Systematic Groupings in Hearing Impaired Consonant Perception

ANDREA C. TREVINO[1] AND JONT B. ALLEN[1]

[1] *Beckman Institute, University of Illinois Urbana-Champaign*

Auditory training programs are currently being explored as a method of improving hearing impaired (HI) speech perception; precise knowledge of a patient's individual differences in speech perception allows one to more accurately diagnose how a training program should be implemented. Remapping or variations in the weighting of acoustic cues, due to auditory plasticity, can be examined with the detailed confusion analyses that we have developed at UIUC. We show an analysis of the responses of 17 ears with sensorineural hearing loss to consonant-vowel stimuli, composed of 14 English consonants followed by the vowel /ɑ/, presented in quiet and speech-shaped noise. Although the tested tokens are noise-robust and unambiguous for normal-hearing listeners, the subtle natural variations in signal properties can lead to systematic differences for HI listeners. Specifically, our recent findings have shown token-dependent individual variability in error and confusion groups for HI listeners. A clustering analysis of the confusion data shows that HI listeners fall into specific groups. Many of the token-dependent confusions that define these groups can also be observed for normal-hearing listeners, under higher noise levels or filtering conditions. These HI-listener groups correspond to different acoustic-cue weighting schemes, highlighting where auditory training should be most effective.

INTRODUCTION

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

In general, auditory training methodologies do not focus on the listener-specific

consonant recognition deficiencies (i.e., individual differences) that are present prior to the training period. Although an identical, overarching approach is desirable when initially assessing the efficacy of a training scheme, it may not be the most beneficial for providing treatment to the patient population. Our previous works [Trevino and Allen, 2013a,b] have shown that patients with mild-to-moderate hearing loss have consonant recognition errors that are usually limited to a small subset of test consonant-vowel tokens. This indicates that, for maximum efficacy and efficiency, a targeted approach is necessary in the implementation of training programs. In addition, we have explored the significant effects of talker variability on HI perception, particularly across tokens of the same consonant (i.e., within-consonant perceptual differences). These within-consonant differences, again, highlight the need for a targeted, patient-specific approach, as well as the importance of considering token variability in the analysis of perceptual data.

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

METHODS

Subjects

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

The majority of the ears in our study have slight-to-moderate hearing loss with high-frequency sloping configurations. One HI ear (14R), has an inverted high-frequency loss, with the most hearing loss $<$2 [kHz] and a threshold within the normal range at 8 [kHz]. For further listener details, including level of hearing loss, age, and most comfortable level, see Trevino and Allen [2013a,b].

Speech materials

All stimuli used in this study were selected from the Linguistic Data Consortium Database (LDC-2005S22). Speech was sampled at 16 [kHz]. Fourteen naturally-

spoken American English consonants (/p, t, k, f, s, ʃ, b, d, g, v, z, ʒ, m, n/) were used as the test stimuli. Each consonant was spoken in an isolated consonant-vowel (CV) context, with the vowel /ɑ/. Two tokens were selected (1 male and 1 female talker) for each consonant, resulting in a total of 28 test tokens (14 consonants x 2 talkers = 28 tokens). The term *token* is used throughout this work to refer to a single CV speech sample from one talker.

The 28 test tokens were selected based on their NH perceptual scores in Quiet and speech-weighted noise. To ensure that tokens were unambiguous and robust to noise, each token was selected based on a criteria of $\leq 3.1\%$ error for a population of 16 NH listeners, calculated by combining results in Quiet and -2 [dB] signal-to-noise ratio (SNR) of noise (i.e., no more than 1 error over a total N=32, per token) [Phatak and Allen, 2007]. Such tokens are representative of the LDC database; Singh and Allen [2012] shows, for the majority of tokens, a ceiling effect for NH listeners ≥ -2 [dB] SNR. One token of /fɑ/ (male talker, label m112) was damaged during the preparation of the tokens, thus it has not been included in this analysis.

The stimuli were presented with flat gain at the *most comfortable level* (MCL) for each individual HI ear. For the majority of the HI ears the MCL was approximately 80 ± 4 [dB] SPL; only two subjects did not choose an MCL within this range (36L/R chose 68/70 [dB] SPL and 14R chose 89 [dB] SPL).

Experimental procedure

The speech was presented at 4 SNRs (0, 6, 12 [dB] and Quiet) using speech-weighted noise, generated as described by Phatak and Allen [2007]. Presentations were randomized over consonant, talker, and SNR. The total number of presentations for each consonant ranged from N =40–80 for each HI ear (total N =5–10 over two adaptive phases x 2 tokens x 4 SNRs). The Vysochanskiĭ–Petunin inequality was used to verify that the number of trials were sufficient to determine correct perception within a 95% confidence interval, as described in the appendix of Singh and Allen [2012].

All of the data-collection sessions were conducted with the subject seated in a single-walled, sound-proof booth. The speech was presented monoaurally via an Etymotic ER–3 insert earphone. The contralateral ear was not masked or occluded. The subject chose their MCL (for non-test speech samples) before testing began. A practice session, with different tokens from those in the test set, was run first in order to familiarize the subject with the testing paradigm and to confirm their MCL setting. After hearing a single presentation of a token, the subject would choose from the 14 possible consonant responses by clicking one of 14 CV-labeled buttons on the graphical user interface, with the option of up to 2 additional token repetitions, to improve accuracy. Additional experimental details are provided in Han [2011] and [Trevino and Allen, 2013a,b].

Data Analysis

The variability of naturally-spoken acoustic cues can lead to HI within-consonant differences in both error and consonant confusions [Trevino and Allen, 2013a,b]; therefore, calculations at the token level are necessary in any analysis that attempts to understand how a HI listener is using and interpreting the acoustic cues that are available to them. In this paper, the data is analyzed at the token level, with individual data points for the HI ears.

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

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

$$< \mathbf{x}, \mathbf{y} >= \sum_i x_i y_i$$

and the norm as

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

Then the square of the Hellinger distance

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

Thus, the spherical k-means algorithm, which forms groups based on $1 - cos(\Theta_{xy})$ between points distributed on the unit sphere, produces results that also minimize the

Hellinger distance. The spherical k-means clustering algorithm is implemented in MATLAB, with the *kmeans()* function. For each token, one of the clusters is always composed of the data points where HI listeners correctly perceived the consonant; the remaining clusters are composed of the data with varying degrees of error. Therefore, assuming there are errors, the minimum possible K for a token is 2.

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

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

RESULTS

For a fixed consonant token, HI listeners vary widely in both the degree of error and the SNR threshold at which errors begin to occur. Despite this individual variability, we have observed that different HI ears tend to have similar token-dependent confusions once an error is made [Trevino and Allen, 2013b]. If HI listeners generally share a similar confusion group for a particular token, then an auditory training scheme that corrects for this confusion should be effective for a broad population of patients. In order to explore the extent of the similarities across HI listeners, we use the spherical k-means clustering algorithm to group the listeners based on confusions. The data at all tested SNRs is used together in the k-means clustering analysis, since the different severities of hearing impairment across the many listeners leads to errors at different SNRs.

Each cluster identified by the k-means algorithm is composed of listeners with similar consonant confusions. The number of clusters, K, for each token is determined by the G-means algorithm, which selects K iteratively based on a statistical test of the cluster distributions [Hamerly and Elkan, 2004]. As a result of incorporating the statistical test, the number of resulting clusters K is the amount of significantly different confusion groups that are present in our data. For example, the case of two resulting clusters, $K = 2$, indicates that all of the listener data is distributed within the cluster of correct-response data points and a second cluster defined by a single confusion group. From the results in Table 1, we see that 17 out of the 27 tokens have $K \leq 3$, indicating that all of the HI data for these tokens falls into one of 3 confusion-based clusters. 22 out of 27 tokens have $K \leq 4$. This small number of clusters for

the majority of tokens indicates that, generally, only a few token-dependent confusion groups are present in the HI data.

For each cluster, the primary confusions that define the kth mean, along with the number N of data points within each cluster, are included in Table 1. Results for the cluster of 'correct' responses (i.e., the cluster of data with no more than 1 error over 5-10 trials) are also included. From the results in Table 1, we see that the confusions that define the clusters can vary across tokens of the same consonant. For example, /d, g, v/ confusions are present for the female /bɑ/ token, while only /v/ confusions dominate the responses for the male /bɑ/ token. In addition, the large number of data points, N, in the 'correct' clusters of all tokens indicates that the mild-to-moderate HI listeners in this study did not have widespread errors. These are observations that have been made previously in Trevino and Allen [2013b]; this analysis shows how these observations can also be made from the results of k-means clustering.

The extent of the similarity across listener responses can be quantified by the angle between the points in the spherical vector space. These angles can be expressed as direction cosines or Hellinger distances, as described in the Methods section, and can range from 0 °to 90°. The angle Θ_{x,μ_k} between a data point $\mathbf{x}$ in the kth cluster and the kth cluster mean μ_k provides a measure of how well each mean represents the overall group of data points. This average of this measure, $\widehat{\Theta}_{x,\mu_k}$, is analogous to the variance within each cluster. Results for $\widehat{\Theta}_{x,\mu_k}$ are shown in Table 1. For reference, when each data point $\mathbf{x}$ is the result of 5-10 presentations, as ours are, an angle of 18°-27°lies between a vector of correct responses and a vector with a single incorrect response. Overall, the clusters defined by a larger number of primary confusions have larger $\widehat{\Theta}_{x,\mu_k}$ values. Systematic groupings of HI data in terms of consonant confusions is observed for all the tested tokens.

DISCUSSION

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

CV	k^{th} Mean (N)	$\widehat{\Theta}_{x,\mu_k}$		CV	k^{th} Mean (N)	$\widehat{\Theta}_{x,\mu_k}$
ba$_{F101}$ $K=2$	k_1 : correct (39) k_2 : b, d, g, v (29)	12° 36°		**ba**$_{M112}$ $K=4$	k_1 : correct (32) k_2 : b, v (21) k_3 : b, v (9)	15° 27° 19°
da$_{F105}$ $K=3$	k_1 : correct (61)	10°		**da**$_{M118}$ $K=2$	k_1 : correct (61) k_2 : d, g, t (7)	10° 25°
fa$_{F109}$ $K=2$	k_1 : correct (39) k_2 : f, s, v (29)	14° 34°		-		
ga$_{F109}$ $K=2$	k_1 : correct (35) k_2 : b, d, f, g, v (33)	8° 48°		**ga**$_{M111}$ $K=4$	k_1 : correct (54)	10°
ka$_{F103}$ $K=3$	k_1 : correct (50) k_2 : k, p, t (11) k_3 : t (7)	11° 25° 22°		**ka**$_{M111}$ $K=2$	k_1 : correct (56) k_2 : k, t (12)	9° 23°
ma$_{F103}$ $K=3$	k_1 : correct (46) k_2 : m, v (12) k_3 : m, n (10)	11° 28° 26°		**ma**$_{M118}$ $K=2$	k_1 : correct (61) k_2 : m, n, v (7)	9° 16°
na$_{F101}$ $K=4$	k_1 : correct (52) k_2 : m, n (9)	10° 25°		**na**$_{M118}$ $K=4$	k_1 : correct (43) k_2 : m, n (15)	12° 4°
pa$_{F103}$ $K=6$	k_1 : correct (59)	13°		**pa**$_{M118}$ $K=2$	k_1 : correct (61) k_2 : f, p, t, z (7)	12° 35°
sa$_{F103}$ $K=3$	k_1 : correct (55) k_2 : s, ʒ, z (7)	11° 26°		**sa**$_{M120}$ $K=5$	k_1 : correct (45) k_2 : s, z (11)	11° 10°
ta$_{F108}$ $K=2$	k_1 : correct (61) k_2 : f, p, s, t (7)	6° 40°		**ta**$_{M112}$ $K=2$	k_1 : correct (62)	6°
va$_{F101}$ $K=3$	k_1 : correct (43) k_2 : f, v (15) k_3 : b, d, m, n, v (10)	11° 32° 38°		**va**$_{M118}$ $K=7$	k_1 : correct (29) k_2 : p, v (12) k_3 : m, n, v (11)	14° 25° 28°
ʃa$_{F103}$ $K=2$	k_1 : correct (60) k_2 : s, ʃ, z (8)	7° 24°		**ʃa**$_{M118}$ $K=2$	k_1 : correct (65)	6°
ʒa$_{F105}$ $K=4$	k_1 : correct (42) k_2 : z (16)	11° 18°		**ʒa**$_{M107}$ $K=3$	k_1 : correct (36) k_2 : g, ʒ, z (17) k_3 : v, ʒ, z (15)	13° 32° 38°
za$_{F106}$ $K=7$	k_1 : correct (35) k_2 : ʒ, z (11) k_3 : s, ʒ, z (8)	14° 9° 19°		**za**$_{M118}$ $K=6$	k_1 : correct (38) k_2 : ʒ, z (11) k_3 : v, ʒ, z (9)	14° 18° 20°

Table 1: Clustering results for 27 CV tokens, talker gender and identification number is indicated by the CV subscript. The resulting total number of clusters K is included in the CV column. Each row shows the data for a single cluster; to focus on clusters with similar listeners, clusters with less than 6 data points are omitted. The main confusions comprising the kth cluster means ($>5\%$) are listed under k^{th} Mean (N), with N being the number of data points within each cluster (out of 68 total). Similarities across HI ears within a cluster are quantified by the average angle, between the members of each cluster and the kth mean, $\widehat{\Theta}_{x,\mu_k}$.

each HI individual.

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

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

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

Greg Hamerly and Charles Elkan. Learning the k in kmeans. *Advances in neural information processing systems*, 16:281, 2004.

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

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

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

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

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

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.

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