50 Years Late: Repeating Miller-Nicely 1955

Andrew Lovitt, Jont Allen

Department of Electrical and Computer Engineering
University of Illinois, Urbana, Il USA
lovitt@uiuc.edu, jontalle@uiuc.edu

Abstract

Portions of the procedure and analysis of the wide-band noise masking experiment in Miller-Nicely's 1955 JASA paper (MN55) was repeated and a new set of data was collected in 2005. This classic paper is a commonly referenced work in which confusion matrices were collected for a set of consonant-vowels (CVs). From an analysis of the original results, they made conclusions about the robustness of various distinctive features when the CVs are degraded in masking noise. Our repeat experiment shows a number of similarities and differences. The two experiments show significantly different amounts of relative information transmitted for each distinctive feature. In the repeat experiment the voicing feature is less robust whereas the place feature is more robust.

Index Terms: phone recognition, distinctive feature, confusion matrix, confusion groups, voicing.

1. Introduction

In 1955 Miller and Nicely [1] (MN55) analyzed the specific confusion patterns for a set of consonant-vowels (CV) masked with wideband noise. Their work consisted of collecting confusion matrices (CM) and analyzing the results using information theory methods. These results showed that as the SNR was lowered, for each CV spoken, subjects only chose from a certain subset of CVs for the response. This analysis showed that certain distinctive features are more robust to degration in noise. These results were then reinterpreted by Soli *et al.* [2], Shepard [3], and by Allen [4]. This new experiment will be referred to as MN05.

1.1. MN55 Procedures

The MN55 procedure consisted of presenting CVs to a subject, who then reported what they heard. Each presentation was a talker saying a certain CV. The subjects listened to the presentation over headphones. The signal-to-noise ratio (SNR) of the presentations was varied. The SNRs tested were -18, -12, -6, 0, 6, and 12 dB SNR. The SNR was created by keeping the speech at a constant level and adjusting the level of the masking noise. The count matrix is composed of entries ($N_{h|s}$) in which each entry is the number of responses based on what was said (where h is the CV heard, s is the CV uttered). These matrices were then row normalized to produce a CM where each element was the probability of the response ,h, given the spoken CV ,s, (i.e. $P_{h|s}(\mathrm{SNR})$).

The CVs presented in MN55 consisted of the consonants /p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/, and /n/ followed by the vowel of /ɑ/. The talker was isolated from the subjects and spoke into a microphone which was connected to a circuit which added the wideband masking noise. The talker would read a list of 200 CVs which were randomized so the probability of any CV being spoken was $\frac{1}{16}$ over the entire test. Once 200 CVs were read the talker was switched. All listeners in the experiment were also talkers. The listeners were trained extensively before data was collected. At any one time there were only 5 subjects.

The system had a frequency response that was flat over the 200-6500 Hz range. Additionally the noise was low pass filtered to 7000 Hz before the SNR was set. In this setup the noise was constantly being added to the sound from the microphone. When the SNR was low a tone was used to synchronize all the listeners before the each CV was said. The subjects listened to the presentations through headphones. There was an average of 2.1 seconds between each spoken CV.

1.2. MN05 System Review

Only the differences between the MN55 and MN05 procedures will be discussed here. The MN05 testing is conducted using a program run in Matlab©. The stimulus are presented in a sound booth through an attenuator circuit that limited the presentation level to a maximum of 80 dB SPL through the headphones. The subject is seated in front of the computer and the person administered the test unsupervised with all University of Illinois at Urbana-Champaign Institutional Review Board procedures followed. There were 2 SNRs added for the MN05 testing which were quiet and -15 dB SNR. The -15 dB SNR was added because in MN55 the most interesting confusions are between -12 and -18 dB SNR.

All CVs were taken from the LDC nonsense speech corpus.[1] Eighteen talkers were chosen from this database and each talker spoke each CV once. These were limits imposed by the database choice. The subject pool consisted of 26 subjects from the area surrounding the University of Illinois at Urbana-Champaign. They all spoke English as a first language and had no ear infections or hearing problems. The subjects were trained for an hour by listening to presentations with no masking noise added. This familiarized the subjects with the experiment. No further training was given to prevent overtraining the subjects. The subjects listened to one block at a time. A block is 18 sounds from one talker at one SNR in which each CV is spoken at least once. This was done to simulate as closely as possible the procedure from MN55.

After the stimulus was presented, the subject clicked a button in the graphical user interface (GUI) corresponding to the CV the subject heard. The subject was allowed to proceed at their own pace, without limits on the response time. The subject was given the ability to repeat the stimuli multiple times as desired. The average number of presentations grew from 0.42 repeats per a trial in quiet to 2.34 at -18 dB SNR.

The subject was also offered a button denoted 'noise only'.

[1] LDC Articulation Index database number LDC2005S2

The subject was instructed to only press this button if they repeated the sound more than once and could not hear anything. If the subject heard something they were instructed to make their best guess. 'Noise only' responses were recorded as $\frac{1}{16}$th of a hit for each response. The percentage of 'noise only' responses is 45% at -18 dB SNR, 12% at -15 dB SNR, and only 1% at -12 dB SNR. This button was included so that at very low SNRs when nothing was heard subject biases would not influence the data.

In MN55 the authors argued for the existence of distinctive features as an explanation of groupings. All distinctive features in this paper are identical to the distinctive features from MN55. For the purposes of this paper there are three 'voicing' and 'nasality' groups, which are the 'voiced non-nasals' (v), 'unvoiced non-nasals' (uv), and 'voiced nasals' (n). The following notation will be used in place of the IPA symbols on the graphical results: /θ/ is /th/, /ʃ/ is /sh/, /ð/ is /dh/, and /ʒ/ is /zh/.

2. Confusion Patterns

For this analysis no effort was made to identify the mispronounced utterances, poor talkers, or poor subjects, because this was not done in the original experiment. The data sets have slightly different confusion patterns. At the highest SNR in MN05, the CVs, which have errors, are predominately confused only with other CVs which have errors. These errors are predominately found in the $P_{uv|v}$ and $P_{v|uv}$ sections. Specifically these errors are in the detection of the voicing feature. In Fig. 1 the confusion matrix for both experiments is shown for the -6 dB SNR case. The grid lines drawn are to delineate the UV, V, and N groupings in both the spoken CV and response CV. The columns and rows are ordered as they are in MN55.

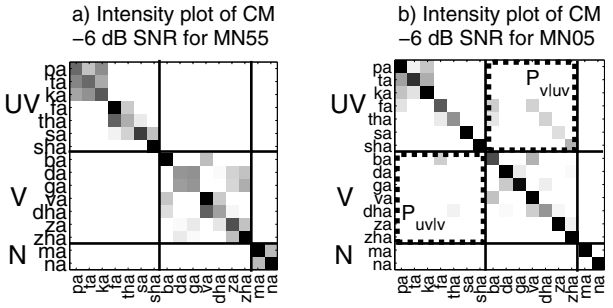

Figure 1: The intensity plots show the confusion matrices for -6 db SNR for both data sets. Figure (a) is the data from the MN55 data and figure (b) is the data from the MN05 data. The intensity is proportional to the $P_{h|s}$ for each cell. The white background corresponds to chance performance. Each row corresponds to a specific s which is the CV said be the talker. Each column corresponds to a specific h which is the responses of the subject. The most obvious difference between the two data sets are the voicing errors. Specifically the $P_{v|uv}$ and $P_{uv|v}$ errors that appear in the MN05 data. These errors are not inherent to the MN55 data.

It is apparent from Fig. 1 that the MN55 ordering of the rows and columns is not the optimal ordering for grouping primary confusions in MN05. This is evident from the amount of responses in the $P_{v|uv}$ and $P_{uv|v}$ blocks in Fig. 1b. If the ordering was optimal, the probability mass would be concentrated near the diagonal of the CM.

The original ordering from MN55 resulted from the primary confusion group members, which was ordered so that the confusion group members were next to each other. Since the ordering of the CM is sub-optimal for the MN05 data a new ordering is found which groups the sounds according the their confusions in MN05.

In order to judge the fitness of any ordering the following Manhattan (taxi-cab) distance metric (Eq. 1) is employed. This metric weights the off-diagonal elements in proportion to their distance from the diagonal. A small value will signify that the majority of the responses are close to the diagonal. If the matrix is uniform the metric will be maximal (85 for a confusion matrix with 16 CVs).

$$W(\text{SNR}) = \sum_{1\le i\le 16}\left(\sum_{1\le j\le 16}|i-j|P_{j|i}(\text{SNR})\right) \qquad (1)$$

The desire is to develop a new ordering for each data set which is optimal over all SNRs. First all the count matrices ($N_{h|s}(\text{SNR})$) were summed over SNR. Next all CVs were analyzed to find their 'primary groupings'. A 'primary group' is when all members of the group have the highest number of responses for each CV in the group. For instance if /pɑ/, /tɑ/, and /kɑ/ are a 'primary group', then when /pɑ/ is said the top 3 responses are /pɑ/, /tɑ/, and /kɑ/. In this way all the 'primary groups' are found. A list was then made of all possible orderings where all members of a 'primary group' are next to each other. Then these possible orderings are brute-force tested and the best ordering was kept.

The results of these new orderings are seen in Fig. 2. The new orderings were found which are unique and minimize the taxi-cab metric (Eq. 1). Plot (a) is the MN55 data for the new MN55 ordering. Likewise plot (b) is the MN05 data for the new MN05 ordering.

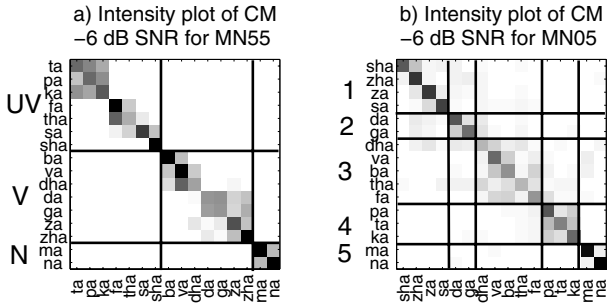

Figure 2: These plots are similar to the plots in fig. 1 except the order of the rows and columns are arranged by the new orderings. Figure (a) is the data from MN55 at -6 dB SNR where the ordering is optimal for the MN55 data. Likewise figure (b) is the data from MN05 for -6 dB SNR for the ordering which is optimal for MN05. The lines drawn delineate the voicing and nasality groups for MN55 and the major confusion groups in MN05.

The new ordering for the MN55 data is: /tɑ/, /pɑ/, /kɑ/, /fɑ/, /θɑ/, /sɑ/, /ʃɑ/, /bɑ/, /vɑ/, /ðɑ/, /dɑ/, /gɑ/, /zɑ/, /ʒɑ/, /mɑ/, /nɑ/. The differences are the /tɑ/ and /pɑ/ are switched, and the group /dɑ/ and /gɑ/ is switched with /vɑ/ and /ðɑ/. The major distinctive feature groups (voicing and nasality) are still the same.

The ordering based on MN05 is: /ʃɑ/, /ʒɑ/, /zɑ/, /sɑ/, /dɑ/, /gɑ/, /ðɑ/, /vɑ/, /bɑ/, /θɑ/, /fɑ/, /pɑ/, /tɑ/, /kɑ/, /mɑ/, /nɑ/. The plot of this data at -6 dB SNR is seen in (b) in Fig. 2. The 5 major groupings are labeled on the plot. There are a few trends which are similar

to MN55. We see the nasal group (/mɑ/ and /nɑ/) and the /dɑ/ and /gɑ/ group in both data sets. We also see the unvoiced plosives (/pɑ/, /tɑ/, and /kɑ/) in both orderings. The different groups are the /sɑ/, /zɑ/, /ʃɑ/, and /ʒɑ/ group (denoted duration group) and the /bɑ/, /fɑ/, /θɑ/, /vɑ/, and /ðɑ/ group (denoted /bɑ/group). Both of these groups contain members from both voiced and unvoiced sets of CVs. This implies a lot of the errors for the CVs in these groups are voicing. This is in contrast to the results from MN55.

The metric is now applied to all orderings of the confusion matrices and the analysis in reported in Tab. 1. The values are reported for each SNR separately.

	MN55 data set			MN05 data set		
SNR	Original ordering	New MN55 Order	New MN05 Order	Original ordering	New MN55 Order	New MN05 Order
12	2.55	2.03	2.15	7.86	6.30	3.88
6	4.43	3.76	4.54	10.61	8.71	5.63
0	8.13	6.77	9.26	14.91	12.75	9.25
-6	19.76	17.22	22.74	25.86	23.51	18.54
-12	36.14	34.47	44.64	47.66	46.05	41.24
-18	78.76	78.87	80.17	80.69	81.05	79.12

Table 1: This table displays the metric values (Eq. 1) for the original ordering and both new orderings applied to both the MN55 and the MN05 data sets. The max value the metric can have is 85.

There are three important results from this table. First, we see that the new orderings do decrease the metric for their respective data sets. Second, the new MN55 ordering decreases the metric for the MN05 data as well. This is due to the switching of /dɑ/ and /gɑ/ with /vɑ/ and /ðɑ/. The /bɑ/ group in MN05 contains /bɑ/, /vɑ/, and /ðɑ/ so the rearrangement in the ordering will bring these closer together. Third, the MN05 order is only slightly worse than the original ordering for the MN55 data.

The largest difference in metric values is at the highest SNRs. This is because the 'primary groupings' are the CVs which are confused at the highest SNRs. Since the order will group these CVs together the metric will decrease significantly at the highest SNRs.

The new ordering from the MN05 data is best for the MN05 data. This ordering rejects the structure in MN55 due to distinctive features. The new ordering based on MN55 also rejects the ordering based on distinctive features. From these new orderings it is apparent that the distinctive features provided in MN55 do not explain the confusion patterns either MN55 or MN05.

2.1. Similarities between data sets

There are 3 groups showing similarities between the two data sets. These groups are (/pɑ/, /tɑ/, and /kɑ/), (/mɑ/, and /nɑ/), and (/dɑ/, and /gɑ/) labeled as 2, 4, and 5 in fig. 2b. Each of these groups are primarily only confused with the other members of their group at high SNRs in both experiments.

The first grouping is the unvoiced plosives (/pɑ/, /kɑ/, and /tɑ/). This group is only primarily confused with its own members down to low SNRs in MN05.

The second grouping seen in both orderings is the /ma/ and /na/ group. Both of the CVs in this group are voiced-nasals. This group was found to be very robust to wideband masking noise in MN55. This is supported by the data from MN05.

The third grouping is the /dɑ/ and /gɑ/ group. From fig. 2b it is apparent that at -6 dB SNR there are very few other competitors in MN05. In MN55 both /dɑ/ and /gɑ/ are confused with /zɑ/ and /ʒɑ/ at over 10% for -6 dB SNR. However, the major confusions were between /dɑ/ and /gɑ/.

2.2. Differences between data sets

There are a few very significant differences in the data sets. There are two groups in MN05 that are not seen in MN55. The first grouping in the MN05 data that shows a slight variation from the MN55 data is the /fɑ/, /bɑ/, /vɑ/, /θɑ/, and /ðɑ/ group. This grouping is actually seen slightly in the reordered MN55 data. In that data set /bɑ/, /ðɑ/ and /vɑ/ are confused. However in the MN05 data the confusions are much stronger. This group contains two sets of CVs where the differences inside the sets are a voicing feature (/vɑ/ & /fɑ/ and /θɑ/ & /ðɑ/). These two sets are highly internally confused at high SNRs. We notice that these sets inside the larger group also have the same distinctive feature for place.

The second group where the confusion patterns are different is the /zɑ/, /ʒɑ/, /ʃɑ/, /sɑ/ group. This group has two smaller sets inside the main grouping. These sets are /sɑ/ & /zɑ/ and /ʃɑ/ & /ʒɑ/. Both of these sets have members with different voicing features with all the other distinctive features the same. These confusions are not seen in the data from MN55. These confusions are so prolific in the data that they dominate the total errors at the highest SNRs.

3. Mutual Information

The differences between the two data sets are analyzed using information theory. This consists of using mutual information and relative information transmitted.

Mutual information measures the relationship between the input (CV spoken) and the output (CV responded). If there are no errors then the input maps directly to the output and the information transmitted is maximum. If there are errors the mutual information is lower. If the information transmitted is 0 then there is no correlation between the input and the output.

Mutual information is defined as follows:

$$T(x;y) = -\sum_{i,j} p_{j|i} \log_2 \left(\frac{p_i p_j}{p_{j|i}} \right) \tag{2}$$

where $T(x;y)$ is a measure of the information transferred from the input to the output in bits per a stimulus. The relative information transmitted ($T_{\text{rel}}(x;y)$) is defined as the portion of the possible information transferred. Thus, the definition of relative information transmitted is:

$$T_{\text{rel}}(x;y) = \frac{T(x;y)}{H(x)} \tag{3}$$

where $H(x)$ is the maximum entropy for the input. Table 2 shows the relative information transmitted ($T_{\text{rel}}(x;y)$) for each feature in both data sets. The $H(x)$ is also placed on the table for each column.

The most obvious result from tab. 2 is that the nasal, frication, and duration features are similar in both experiments and that place and voicing are significantly different. The information transmitted is used to analyze the robustness of voicing and place. These two distinctive features were chosen because of the strong conclusions in MN55. Using the information in this table the graph in fig. 3 is created. This figure shows the relative information transmitted for voicing and place for both experiments.

Figure 3 shows the difference in relative information transmitted for the voicing feature between MN55 and MN05. In the original MN55 analysis the voicing feature was found to be the most robust in noise. However in the MN05 data the voicing feature does not have the same relative information transmitted. In fact,

	MN55					MN05				
SNR	Voice	Nasal	Fricat	Durat	Place	Voice	Nasal	Fricat	Durat	Place
-18	.021	.015	.000	.001	.001	.012	.033	.011	.016	.012
-12	.522	.485	.069	.107	.037	.177	.351	.128	.240	.130
-6	.806	.730	.279	.307	.161	.372	.750	.328	.564	.328
0	.955	.910	.620	.596	.373	.528	.906	.527	.774	.643
6	.962	.998	.782	.784	.552	.596	.965	.680	.866	.820
12	.966	.999	.853	.926	.703	.667	.980	.823	.924	.884
$H(x)$	.989	.544	1.00	.811	1.55	.989	.544	1.00	.811	1.55

Table 2: This table shows the relative information transmitted between the input and the output for each distinctive feature. The entropy ($H(x)$) is also placed on the table for each column. If the value is 0 then there was no information transmitted. Conversely if the value is 1 all the information for that feature is correctly transmitted (no errors).

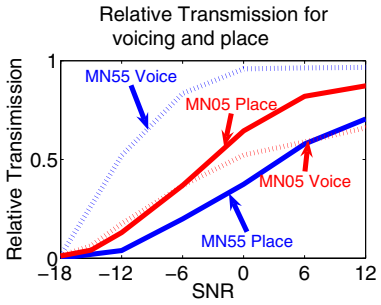

Figure 3: This is a plot of the relative information transmitted for voicing and place in both experiments. The voicing feature is transmitted better for MN55 and the place feature is transmitted better for MN05.

the relative information transmitted is less than the place distinctive feature. In the MN05 data the place information is transmitted significantly better than in MN55. The place feature is transmitted better than the voicing feature for the MN05 data. This is mostly due to the errors in the /zɑ/, /ʒɑ/, /ʃɑ/, /sɑ/ group, since this group's errors are more likely voicing errors than place (i.e. /sɑ/ is confused with /zɑ/ but not /ʒɑ/) at high SNRs. Additionally most of the errors in the /bɑ/group are voicing errors at the highest SNRs. The errors highlighted in fig. 1b are the voicing errors. Voicing errors are very common in MN05.

4. Discussion

The confusion patterns and primary competitors are different for both experiments. This is significant because the MN05 data shows errors in distinctive features that are extremely robust in MN55. In MN55 it is unknown (unstated in the paper) how the subjects were trained. It is known that they were trained extensively. In that protocol the subjects were required to answer something even if they heard only noise. This forced the subjects to decide somehow how to respond. It is possible, since the subjects were highly-trained, the subjects employed a specialized guessing procedure when they heard only noise. Some of the differences in the data might be attributed to this approach to random guessing. In contrast the subjects in MN05 were not highly trained and they were offered the 'noise only' button.

Another difference that may have biased the data is that the subjects and talkers knew each other in MN55. In MN55 all the subjects and talkers were females from the Boston area. This familiarity and daily interaction between the talkers and listeners may have biased the subject responses. In MN05 the subjects did not know the talkers. There was also a larger subject pool so the per subject biases should not affect the data as much in MN05. The subject and talker pool is heterogeneous in the MN05 data.

5. Conclusions

MN55 and MN05 provide different results, both in error patterns and mutual information. The CVs /bɑ/, /vɑ/, /fɑ/, /θɑ/, /ðɑ/, /sɑ/, /zɑ/, /ʒɑ/, /ʃɑ/ have difference confusion groups between the two experiments. These differences are due to various factors. The presentations were prerecorded in MN05 as opposed to live in MN55. A second factor is the composition of the talkers and subjects. The subjects were highly trained in MN55 and knew their talkers. The subjects in MN05 were heterogeneous in that they came from all backgrounds and accents. Additionally the corpus contained a heterogeneous selection of talkers.

The CVs specific distinctive feature errors are different between the two experiments. In both experiments nasality, duration, and frication have approximately the same relative information transmitted. In MN55 the distinctive feature errors were systematic as the SNR was decreased. For instance, the place distinctive feature is in error at higher SNRs than voicing for all CVs. However in the MN05 data there are no such global trends. Some CVs (/zɑ/, /sɑ/, /ʒɑ/, and /ʃɑ/) had all the distinctive features properly recognized at high SNRs except voicing. In contrast other CVs (/dɑ/, and /gɑ/) correctly transmitted all distinctive features except for place at the highest SNRs. The distinctive feature errors are confusion group specific in MN05. Additionally only certain groups contained errors at the highest SNRs in MN05. These groups are the /bɑ/group and the /zɑ/, /sɑ/, /ʃɑ/, and /ʒɑ/ group. These errors were predominantly voicing errors, so this explains the presence of voicing errors at the highest SNRs. In both data sets /bɑ/ is more likely to be confused with /vɑ/ and /θɑ/ than /dɑ/ and /gɑ/. Additionally, the errors are more likely to be voicing than place at high SNRs in the new experiment.

6. References

[1] George A. Miller and Patricia E. Nicely, "An analysis of perceptual confusion amoung English consonants," *Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.

[2] Sigfrid D. Soli and Phipps Arabie, "Auditory versus phonetic accounts of observed confusions between consonant phonemes," *Journal of the Acoustical Society of America*, vol. 66, pp. 46–59, 1979.

[3] R. Shepard, "Psychological representation of speech sounds," in *Human Communication: A Unified View*, E. David and P. Denies, Eds., chapter 4, pp. 67–113. McGraw-Hill, New York, 1972.

[4] Jont B. Allen, "Consonant recognition and the articulation index," *Journal of the Acoustical Society of America*, vol. 117, pp. 2212–2223, April 2005.