Effects of NALR on consonant-vowel perception

CHRISTOPH SCHEIDIGER* AND JONT B. ALLEN

Human Speech Recognition Group, University of Illinois, Urbana, IL, USA

Consonant vowel (CV) identification experiments in masking noise with 16 hearing-impaired (HI) ears at two different gain conditions, i.e., flat-gain (FG) and spectral correction (National Acoustic Laboratory Revised prescriptive procedure, NALR), were administered (Han, 2011). In both gain conditions, listeners were directed to adjust the presentation level to their most comfortable loudness (MCL). MCL testing runs 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) (Posner and Ventry, 1977; Zurek and Delhorne, 1987). The results, however, prove that for speech testing MCL is justified. A more rigorous definition for audibility based on entropy in recognition experiments is provided. Furthermore, the effectiveness of NALR for CV perception is investigated. The average error went down from 20.1% ($\sigma = 3.7$) to 16.3% ($\sigma = 2.8$). For 50.5% of the token[1]-ear pairs (TEPs) the error and entropy both went down, while for 15.1% of the TEPs the entropy and error went up with NALR. In order to evaluate statistically siginificant effects of NALR, the confusion matrix data were clustered, and the number of ears which switched clusters when NALR was applied were investigated. In addition, the subjects' confusions under both conditions were studied and compared to the confusions of other HI and normal-hearing (NH) subjects.

INTRODUCTION

The goal of this research is to better understand speech perception in hearing-impaired ears. The human speech recognition (HSR) group at the University of Illinois at Urbana-Champaign takes the approach to look at CV recognition tasks of NH as well as HI subjects. CVs are chosen, as opposed to words, phrases, or sentences in order to reduce the influence of higher-order (context) processing in the auditory pathway, which permits the control of differences in cognitive abilities (e.g., memory, semantics) (Miller *et al.*, 1951). A second goal of this paper is to address what audibility means in speech perception experiments and to determine how it can be verified. Lastly, we will show that, despite its high variability, speech as a test for hearing loss and hearing-aid evaluation can deliver more detailed insights than the commonly used pure tones, which is in contrast to what Walden *et al.* (1983) and Zurek and Delhorne (1987) suggested.

*Corresponding author: csche@elektro.dtu.dk

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

Proceedings of ISAAR 2013: Auditory Plasticity – Listening with the Brain. 4th symposium on Auditory and Audiological Research. August 2013, Nyborg, Denmark. Edited by T. Dau, S. Santurette, J. C. Dalsgaard, L. Tranebjærg, T. Andersen, and T. Poulsen. ISBN: 978-87-990013-4-7. The Danavox Jubilee Foundation, 2014.

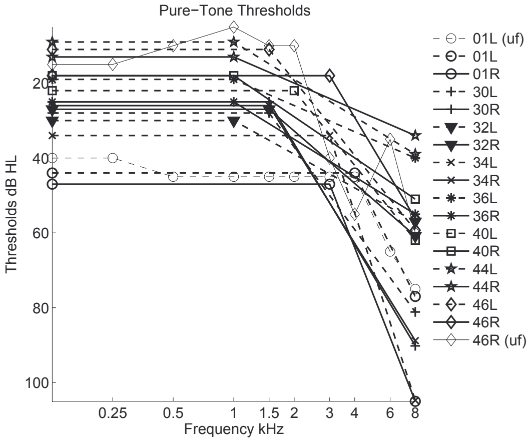

Fig. 1: PTTs for the sixteen ears

HI ear	Age	PTA	MCL	
			FG	NALR
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)

Table 1: Subjects' age, PTAs and MCLs (dB SPL)

METHODS

The two conditions (i.e., FG and NALR) were administered as separate experiments (Han, 2011). Each of the 8 subjects (16 HI ears) passed a middle-ear examination and their hearing thresholds were measured before each experiment. All 16 ears had mild-to-moderate hearing loss. Fig. 1 shows the fitted PTTs according to Trevino (2013).

The CV syllables consisted of 14 consonants (6 stops, 6 fricatives, and 2 nasals) followed by /a/. Two talkers (1 male and 1 female) were selected per consonant. The tokens were chosen from those for which there was less than 3% error at SNRs < -2 dB in previous NH experiments. The male tokens for /f, s, ʒ, n/ + /a/ were removed from the analysis, because they had to be changed between the two conditions, leaving 12 CVs for comparison (24 tokens). The tokens used for the experiments are well characterized: the perceptual cues have been identified by the 3DDS method (Li *et al.*, 2012) and the CMs at 6 SNRs were previously determined in both white noise and speech weighted noise.

The subjects were able to adjust the presentation level at any time during the experiment, however, as seen in Table 1 only 40L made use of this option. All of the subjects had one practice session before they began the experiment. Syllable presentation was randomized over consonants, speakers, and SNRs (12, 6, and 0 dB, plus quiet). For each condition, SNR and subject, a token was presented between 5 and 10 times (depending on the error); this resulted in 800-1000 trials per subject.

RESULTS

The resulting *confusion matrices* (CM) were analyzed using the following tools.

Entropy

Information theory and entropy were introduced by Shannon (1948). Miller and Nicely (1955) were the first to apply an entropy analysis to speech confusion data. Entropy, a measure of the randomness of a response, is defined as the expected value of the information ($\log(1/p_i)$), the CM row sum is $\sum_i p_i = 1$, where $i = 1 \ldots 14$ (14 is the number of possible responses):

$$\mathscr{H}(\mathbf{p}) = \sum_{i=1}^{I} p_i \log_2 \left(\frac{1}{p_i} \right). \qquad \text{(Eq. 1)}$$

Audibility: Posner and Ventry (1977) found that subjects perform below their maximum speech discrimination abilities if tested under MCL conditions. The data, however, suggest that most tokens were fully audible to all the subjects under both conditions. We suggest that calculating the entropy in quiet is a more meaningful audibility measure for CV identification experiments than LTASS and PTTs. Low entropy implies consistency, which is a strong test of audibility, even if the error is high (cf. Fig. 2 (a)).

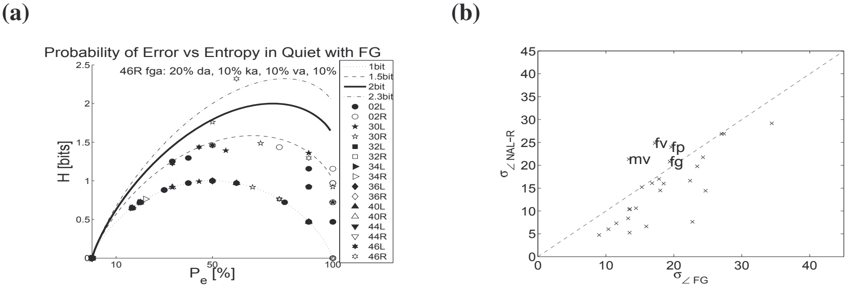

Fig. 2: **(a)** P_e vs $\mathscr{H}$ plot for all subjects and tokens in quiet (FG condition). Entropy is low, even though in many cases the error is high, thus audibility is not an issue. The 2-bit curve is a reasonable audibility threshold based on the Miller and Nicely (1955) confusion groups. According to this definition only the female token of /gɑ/ is not audible for subject 46R. All the other sounds are audible for all subjects. **(b)** The standard deviation of the angles between the correct response and the ears decreases with NALR in all but the four labeled cases (fg, fp,fv, mv) subjects responded more consistently, (mv = male /va/ token).

Effects of NALR: 24 tokens can be compared between the two experiments and 16 ears. This results in 384 cases, when collapsed over SNR. Those cases can be categorized according to how the entropy and error changed from the FG to the NALR experiment. Most of the cases (50.5%) are the cases where NALR decreased both the entropy and error. The second largest group is the one where NALR increased both the entropy and error (15.1%). The other two categories only contain the few remaining TEPs (cf. Fig. 3).

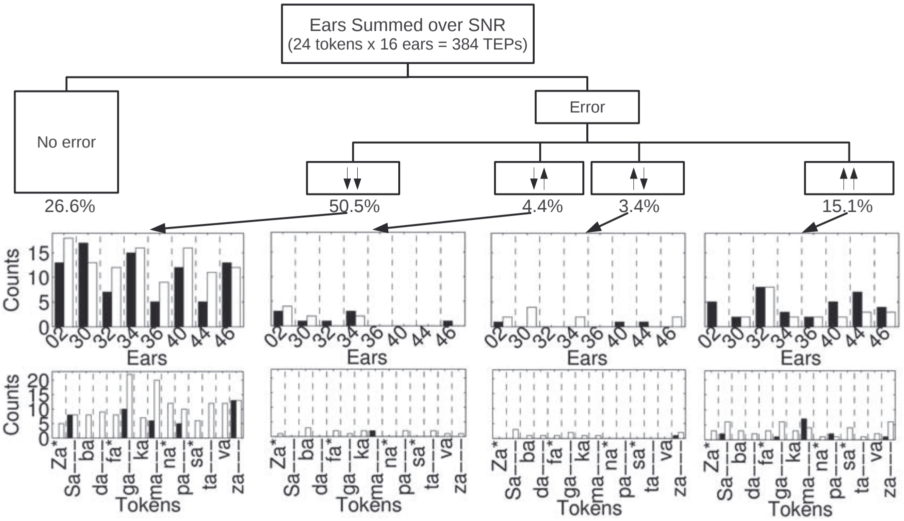

Fig. 3: Categorization of the CV perception data for the 24 tokens and 16 listeners collapsed over four SNRs. For 102 (26.6%) of the 384 TEPs there was zero error in both conditions. The remaining 282 TEPs are grouped into one of 4 major categories, in the category labels the first arrow indicates what affect NALR had on the entropy, the second one indicates what happen 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*).

Direction cosine

Every confusion matrix defines a vector space, where each row is a vector in that space. In order to find the distance between two tokens (rows), a norm must be defined: we chose a metric called the *Hellinger Distance* (HD), which uses the square roots of the probability vectors **p**. Via *Schwartz's inequality*, it is possible to calculate an angle θ_{lm} between any two tokens in the vector space. The angle is a measure of how different two response vectors are. The HD can also be used to measure the difference between the two experiments or between a listener's response and the correct answer.

The HD seems to be an underutilized measure for the analysis of CMs.

$$cos\ \theta_{lk} = \mathbf{p}_l \cdot \mathbf{p}_m = \sum_{i=1}^{I} \sqrt{p_{l,i}}\sqrt{p_{m,i}} \qquad \text{(Eq. 2)}$$

Confusions: The angle between the correct answer and the response is a measure of the change in confusions. The mean angle ($\mu_{\angle}$) and the standard deviation ($\sigma_{\angle}$) of the angles for one token are expected to decrease if the ears become more accurate in their response or if they become more consistent in their answers, respectively.

NALR 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 2 (b) one can see that the variance of the angles ($\sigma_{\angle}$) goes down with NALR in all but 3 cases: fv (i.e., female /va/), fp and fg. The mean angle ($\mu_{\angle}$) goes down with NALR for all 24 tokens.

K-means clustering

Once normed vector space is defined, the elements in this space may be clustered. For each of the 24 tokens, there are $2 \times 4 \times 14 = 112$ (2 conditions, 4 SNRs and 14 ears) data points in the fourteen dimensional space. The *k-means* algorithm is then 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 $\mathbf{c}_k$s are then sorted according to their entropy (Eq. 1).

Classifying NALR: 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 NALR. 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 NALR, are considered cases where NALR worked. 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 NALR are cases where NALR 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: $\mathbf{c}_y \rightarrow \mathbf{c}_x$, $x < y$).

In the k-means analysis, listeners' responses are not collapsed over SNR, but they are grouped according to proximity in the vector space. Restricting to $K = 4$ clusters helps to come to statistically more meaningful results. 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. When examining all 1568 cases ($4 \times 14 \times 24 = 1344$), 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%.

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

Table 2: 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.

Comparison to NH subjects

Table 2 shows split up by token (i) how many ears have a sufficient number of errors in order to be considered significant, (ii) what the main confusions are for both experiments (are they consistent across ears?), (iii) how the entropy of the listeners change with NALR and (iv) for how many subjects the two ears are remarkably different (as measured by angle between the responses), (v) what the average error is for the token.

For each token it is interesting to know (i) how many ears have a sufficient number of errors in order to be considered significant, (ii) what the average error is for the token, (iii) what the main confusions are for both experiments (are they consistent across ears?), (iv) if the confusions that were made in the NALR experiment were expected (same Miller and Nicely confusion groups, expected from the normal-hearing 3DDS data of the particular token), (v) how the entropy of the listeners change with NALR, and (vi) for how many subjects the two ears are remarkably different (as measured by angle between the responses). The results for all 24 tokens are summarized in Table 2.

CONCLUSIONS

Audibility

Despite the uncommon approach of measuring CV confusions at MCL, the data demonstrates, based on the low entropy in quiet, that audibility was not an issue. Audibility is not rigorously defined. Given the results of our CV recognition experiments, we propose the use of entropy as means of defining 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 found to be bursts or frequency edges (Li, 2010; Li *et al.*, 2012), whereas the long-time speech spectrum is dominated by vowels.
2. CV perception is binary: the acoustic speech cue either can be heard or cannot be heard (Singh and Allen, 2012).
3. PTTs do not characterize the audibility of acoustic speech cues as indicated by the 3DDS method (Li, 2010). PTTs for example are an inadequate predictor of the audibility of a plosive burst, which can be much more intense than the LTASS in a critical band over a few centi-seconds (Wright, 2004).

From the reasoning stated above, it follows that a sound with 100% error ($\mathscr{H} = 0$ bit) must be audible. This is plausible since the ear must be listening to some signal properties, otherwise it would not be so consistent. On the other hand, a listener who responds randomly across all 14 consonants has $P_e = 0.93$ and $\mathscr{H} = 3.8$bits, indicating the listener cannot hear the signal. 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 3, therefore a response with 3 equally likely responses can be taken as an audibility threshold. The subject is most likely guessing when confusions outside of a known confusion group appear. In Fig´. 2 (a) the 2-bit curve representing the audibility threshold is plotted thicker. Only one point (ear 46R female /gɑ/) lies above the line, for all the other ears and tokens audibility can be assumed not to be the problem.

Effects of NALR

NALR generally, decreases the entropy (see *NALR* column in Table 2, Fig. 3 and also, the k-means result). This means the responses with NALR, show on average smaller confusion groups. The ears become more consistent in their responses, based on the decreasing standard deviation $\sigma_{\angle}$, which means the angles in the 14 dimensional space between the responses and the correct answer become more similar for all ears. In addition, the responses become closer to the correct answer, since the mean angle ($\mu_{\angle}$) of all ears per token decreases with NALR. Therefore, NALR not only decreases the randomness of the answers but also causes the ears to agree more on a token basis. This gives hope for training of listeners with their specific problems, since they all seem to agree on the signal they hear. Given the presented data, we have demonstrated the effectiveness of NALR using a speech test instead of pure tone tests. This suggests that a carefully constructed speech test can be used as a diagnostic tool: From the results listed in Table 2, we know all listeners for whom CV tokens cause problems and therefore can get detailed information about their hearing loss. Carefully characterized CVs can be used to find specific problems in HI subjects, that PTTs cannot.

REFERENCES

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

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

Li, F., Trevino, A., Menon, A., and Allen, J.B. (**2012**). "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," J. Acoust. Soc. Am., *132*, 2663-2675.

Miller, G.A., Heise, G.A., and Lichten, W. (**1951**). "The intelligibility of speech as a function of the context of the test materials," J. Exp. Psychol., **41**, 329-335.

Miller, G.A., and Nicely, P.E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am., **27**, 338-352.

Posner, J., and Ventry, I.M. (**1977**). "Relationships between comfortable loudness levels for speech and speech discrimination in sensorineural hearing loss," J. Speech Hear. Disord., **42**, 370-375.

Shannon, C.E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J., **27**, 379-423.

Singh, R., and Allen, J.B. (**2012**). "The influence of stop consonants' perceptual features on the Articulation Index model, J. Acoust. Soc. Am., **131**, 3051-3068.

Trevino, A. (**2013**). *Techniques for understanding hearing impaired perception of consonant cues*. PhD thesis, University of Illinois at Urbana-Champaign.

Walden, B.E., Holum-Hardegen, L.L., Crowley, J.M., Schwartz, D.M., and Williams, D.L. (**1983**). "Test of the assumptions underlying comparative hearing aid evaluations," J. Speech Hear. Disord., **48**, 264-273.

Wright, R. (**2004**). "A review of perceptual cues and cue robustness," in *Phonetically-Based Phonology*. Edited by B. Hayes, R. Kirchner, and D. Steriade (Cambridge University Press), pp. 34-57.

Zurek, P., and Delhorne, L. (**1987**). "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," J. Acoust. Soc. Am., **82**, 1548-1559.