50 Years Late: Repeating Miller-Nicely 1995

Andrew Lovitt, and Jont B. Allen

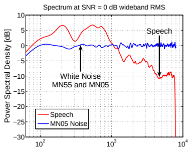
University of Illinois at Urbana-Champaign

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 (CM) 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.

Similar Procedures

- Subjects reported which CV they heard when the CVs were spoken in white masking noise
- CVs: p, t, k, f, θ, s, ʃ, b, d, g, v, ð, z, ʒ, m, and n with ɑ
- SNR = $\frac{\text{VU reading of Speech}}{\text{RMS of Noise}}$
- Noise LPF to 7000, Speech BPF 200-6500 Hz
- All listeners spoke English as a first language



Differences in Procedures

	MN55	MN05
SNRs (dB)	12, 6, 0, -6, -12, -18	Quiet, 12, 6, 0, -6, -12, -15, -18
Tokens	Live	Prerecorded [a]
Talkers	5 (same as listeners)	18
Listeners	5 (same as talkers)	23
Gender of subjects	Female	Male and Female
Repeats	No Repetition	Allowed to Repeat
VU Meter	Hardware	Software [b]
Administered by	Personal	Matlab©(supervised)
Training	Trained Heavily	1 hour

- Additional response for MN05: *Noise Only* response which was pressed by the subject if they didn't hear any speech. These responses were distributed uniformly over all possible responses

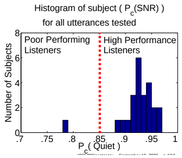
[a] LDC Corpus #LDC2005S2

[b] (submitted) VU-soft, Lobdell and Allen, 2006

Listener Selection

MN55 used highly trained subjects. The subjects are assumed to operate at high P_c in quiet, and the talkers are assumed to properly pronounce the CVs. Thus the *utterances* (a talker speaking a CV) and listeners are analyzed and poor performance subjects are pruned. (Procedures follow Phatak 06 [5])

- Kept 23/24 listeners who:
  - Completed the Experiment
  - $P_c(\text{Quiet}) > .85$

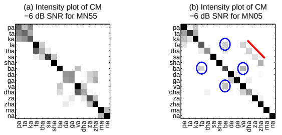


Utterance Selection

- Mislabeled: Utterances where the listeners report a CV which is not what the utterance is labeled as
  - Utterance are relabeled so $P_c(\text{Quiet})$ is Maximum
- After relabeling, mispronounced utterances are where the $P_c(\text{Quiet}) < .8$ (Listeners do not agree in Quiet)
  - Mispronounced: removed from analysis
  - 10/12 mislabeled are also mispronounced

s	Mislabeled $s*$	Number of Mispronounced	Number of presentations after pruning
/ðɑ/	/bɑ/, /vɑ/, /θɑ/, /θɑ/	11	6
/θɑ/	/fɑ/, /ðɑ/	8	9
/zɑ/	/ðɑ/, /sɑ/	1	15
/vɑ/	/fɑ/	3	15
/fɑ/	/θɑ/	3	16
/ʒɑ/	/ʃɑ/	2	16
/bɑ/		2	16
/sɑ/		1	18
others			18

Confusion Differences



- The confusion matrices are different between each experiment specifically:
  - /fɑ/, /θɑ/, /sɑ/, and /ʃɑ/ show confusions with their voiced counterpart in MN05 (red line)
  - /fɑ/, /bɑ/, and /vɑ/ are confused with each other in MN05 (blue circles), these confusions are with both voiced and unvoiced consonants

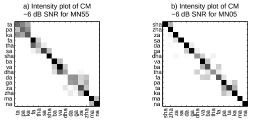
Reordering of the Confusion Matrix

- **Is there an ordering that has the CVs ordered next to their confusions?**
- The weighting metric used is the taxi-cab (Manhattan) distance weighted by the $P_{j|i}(SNR)$ (eq. 1)

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

- All orderings are pruned of obviously poor orderings; then remaining orderings are analyzed across all SNRs

Reordered Confusion Matrices



- Both orderings contain the following CVs together
  - /pɑ/, /tɑ/, and /kɑ/ (Unvoiced Plosive)
  - /mɑ/, and /nɑ/ (Nasal)
  - /dɑ/, and /gɑ/
  - /bɑ/, /vɑ/, and /ðɑ/
- These groups have the same major confusions in both experiments and thus were grouped together in both experiments

Differences in Orders

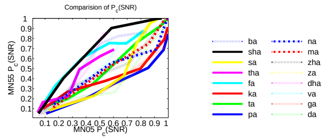
- Differences in new MN55 order
  - Reorders /vɑ/ and /ðɑ/ with /dɑ/ and /gɑ/ because /bɑ/ is more confused with /vɑ/ and /ðɑ/, and also /dɑ/ and /gɑ/ are more confused with /zɑ/ and /ʒɑ/
  - Reorders /pɑ/ with /tɑ/ because /kɑ/ is more confused with /pɑ/ than /tɑ/
- Groupings in MN05 not found in MN55
  - /sɑ/ and /zɑ/: Represent a confusion across voicing
  - /ʃɑ/ and /ʒɑ/: Represent a confusion across voicing
  - /bɑ/, /fɑ/, and /vɑ/: Same place, but frication and voicing are different
- Thus MN05 shows confusion not predicted from the MN55 results that voicing is most robust and place is less robust.

Distinctive Features

CV	Voicing	Nasality	Frication	Duration	Place
/pɑ/	0	0	0	0	0
/tɑ/	0	0	0	0	1
/kɑ/	0	0	0	0	2
/fɑ/	0	0	1	0	0
/θɑ/	0	0	1	0	1
/ʃɑ/	0	0	1	1	1
/sɑ/	0	0	1	1	2
/bɑ/	1	0	0	0	0
/dɑ/	1	0	0	0	1
/gɑ/	1	0	0	0	2
/vɑ/	1	0	1	0	0
/ðɑ/	1	0	1	0	1
/ʒɑ/	1	0	1	1	1
/zɑ/	1	0	1	1	2
/mɑ/	1	1	0	0	0
/nɑ/	1	1	0	0	1

- The distinctive features of MN55 are voicing, nasality, frication, duration, and place.
- MN55's results indicated a structure in the confusion patterns that was similar to the structure of the distinctive features.
- The information theoretic analysis of MN55 was repeated on MN05

P_c (MN55) vs. P_c (MN05)



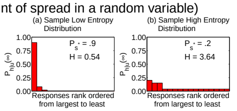
- CVs: P_c (MN05) $\geq P_c$ (MN55), same error patterns as MN55. However the P_e for these CVs are lower in MN05. (/mɑ/, /nɑ/, /tɑ/, /pɑ/, /kɑ/, /ʒɑ/, /dɑ/, and /gɑ/)
- CVs: P_c (MN55) $> P_c$ (MN05), different error patterns and these errors tend to be voicing errors and not place errors. (/ʃɑ/, /bɑ/, /θɑ/, /ðɑ/, /fɑ/, /vɑ/, (and /zɑ/, /sɑ/ at high SNRs))

Entropy and Mutual Information

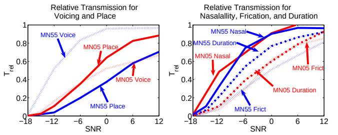
- *Mutual information* measures the amount of information sent through a channel
- Relative Information Transmitted ($T_{\text{rel}}(i;j)$) is defined as follows: (P_{ij} represents the probability that i was sent and that j was received.)

$$T_{\text{rel}}(x;y) = \frac{-\sum_{i,j} p_{ij}\log_2(\frac{p_i p_j}{p_{ij}})}{H(i)} \quad (2)$$

- $H(i)$ is the entropy of the input (entropy represents the amount of spread in a random variable)



Comparison of the Mutual Information



- Nasality, Frication, and Duration all have approximately the same amount of information transmitted in both experiments even with the large differences between the experiments
- Voicing and Place have significantly different amounts of relative information transmitted
  - Place and Voicing reverse their importance

Discussion

- The differences in information transmitted and confusion patterns may be due to:
  - Subject differences:
    - Talker differences: MN55 used subjects (listeners and talkers) who knew each other, this was not true in MN05
    - Listener differences: MN55 were highly trained, MN05 had no training.
    - MN55 used only female talkers and listeners, MN05 used talkers from diverse linguistic backgrounds and listeners born all over the USA
  - Corpus differences:
    - MN55 was live whereas MN05 used recordings
    - MN55 tokens had a (assumed) speech-shaped spectra on average, MN05 did not

Conclusions

- The information transmitted for place is higher than voicing in MN05
- In MN55 place was the least robust and voicing was the 2nd most robust to masking noise
- In MN05 this is reversed and voicing is the least robust to noise and place is the 2nd most robust to masking noise
- These differences are due to a fundamental difference in the talkers and listeners
- The listener and utterance variability was analyzed for MN05 (not shown here) and was found not be random, but having a structure to it. Such analysis is not possible with MN55 data. This leads to further work on the structure of the confusion patterns.

References

[1] George A. Miller and Patricia E. Nicely, "An analysis of perceptual confusion among 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.

[5] S. A. Phatak and J. B. Allen„ "Consonant and vowel confusion patterns in speech-weighted noise," *Journal of the Acoustical Society of America*, **SUBMITTED**.