contains /fɑ/ and /bɑ/.

B.1.17 f106/zɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the female token of /zɑ/ can be found in Figure B.17.

Listeners Ten out of 16 listeners have enough errors for further analyses. The average error is 34.4% and 28% for the FG and NAL-R experiments respectively.

Confusions The main confusions are /ʒɑ/ and /vɑ/. Furthermore, for 3 ears there are strong /dɑ/ confusions and subject 34 also has /nɑ/ confusions (Figure B.17 left).

Normal Hearing NH people make /ʒɑ/confusions at SNRs around -10 [dB] on the same token; other /zɑ/ tokens show additional confusions with /sɑ/. The other confusions, however, are hard to explain.

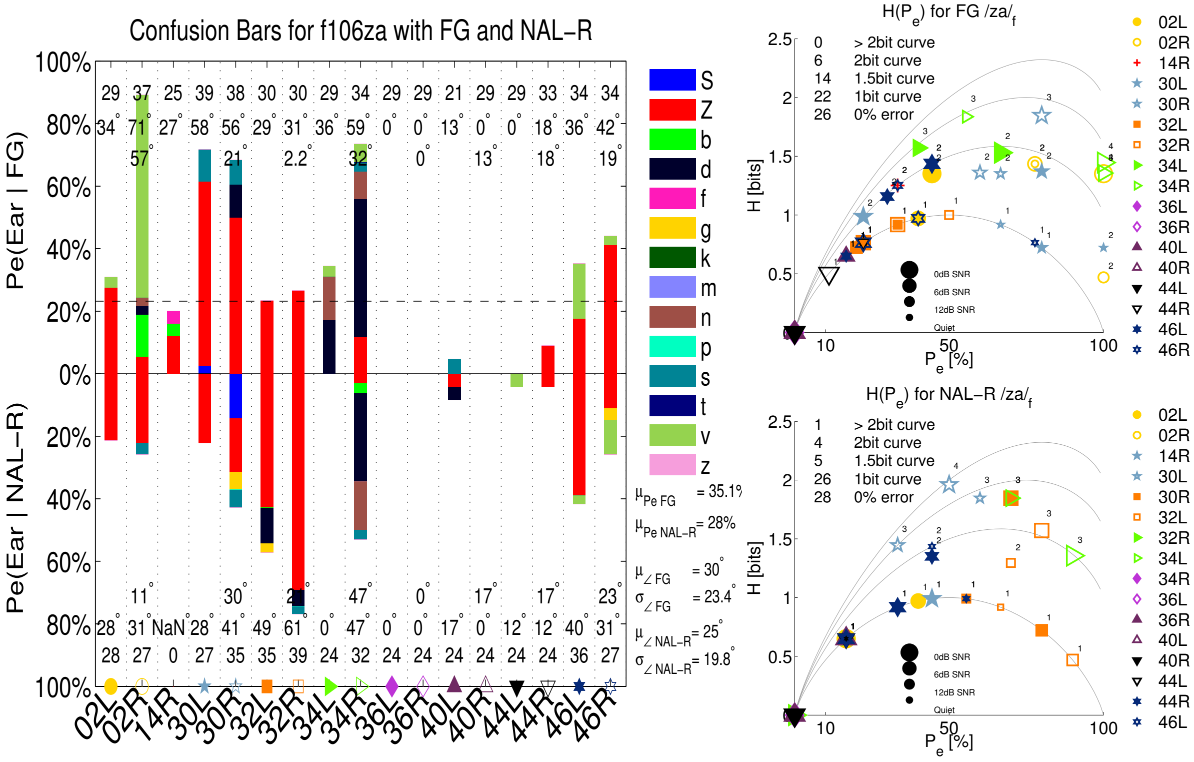


Figure B.17: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token f106/zɑ/. For a detailed description see Figure 4.5.

Entropy Curves There is no general trend in entropy. It decreases for 02R for example and the confusions change from mostly /vɑ/ to mostly /ʒɑ/, but it does not decrease in general. For 34R it stays about the same and it even increases for 30R for example.

Ears The ears of 02 are different in the FG experiment. The left ear shows the expected /ʒɑ/ confusions, while the right ear has a higher entropy with large /vɑ/ confusions. Also, 32 has two differently performing ears; one of them has no error at all in both experiments, while the other shows /ʒɑ/ confusions that get stronger in the NAL-R experiment. In the NAL-R experiment the ears of subject 30 are different. The left ear shows the expected /ʒɑ/ confusions while the right ear shows a high entropy response with //ʒɑ/,/ʃɑ/,/gɑ// and /sɑ/ confusions.

B.1.18 m118/zɑ/

The confusion bars as well as the $\mathcal{H}(P_e)$ charts for the male token of /zɑ/ can be found in Figure B.18.

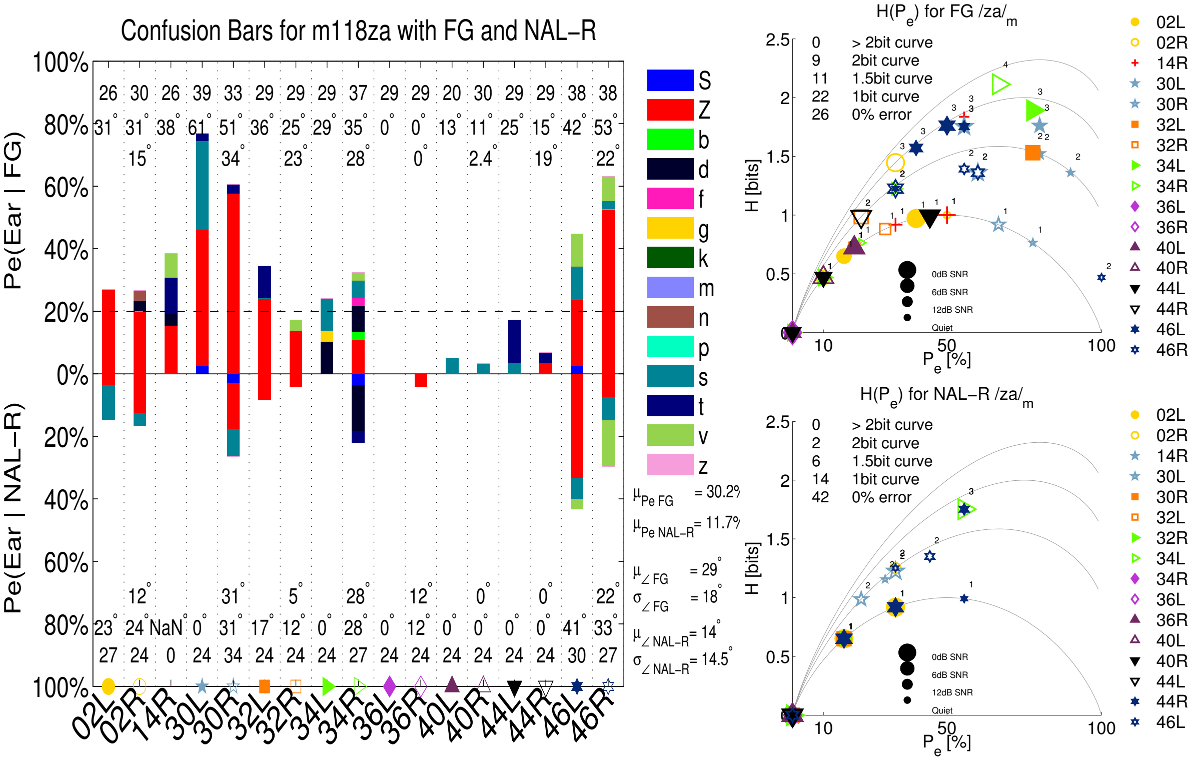


Figure B.18: The confusion bars (left) and $\mathcal{H}(P_e)$ plots for the FG and NAL-R condition (right) are displayed for token m118/zɑ/. For a detailed description see Figure 4.5.

Listeners Ten out of 16 listeners have enough errors for further analyses. The average error is 30.6% and 11.7% for the FG and NAL-R experiments respectively.

Confusions The main confusions are /ʒɑ/ and /sɑ/ (Figure B.18 left).

Normal Hearing Even though this specific token showed perfect recognition in the NH experiments all the way to -16 [dB], the confusions made by HI people are seen in other /zɑ/ tokens. /ʒɑ/confusions are the most common confusions in other tokens, but confusions with /sɑ/ are also present.

Entropy Curves In contrast to the female token, NAL-R decreases both the error and the entropy of the responses. For all the ears, except 34R, all confusions other than /ʒɑ/ and /sɑ/ reduce to an insignificant level.

Ears The two ears of 34 are different. Both have high entropy but the confusions that they are making are different.

Appendix C

Binomial Confidence Interval

C.1 Introduction

In the research of the Human Speech Recognition (HSR) group, a longstanding problem is to determine how many trials are necessary in order to estimate the probability $P_{h|s}$ (/h/ heard given /s/ spoken) with a specific confidence level (e.g. $1 - \alpha = 0.95$). Several approaches were taken over the past couple of years.

C.2 Vysochanskij-Petunin (VP)

The approach taken in Han (2011) is based on the Vysochanskij-Petunin Inequality.[1] The inequality gives a lower and upper bound for a probability of a random variable (X) with a unimodal distribution and a finite variance (σ^2). The probability of the mean of X being in the interval $[\mu \pm \lambda\sigma]$ is given as

$$P\left[|X - \mu| \geq \lambda\sigma\right] \leq \frac{4}{9\lambda^2} \tag{C.1}$$

In Han's thesis, $\lambda = 3$ was chosen, which gives a probability of $p_{3\sigma} = \frac{4}{9 \cdot 3^2} = 0.0494$ of X laying in the interval $[\mu \pm 3\sigma]$.

Adapted for Bernoulli trials, the following procedure was proposed:

[1] http://en.wikipedia.org/wiki/Vysochanski%C3%AF%E2%80%93Petunin_inequality

1. Best estimate of the true probability $P_{h|s}$ is

$$\hat{p} = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{C.2}$$

2. The standard deviation for the binomial distribution is

$$\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \tag{C.3}$$

3. The lower bound $\hat{p}_{lb}$ for $\hat{p}$ was then calculated by

$$\hat{p}_{lb} = \hat{p} - 3\hat{\sigma}_{\hat{p}} = \hat{p} - 3\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \tag{C.4}$$

C.3 Wald (Normal Approximation) Interval

A very similar approach is the Wald interval introduced by Wald and Wolfowitz 1939.[2]

It is derived as follows. From the central limit theorem we get that (N_S number of successes, N sample size)

$$\frac{N_S - Np}{\sqrt{Np(1-p)}} = \frac{(N_S/N) - p}{\sqrt{p(1-p)/N}} \tag{C.5}$$

has an approximate normal distribution $N(0,1)$ provided that N is large enough. That means for a given probability $1-\alpha$ we can find the $z_{\alpha/2}$ percentile of a standard normal distribution such that

$$P\left[-z_{\alpha/2} \leq \frac{(N_S/N) - p}{\sqrt{p(1-p)/N}} \leq z_{\alpha/2}\right] \approx 1 - \alpha. \tag{C.6}$$

[2] http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval#Normal_approximation_interval

With some algebra

$$P\left[-z_{\alpha/2}\sqrt{p(1-p)/N} \leq (N_S/N) - p \leq z_{\alpha/2}\sqrt{p(1-p)/N}\right] \approx 1-\alpha$$

$$P\left[-(N_S/N) - z_{\alpha/2}\sqrt{p(1-p)/N} \leq -p \leq -(N_S/N) + z_{\alpha/2}\sqrt{p(1-p)/N}\right] \approx 1-\alpha$$

$$P\left[(N_S/N) + z_{\alpha/2}\sqrt{p(1-p)/N} \geq p \geq (N_S/N) - z_{\alpha/2}\sqrt{p(1-p)/N}\right] \approx 1-\alpha$$

we get

$$P\left[\frac{N_S}{N} - z_{\alpha/2}\sqrt{p(1-p)/N} \leq p \leq \frac{Y}{N} + z_{\alpha/2}\sqrt{p(1-p)/N}\right] \approx 1-\alpha. \tag{C.7}$$

If we replace p with the best estimate for p, which is $\hat{p} = N_S/N$, we get the $100(1-\alpha)$ % confidence interval for our estimate $\hat{p}$:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \tag{C.8}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of a standard normal distribution. For a confidence level of 95% ($\alpha = 0.05$) $z_{\alpha/2} = 1.96$.

The 3σ from the Vysochanskij-Petunin can therefore be reduced to 1.96 by the assumption that the distribution of X approaches a normal distribution. However, this assumption is only true for large n; otherwise, the Wald interval is known to perform poorly (see Ghosh (1979) and Blyth and Still (1983)).

C.4 Clopper-Pearson Interval

Instead of assuming the distribution to be unimodal (Vysochanskij-Petunin) or normal (Wald) the distribution can be taken to be what it is, namely discrete and binomial. That is the reason why the Clopper-Pearson interval is often referred to as the 'exact method'. This method is for example used in the Matlab function `binofit`. Unfortunately, the formulas get quite messy for calculating these intervals. The reasoning behind them, however, is easy to understand. We set the cumulative distribution function (CDF) of our bound probabilities equal to half of our significance level α (half of α for the lower bound (lb) and half for the upper bound (ub)).

$$\sum_{0 \leq k < N_S} \binom{N}{k} p_{lb}^k (1 - p_{lb})^{(N-k)} = \frac{\alpha}{2} \tag{C.9}$$

$$\sum_{N_S < k \leq N} \binom{N}{k} p_{ub}^k (1 - p_{ub})^{(N-k)} = \frac{\alpha}{2} \tag{C.10}$$

where p_{lb} is the lower bound and p_{ub} is the upper bound. N_S is the number of successes out of N trials.

To solve for p_{lb} and p_{ub} is not a simple task. However, the fact that the Binomial CDF (as shown above) can be calculated with the beta distribution[3] enables us to write:

$$p_{ub} = 1 - \text{BetaInv}\left(\frac{1-\alpha}{2}, N-k, k+1\right) \tag{C.11}$$

$$p_{lb} = 1 - \text{BetaInv}\left(1 - \frac{1-\alpha}{2}, N-k, k+1\right) \tag{C.12}$$

For the case of $N_S = N$ the calculations for the lower bound can be simplified. We know $P(k, p) = \binom{N}{k} p^k (1-p)^{(N-k)}$ is the probability of seeing k successes out of N trials if the probability for a success is given by p. Since we had $N_S = N$ successes we get $\hat{p} = N_S/N = N/N = 1$; however, we assume that this result is only correct

[3] Alternatively the F distribution can also be used.

with a $1-\alpha$ confidence. So that means if we would repeat the experiment 100 times we would only see 95 times $N_S = N$ and 5 times we would see $N_S < N$. Therefore we are only $100(1-\alpha)$ % sure to see $P(k = N, p) = \binom{N}{N} = p^N(1-p)^0 = 1 \cdot p^N \cdot 1 = p^N$. We also know that $\sum_{k=0 \ldots n} \binom{n}{k} p_{lb}^k (1-p_{lb})^{(n-k)} = 1$; therefore, the probability to see $N_S < N$ is $\sum_{k=0 \ldots N-1} \binom{N}{k} p^k (1-p)^{(N-k)} = 1 - p^N$ and we expect it to be α:

$$p_{lb} = \sqrt[N]{1-\alpha} \tag{C.13}$$

Instead of expecting it to be α we could also expect it to be $\alpha/2$ since that is the probability for the lower bound according to the definition in Equation C.9.

$$p_{lb} = \sqrt[N]{1-\alpha/2} \tag{C.14}$$

C.5 Wilson Interval

Agresti and Coull (1998); Coull and Agresti (2000) claim that the Wilson interval is a more accurate confidence interval for a binomial proportion than the 'exact method'. The Wilson interval is defined as

$$p_{lb} = \frac{\hat{p} + \frac{z^2}{2N} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \tag{C.15}$$

where z is the $\alpha/2$ quantile of the standard normal distribution.

For the derivation of the Wilson interval we note from Equation C.5

$$\frac{|(N_S/N) - p|}{\sqrt{p(1-p)/N}} \leq z_{\alpha/2}, \tag{C.16}$$

this can be written as

$$\left(\frac{N_S}{N} - p\right) - \frac{p(1-p)z_{\alpha/2}^2}{N_S} \leq 0. \tag{C.17}$$

Equation is a quadratic expression in p. With the quadratic formula and $z_{\alpha/2} = z$

and $\hat{p} = N_S/N = p$ we get

$$\frac{\hat{p} + z^2/(2N) \pm z\sqrt{\hat{p}(1-\hat{p})/N + z^2/(4N^2)}}{1 + z^2/N}. \tag{C.18}$$

The zeros give endpoints for an approximate $100(1-\alpha)\%$ confidence interval for p. If n is large $z^2/(2N)$, $z^2/(4N^2)$ and z^2/N are small and the Wilson interval approaches the Wald interval.

C.6 Monte Carlo Simulation

In order to get an idea how the different intervals perform for our case of low N and extreme probabilities (close to one), a Monte Carlo simulation was performed. It simulates a coin toss with a biased coin. The bias can be expressed as the probability of the coin showing a head after being tossed p_{head}. The probability of the coin landing tail up is $p_{tail} = 1 - p_{head}$. The simulation performs one million coin tossing experiments with different biases p_{head} (ranging from 0.96–0.999 in steps of 0.001). The experiments consist of N coin tosses with $N = \{2,\ 5,\ 10,\ 20\}$. They are simulated by a random number generator, which generates a matrix with N rows and one million columns of uniformly distributed values between 0 and 1. The values that are smaller than p_{head} are counted as heads, the rest as tails. This results in curves (e.g., lines in Figure C.1) showing the number proportion of the one million experiments where not all coin tosses showed head despite the high bias, which can be interpreted as significance level α, as a function of the coin bias. By design all the curves are close to $p_{head} = 1$, in order to make the plot easier to read, the abscissa shows $1 - p_{head}$ on a log scale. In the legend $\hat{p}$ is denoted as 'p. By comparing the points where the lines from the Monte Carlo simulation cross $\alpha = 0.05$, we obtain an answer for which CI to use. The CP confidence interval lines up exactly with the simulation, and thus should be used. In addition, one can see that, the line for $N = 10$ crosses $\alpha = 0.05$ at $1 - p_{head} = 0.05$. This suggests that $N = 10$ trials should be enough to say a coin has a bias of $p_{head} = 1$ on a 5% significance level.

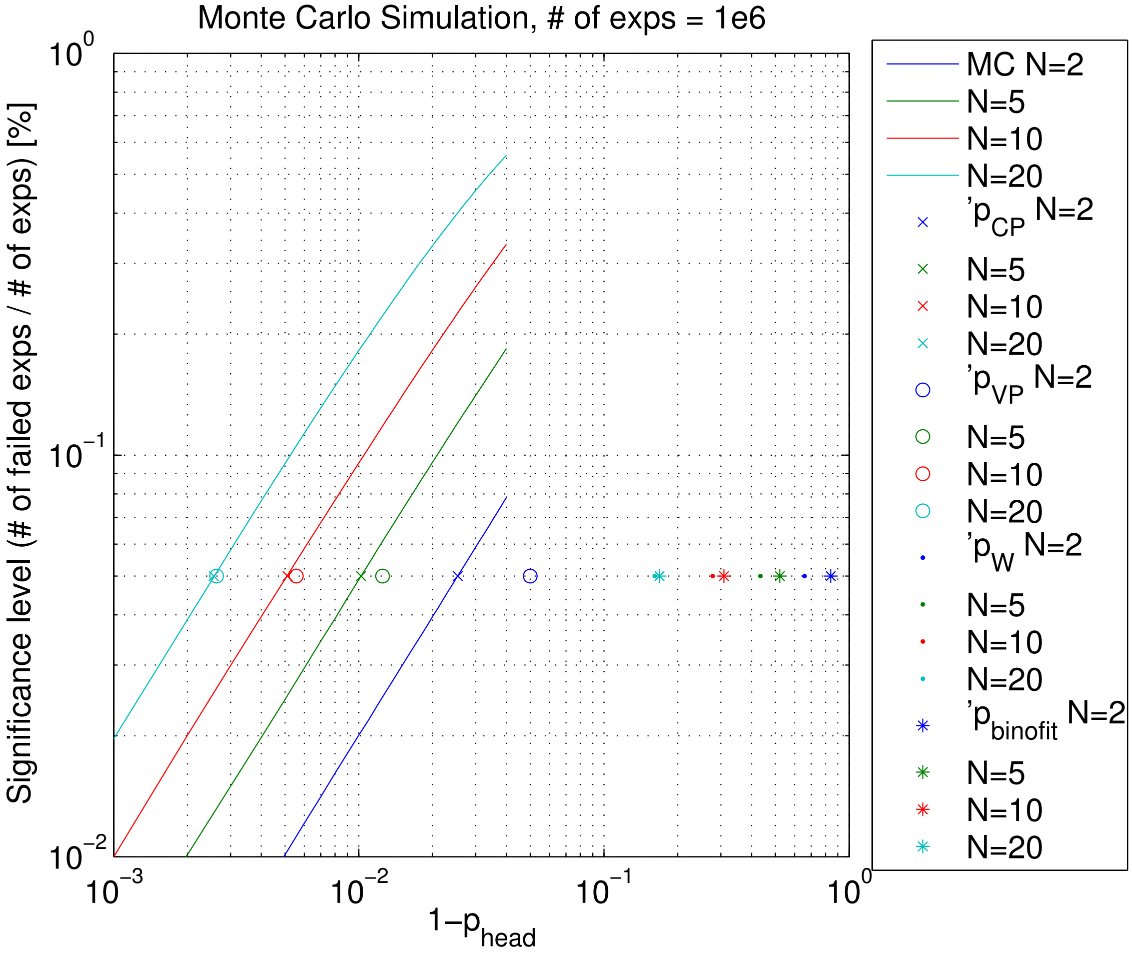


Figure C.1: Monte Carlo simulation of one million coin toss experiments consisting of different number of tosses N. Plotted is the significance level (e.g. number of experiments in which not all tosses were heads divided by the number of experiments, i.e. one million) as a function of the bias of the coin (p_{head}). Plotted along the values of the Monte Carlo simulation are the different CI estimates. In order to make the difference between the values more visible, $1 - p_{head}$ is displayed on a log scale rather than p_{head} on a linear scale.

Appendix D

Matlab Code

D.1 plsi.m

```
  function [P_wz,P_zd,P_z] = plsi(X,k,m,s)
%% Probabilistic Latent Semantic Indexing
% function [P_wz,P_zd,P_z] = plsi(X,k,m,s)
%
% Inputs: X - NxD data matrix
%         k - number of endmembers
%         m - maximum iterations allowed (default: 500)
%         s - sparsity (default: 1) (<1 gives anti-sparsity)
% Outputs: P_wz - kxD endmember matrix
%          P_zd - Nxk mixing weight matrix
%          P_z  - kx1 endmember weights matrix
%
% PLSI uses EM to maximize sum[sum[X(w,d)*log(P(w,d))]].
% This is equivalent to minimizing the total cross-entropy
% between the observed data and its projection to the convex hull.
% The hull is learned jointly with the optimal mixing weights.
% If the endmembers are known, PLSI can learn the optimal mixing weights.
% This is the process of "folding in" or "cross-entropy unmixing."
% The data error is assumed to be high-precision Dirichlet-distributed.
% Reconstruction of data is Xp = P_zd*P_wz
% Code is based on Johannes Traa implementation
%
% Fast multiplicative updates from
%   Zhong-Yuan Zhang, NMF: Models, Algorithms and Applications

[N,D] = size(X);

%% check inputs
if nargin < 3 || isempty(m); m = 500; end
if nargin < 4 || isempty(s); s = 1;   end

%% initialize
% P_wz = X(randperm(N,k),:); % k-by-D matrix of endmembers
P_wz = rand(k,D)+1; P_wz = bsxfun(@rdivide,P_wz,sum(P_wz,2));
P_dz = ones(N,k)/N; % P(d|z), N-by-k matrix of doc responsibilities
P_z = ones(k,1)/k; % P(z), k-by-1 vector of responsibilities

%% iterate
for i=1:m
    P = X./(P_dz*bsxfun(@times,P_z,P_wz)+eps);
    P1 = P_dz'*P;
    P2 = P*P_wz';
    P3 = sum(P1.*P_wz,2);
    P_wz = P_wz.*bsxfun(@rdivide,P1,P3+eps);
    P_dz = P_dz.*bsxfun(@rdivide,P2,P3'+eps);
    if s ~= 1
        P_dz = P_dz.^s;
        P_dz = bsxfun(@rdivide,P_dz,sum(P_dz,1)+eps);
    end
    P_z = P_z.*P3;
end
```

```
%% get mixing weights P(z|d) from P(d|z) via Bayes's formula
if nargout > 1
    P_zd = bsxfun(@times,P_dz,P_z');
    P_zd = bsxfun(@rdivide,P_zd,sum(P_zd,2)+eps);
end
```

D.2 simplex.m

```
 function [h E] = simplex(id)
%% set up 2D 3-simplex or 3D 4-simplex
% function h = simplex(id)
%
% Input: id - dimensionality of simplex
%               1: 2D (default)
%               2: 3D
% Output: h - fill3 handle of simplex surface (if id == 1)
% Code was written with great support from Johannes Traa

%% check input
if nargin < 1 || isempty(id); id = 1; end

%% set background color to black
figure
whitebg([1 1 1])

%% plot
h = -1;

if id == 1
```

```
    % plot 3-simplex
    h = fill3([1 0 0 1],[0 0 1 0],[0 1 0 0],’w’);
    axis([0 1 0 1 0 1])
    view(135,35)
    axis off
    set(gca,’CameraTarget’,[0 0 .25])
    zoom(1.75)
    %set(gcf,’Color’,[0 0 0])
    E=eye(3);
elseif id == 2
    % plot 4-simplex
    E = eye(4); % original endpoints
    R = [-1 -1 ; ...
          1 -1 ; ...
          0  1 ];
    E = [[E(1:3,1:3)*R+1 zeros(3,1)]; ...
         [1 1/sqrt(2) sqrt(3)/2]]; % transformed endpoints
    % to draw upward edges
    Ep = [E(1:3,:); E(1,:); E(4,:); E(2,:); E(4,:); E(3,:)];

    fill([0 1 2],[0 2 0],zeros(1,3)+0.2); % base surface
    % ensure equilateral triangle
    set(gca,’PlotBoxAspectRatio’,[2/sqrt(3) 1 1])
    plot3(Ep(:,1),Ep(:,2),Ep(:,3),’Color’,’k’)
    view(-15,10)
    axis off
    axis vis3d
    whitebg([1 1 1])
    set(gcf,’Color’,[1 1 1])
end
hold on
```

References

Alan Agresti and Brent A Coull. Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

Jont Allen and Feipeng Li. Speech perception and cochlear signal processing [life sciences]. *Signal Processing Magazine, IEEE*, 26(4):73–77, 2009.

Jont B Allen. *Articulation and intelligibility*, volume 1. Morgan & Claypool Publishers, 2005a.

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

Jont B Allen, JL Hall, and PS Jeng. Loudness growth in 1/2-octave bands (lgob)a procedure for the assessment of loudness. *The Journal of the Acoustical Society of America*, 88:745, 1990.

Shari R Baum and Sheila 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.

Susan Behrens and Sheila 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.

CG Bell, H Fujisaki, JM Heinz, KN Stevens, and AS House. Reduction of speech spectra by analysis-by-synthesis techniques. *The Journal of the Acoustical Society of America*, 33:1725, 1961.

Kenneth Walter Berger, Eric N Hagberg, and Robert L Rane. *Prescription of hearing aids: rationale, procedure, and results.* Herald Publishing House Kent, Ohio, USA, 1977.

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

Sheila E Blumstein and Kenneth N Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America*, 66:1001, 1979.

Sheila E Blumstein and Kenneth N Stevens. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *The Journal of the Acoustical Society of America*, 67:648, 1980.

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

Colin R Blyth and Harold A Still. Binomial confidence intervals. *Journal of the American Statistical Association*, 78(381):108–116, 1983.

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

Denis Byrne. Effects of frequency response characteristics on speech discrimination and perceived intelligibility and pleasantness of speech for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 80:494, 1986.

Denis Byrne and Harvey Dillon. The national acoustic laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid. *Ear and hearing*, 7(4):257–265, 1986.

Raymond Carhart. Tests for selection of hearing aids. *The Laryngoscope*, 56(12): 780–794, 1946.

Ronald A Cole and Brian Scott. Toward a theory of speech perception. *Psychological review*, 81(4):348, 1974.

Ronald A Cole, Yonghong Yan, Brian Mak, Mark Fanty, and Troy Bailey. The contribution of consonants versus vowels to word recognition in fluent speech. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 2: 853–856. IEEE, 1996.

Franklin S Cooper, Pierre C Delattre, Alvin M Liberman, John M Borst, and Louis J Gerstman. Some experiments on the perception of synthetic speech sounds. *The Journal of the Acoustical Society of America*, 24:597, 1952.

Brent A Coull and Alan Agresti. Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics*, 56(1):73–80, 2000.

RM Cox. Using ULCL measures to find frequency/gain and sspl90. *Hearing Instruments*, 34(7):17–21, 1983.

RM Cox. The MSU hearing instrument prescription procedure. *Hearing Instruments*, 39(1):6–10, 1988.

Robyn M Cox. A structured approach to hearing aid selection. *Ear and hearing*, 6 (5):226–239, 1985.

Robert M. Cvengros and Jont B Allen. A verification experiment of the second formant transition feature as a perceptual cue in natural speech. Master's thesis, 2011.

Torsten Dau, Birger Kollmeier, and Armin Kohlrausch. Modeling auditory processing of amplitude modulation. ii. spectral and temporal integration. *The Journal of the Acoustical Society of America*, 102:2906, 1997.

Hallowell Davis, CV Hudgins, RJ Marquis, RH Nichols, GE Peterson, DA Ross, and SS Stevens. The selection of hearing aids part ii. *The Laryngoscope*, 56(4):135–163, 1946.

Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

Pierre C Delattre, Alvin M Liberman, and Franklin S Cooper. Acoustic loci and transitional cues for consonants. *The Journal of the Acoustical Society of America*, 27:769, 1955.

H. Dillon. NAL-NL1: a new procedure for fitting non-linear hearing aids. *Hear J*, 52(4):1016, 1999. URL `http://frye.com/download/NAL-NL1%20article.pdf`.

Harvey Dillon. *Hearing aids*. Thieme, second edition, 2012.

Robert A Dobie. The AMA method of estimation of hearing disability: a validation study. *Ear and hearing*, 32(6):732–740, 2011.

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

Rob Drullman, Joost M Festen, and Reinier Plomp. Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95:1053, 1994a.

Rob Drullman, Joost M Festen, and Reinier Plomp. Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95:2670, 1994b.

Judy R Dubno and Harry Levitt. Predicting consonant confusions from acoustic analysis. *The Journal of the Acoustical Society of America*, 69:249, 1981.

Judy R Dubno, Donald D Dirks, and Donald 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.

Judy R Dubno, Donald D Dirks, and Amy B Schaefer. Effects of hearing loss on utilization of short-duration spectral cues in stop consonant recognition. *The Journal of the Acoustical Society of America*, 81:1940, 1987.

Peter D Eimas and John D Corbit. Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 4(1):99–109, 1973.

JL Flanagan. Parametric coding of speech spectra. *The Journal of the Acoustical Society of America*, 68:412, 1980.

Petr Fousek, Frantisek Grezl, Hynek Hermansky, and Petr Svojanovsky. New nonsense syllables database-analyses and preliminary asr experiments. In *INTERSPEECH*, 2004.

Carol A Fowler. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14(1):3–28, 1986.

Carol A Fowler. Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, 99:1730, 1996.

NR French and JC Steinberg. Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, 19:90, 1947.

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

Bruno Galantucci, Carol A Fowler, and Michael T Turvey. The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, 13(3):361–377, 2006.

Oded Ghitza. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. *The Journal of the Acoustical Society of America*, 110:1628, 2001.

BK Ghosh. A comparison of some approximate confidence intervals for the binomial parameter. *Journal of the American Statistical Association*, 74(368):894–900, 1979.

Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and audio signal processing: processing and perception of speech and music*. Wiley-Interscience, 2000.

Chris Halpin and Steven 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.

Woojae Han. *Methods for robust characterization of consonant perception in hearing-impaired listeners*. PhD thesis, University of Illinois, 2011.

Valerie Hazan and Andrew Simpson. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24(3):211–226, 1998.

Peter Heil. Coding of temporal onset envelope in the auditory system. *Speech Communication*, 41(1):123–134, 2003.

John M Heinz and Kenneth N Stevens. On the properties of voiceless fricative consonants. *The Journal of the Acoustical Society of America*, 33:589, 1961.

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

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

George W Hughes and Morris Halle. Spectral properties of fricative consonants. *The journal of the acoustical society of America*, 28:303, 1956.

Larry Humes, Troy Hackett, et al. Comparison of frequency response and aided speech-recognition performance for hearing aids selected by three different prescriptive methods. *Journal of the American Academy of Audiology*, 1(2):101–108, 1990.

Larry E Humes. Evolution of prescriptive fitting approaches. *American Journal of Audiology*, 5(2):19, 1996.

Allard Jongman, Ratree Wayland, and Serena Wong. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108:1252, 2000.

Søren Jørgensen and Torsten Dau. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. *The Journal of the Acoustical Society of America*, 130:1475, 2011.

Søren Jørgensen, Stephan D Ewert, and Torsten Dau. A multi-resolution envelope-power based model for speech intelligibility. *The Journal of the Acoustical Society of America*, 134:436, 2013.

Candace A Kamm, Donald D Dirks, and Theodore S Bell. Speech recognition and the articulation index for normal and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 77:281, 1985.

Diane Kewley-Port, T Zachary Burkle, and Jae Hee Lee. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 122:2365, 2007.

Mead C Killion and Patricia A Niquette. What can the pure-tone audiogram tell us about a patient's SNR loss? *The Hearing Journal*, 53(3):46–48, 2000.

Vern O Knudsen and Isaac H Jones. Symposium on the viiith nerve. i.basic principles underlying tests of hearing. *The Laryngoscope*, 45(1):1–23, 1935.

Sergei Kochkin. MarkeTrak VI: The VA and direct mail sales spark growth in hearing aid market. *The Hearing Review*, 8(12):16–24, 2001.

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

Feipeng Li. *Perceptual cues of consonant sounds and impact of sensorineural hearing loss on speech perception.* PhD thesis, University of Illinois at Urbana-Champaign, 2010.

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

Feipeng Li, Anjali Menon, and Jont 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.

Feipeng Li, Andrea Trevino, Anjali Menon, and Jont B. Allen. A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. *The Journal of the Acoustical Society of America*, 132(4):2663–2675, 2012.

Alvin M Liberman. Some results of research on speech perception. *The Journal of the Acoustical Society of America*, 29:117, 1957.

Alvin M Liberman and Ignatius G Mattingly. The motor theory of speech perception revised. *Cognition*, 21(1):1–36, 1985.

Alvin M Liberman, Pierre C Delattre, Franklin S Cooper, and Louis J Gerstman. The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied*, 68(8):1–13, 1954.

Alvin M Liberman, Katherine S Harris, Howard S Hoffman, and Belver C Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358–368, 1957.

Alvin M Liberman, Ignatius G Mattingly, et al. A specialization for speech perception. *Science*, 243(4890):489–494, 1989.

Alvin Meyer Liberman. *Speech: A special code.* The MIT Press, 1996.

AM Liberman, FS Cooper, DP Shankweiler, and M Studdert-Kennedy. Perception of the speech code1. *Psychological review*, 74(6):431–461, 1967.

Bryce Lobdell. *Models of human phone transcription in noise based on intelligibility predictors.* PhD thesis, University of Illinois at Urbana-Champaign, 2009.

Bryce Lobdell and Jont B Allen. An information theoretic tool for investigating speech perception. In *Ninth International Conference on Spoken Language Processing*, 2006.

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

S Lybarger. United states patent application no. *FN*, 543:278, 1944.

GA McCandless and PE Lyregaard. Prescription of gain/output (pogo) for hearing aids. *Hearing Instruments*, 34(1):16–21, 1983.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature* 264: 746-748, 1976.

George A Miller. Language and communication. McGraw-Hill, 1951.

George A Miller and Patricia E Nicely. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27 (2):338–352, 1955.

George A Miller, George A Heise, and William Lichten. The intelligibility of speech as a function of the context of the test materials. *Journal of experimental Psychology*, 41(5):329, 1951.

M Ardussi Mines, Barbara F Hanson, and June E Shoup. Frequency of occurrence of phonemes in conversational English. *Language and speech*, 21(3):221–241, 1978.

John J Ohala. Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, 99:1718, 1996.

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

Sandeep A. Phatak and Jont B. Allen. Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America*, 121(4):2312, 2007. ISSN 00014966. doi: 10.1121/1.2642397.

Judd Posner and Ira M Ventry. Relationships between comfortable loudness levels for speech and speech discrimination in sensorineural hearing loss. *Journal of Speech and Hearing Disorders*, 42(3):370, 1977.

WG Radley, WL Bragg, RS Dadson, CS Hallpike, D McMillan, LC Pocock, and TS Littler. Hearing aids and audiometers: Report of the committee on electroacoustics. *London:: His Majesty's Stationery Office*, 1947.

Marion S Régnier and Jont 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.

Robert E Remez, Philip E Rubin, David B Pisoni, Thomas D Carrell, et al. Speech perception without traditional speech cues. *Science*, 212(4497):947–949, 1981.

SO Rice. Distortion produced by band limitation of an fm wave. *Bell System Technical Journal*, 52(5):605–626, 1973.

Ross Roeser and Michael Valente. *Audiology diagnosis*. TNY, 2007.

D Schwartz, P Lyregaard, and P Lundh. Hearing aid selection for severe-to-profound hearing loss. *Hearing Journal*, 41(2):401–406, 1988.

Claude Elwood Shannon. A mathematical theory of communication. The Bell System Technical Journal, 27: 656, 1948.

Robert V Shannon, Fan-Gang Zeng, Vivek Kamath, John Wygonski, and Michael Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234): 303–304, 1995.

Irvin Shore, Robert C Bilger, and Ira J Hirsh. Hearing aid evaluation: Reliability of repeated measurements. *Journal of Speech and Hearing Disorders*, 25(2):152, 1960.

Riya Singh and Jont B Allen. The influence of stop consonants' perceptual features on the Articulation Index model. *The Journal of the Acoustical Society of America*, 131:3051, 2012.

Margaret W Skinner and James 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.

Margaret Walker Skinner. *Speech intelligibility in noise-induced hearing loss: Effects of high-frequency compensation.* PhD thesis, Program in Audiology and Communication Sciences, Washington University School of Medicine, 1976.

Guido 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.

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

Jean A Sullivan, Harry Levitt, Jian-Yih Hwang, and Ann-Marie Hennessey. An experimental comparison of four hearing aid prescription methods. *Ear and hearing*, 9(1):22–32, 1988.

Andrea Trevino and Jont B Allen. Individual variability of hearing-impaired consonant perception. In *Seminars in Hearing*, volume 34, pages 074–085. Thieme Medical Publishers, 2013a.

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

Christopher W Turner, David A Fabry, Stephanie Barrett, and Amy R Horwitz. Detection and recognition of stop consonants by normal hearing and hearing impaired listeners. *Journal of Speech, Language and Hearing Research*, 35(4):942, 1992.

Herbert B Voelcker. Toward a unified theory of modulation part i: Phase-envelope relationships. *Proceedings of the IEEE*, 54(3):340–353, 1966.

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

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

NA Watson and VO Knudsen. Selective amplification in hearing aids. *The Journal of the Acoustical Society of America*, 11(4):406–419, 1940.

Richard Wright. A review of perceptual cues and cue robustness. *Phonetically based phonology*, pages 34–57, 2004.

Zhong-Yuan Zhang. Nonnegative matrix factorization: Models, algorithms and applications. In *Data Mining: Foundations and Intelligent Paradigms*, pages 99–134. Springer, 2012.

PM Zurek and LA Delhorne. Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. *The Journal of the Acoustical Society of America*, 82:1548, 1987.