

José R. Benki^{a)} and Robert Felty^{b)} (2005) Recognition of English phonemes in noise, *J. Acoust. Soc. Am.* **117**, 2568.

^{a)}Department of Otolaryngology, University of Michigan Medical School; benki@umich.edu

^{b)}Departments of German and Linguistics, University of Michigan; robfelty@umich.edu

Introduction

The purpose of the present study is to develop a set of English phoneme confusion matrices in order to assess phonetic neighborhood densities for spoken English words presented in noise. The Shannon et al. (1999) CV and VCV stimuli and Hillenbrand et al. (1995) hVd stimuli sample nearly all English phonemes in a fixed phonological context, and are thus ideal for our application. This poster presents the results from the fixed-response identification of those stimuli as recorded by Shannon et al. and Hillenbrand et al. (S/N ratio=max) as well as embedded in three levels of noise (0 dB, -3 dB, -6 dB).

These widely used multitalker stimulus sets required some modifications for the purpose of the present study. Previous research in both fixed and open response formats indicates differences between utterance-initial and final position for consonants (Redford and Diehl 1999; Benki 2003a,b; Cutler et al. 2004), so a VC condition was created by removing the final vowels the VCV stimuli. Secondly, in addition to consonant substitution errors, deletion (and some insertion) errors are reported in the literature, particularly the deletion of sonorants in final position (Benki 2003b). Accordingly, vowel-only stimuli (V) were created from selected VCV stimuli and presented to listeners in both consonant conditions in order to measure deletion error rates. Finally, the Hillenbrand et al. hVd stimuli do not include the diphthongs /aI aU cI/, for which we recorded 10 talkers to complete the stimulus set.

Procedure

Stimuli mixed with signal correlated noise (see Schroeder 1967 for full description - see Benki 2003 for evidence showing that perceptual effects are similar to broadband white noise) were presented to listeners at four different S/N ratios - -6, -3, 0 and max (i.e. containing only noise that was in the original recording). Stimuli were presented over AKG headphones connected to an iMic D/A USB device on laptops running Windows XP in an anechoic chamber. Results were collected with a forced-choice design using a slightly modified English orthography, implemented in the MATLAB programming environment.

Stimuli

Stimuli consisted of the entire set of possible English consonants and vowels - 23 initial consonants, 15 vowels and diphthongs, and 21 final consonants, (plus a no-consonant condition for both initial and final position). Stimuli were taken from two published sources. The consonant stimuli taken from Shannon et al. 1999, which consisted of CV and VCV stimuli, with all possible consonants, and the vowels [a], [i], and [u]; the hVd stimuli were those recorded by Hillenbrand et al. 1995, which consisted of the 12 English vowels in the environment hVd. Since we were interested in final, not medial, consonants, the VC portions of the VCV stimuli were extracted to construct the VC stimuli. For obstruents, the final vowel was cut off at the onset of glottal pulsing in the waveform; stop bursts were preserved. For sonorants, final vowels were cut off at the beginning of F1 transition, and the consonant amplitude was ramped down slightly. We used all of the talkers from Shannon et al. - 5 female and 5 male, and a random subset of 5 female and 5 male talkers from the Hillenbrand et al. data. This resulted in a stimulus list of 1530 tokens (((24 Initial C + 22 Final C)*3 Vowels + 15 Vowels)*10 Talkers). In order to keep the experiment under one hour, the list was split into 5 random sublists, such that each list contained 2 instances of each phonologically unique syllable drawn roughly equally from the complete set of tokens from each talker. All editing was done with Praat; all stimuli were normalized by peak amplitude.

Pilot testing revealed that final /r/ and /ŋ/sounded very unnatural, and therefore we recorded these stimuli in our lab. As did Shannon et al., we recorded 5 male and 5 female talkers, saying the syllables /ar/, /ir/, and /ur/, and /aŋa/, /iŋi/, and /uŋu/10 times, and selected the best token from each talker. We also recorded the diphthong stimuli /hard/, /haud/, and /hcid/, which Hillenbrand et al. did not.

Listeners

Native speakers of English were recruited via flier at the University of Michigan, and consisted of both undergraduate and graduate students, with no reported speech pathologies. Some listeners had some background in linguistics but none were experienced in identifying speech in noise. A total of 28 listeners took part in the experiment. Since there were 5 different lists, listeners were allowed to participate up to 5 times. The present data are drawn from 40 tests, 10 at each S/N ratio. Later analyses will include investigating a possible learning effect.

Results and Discussion

The detailed results are presented in the **confusion matrices (Table 1 on separate sheet)**. In general, place of articulation errors predominate among the consonants, while among vowels, tense/lax errors predominate among the vowels. Vowel performance /i a u/ in the CV and VC identification tasks was near-ceiling for all S/N ratios.

The results are summarized in **Figure 1 by percent correct vs. S/N ratio**. Performance for all phonemes rose with S/N ratio. Vowels and initial consonants were identified better than final consonants, consistent with Redford and Diehl (1999) and Benki (2003), but opposite the results reported by Cutler et al. (2004). Overall performance for initial consonants for SNR=max (85%) is lower than reported for identification of initial and medial consonants by Shannon et al. (97%). In that study, the listeners were highly practiced (each performed about 10 times as many trials as the listeners in the present study) and were not presented with the /θ h Ø ŋ ʒ/ phonemes.

For the vowel hVd stimuli, talker intelligibility ranged from 63 to 79% correct, averaged over S/N ratio and listener. For the consonants, Talker 7 was significantly less intelligible than the rest of the talkers as shown in the **consonant talker analysis in Figure 2**. This talker was also the least intelligible talker (Female 2) as reported by Shannon et al.

Following Miller and Nicely (1955), the results by phonological feature are summarized by the **information analysis (Figure 3)**, for consonant voicing, place of articulation, and manner contrasts; and vowel height, front/back, and tense/lax contrasts, using the same feature scheme as Cutler et al. (2004). The right axis of each panel is the approximate number of bits transmitted per feature, while the left axis is the percent of the maximum for each feature. For consonants, place of articulation is the most vulnerable to noise, although the absolute number of bits transmitted by place is greater than voicing for all noise levels. For vowels, the tense/lax contrast is most vulnerable to noise.

While the information analysis and percent correct plots (Figures 1 & 2) offer a useful summary of the perceptibility of the stimuli by phoneme, feature, and context, a number of important details are only available in the confusion matrices, in part because trials with null consonants (/Ø/) or diphthongs (/aI aU aI/) as responses or stimuli were omitted from the information analysis. Some of these details include:

- Rates of consonant deletion (highest among final nasals) and insertion
- Confusions among diphthongs
- Interactions between features, such as the unusual vulnerability of manner of articulation in bilabial consonants

Acknowledgements

Special thanks to Alicia Harris, Christina Li, Ariel Moses, and Gloria Redondo for preparing stimuli and running perceptual experiments. This work was supported by grant R03 DC05913 from the NIDCD and the University of Michigan Undergraduate Research Opportunity Program.

References

- Benki, J.R. (2003a) Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition. *JASA* **113** (3), 1689-1705.
- Benki, J.R. (2003b) Analysis of English Nonsense Syllable Recognition in Noise. *Phonetica* **60**, 129-157.
- Cutler, A., Weber, A., Smits, R., and Cooper, N. (2004) Patterns of English phoneme confusions by native and non-native listeners. *JASA* **116** (6), 3668-3678.
- Hillenbrand, J.M., Getty, L.A., Clark, M.J., and Wheeler, K. (1995) Acoustic characteristics of American English vowels. *JASA* **97** (5), 3099-3111.
- Miller, G. and Nicely, P.E. (1955) An analysis of perceptual confusions among some English consonants. *JASA* **27**, 338-352.
- Redford, M., and Diehl, R. (1999) The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *JASA* **106** (3), 1555-1565.
- Schroeder, M.R. (1968) Reference signal for signal quality studies. *JASA* **44**, 1735-1736.
- Shannon, R.V., Jensvold, A., Padilla, M., Robert, M.E., and Wang, X. (1999) Consonant recordings for speech testing. *JASA* **106** (6), L71-L74.

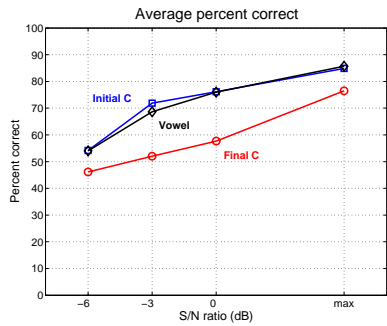


Figure 1. Percent correct at each S/N ratio for CV (initial C), hVd (vowel), and VC (final C) stimuli

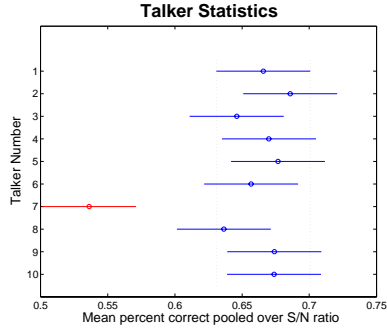


Figure 2. Mean p(c) by talker for consonant data. Error bars represent 95% confidence intervals

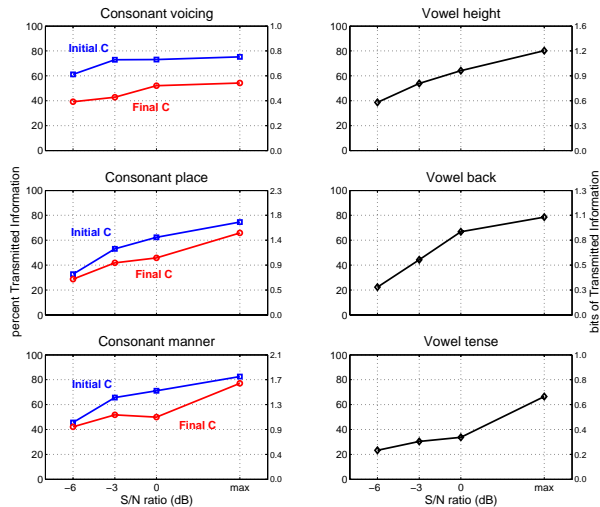


Figure 3. TI featural analysis, pooled across S/N ratio, listener, and talker. The right axis of each panel is the approximate number of bits transmitted per feature, while the left axis is the percent of the maximum for each feature. (cf. Miller & Nicely 1955)

Results and Discussion (cont.)

The proportion correct score in the maximum S/N ratio for each vowel, pooled across listeners, is plotted against the comparable score reported by Hillenbrand et al. in the **vowel analysis (Figure 4)**. The correlation is high even with the outlier phoneme /a/ (with /a/, $r=0.640$, $p<0.025$; without /a/, $r=0.716$, $p<0.013$), which was identified much better by the trained listeners in the Hillenbrand et al. study. Unlike the phonetically trained listeners used by Hillenbrand et al., the naive listeners in the present study identified 20% of the /a/ stimuli as /ɔ/ (consistent with an /a ɔ/ merger) and 35% as /æ/ (consistent with the Northern Cities Shift).

Consonant percent correct by vowel context is presented in Figure 5, and presented in terms of features in the **information analysis in Figure 6**. In general, performance in the SNR=max condition did not vary by vowel context. However, for the noise conditions, both initial and final consonants in the /a/ context were easier to identify than the /u/ and /i/ contexts, with mean differences increasing from 10 to 19 percentage points. The information analysis in Figure 6 indicates that variation by vowel context was primarily due to manner and place of articulation errors, with no systematic vowel effects for the voicing feature. Currently we are systematically measuring duration and formant values in search of an explanation for this finding.

Feature	Values	Phonemes
Consonant manner	stop	/p t k b d g/
	affricate	/tʃ dʒ/
	fricative	/f v θ ð s ʃ h ʊ z ʒ/
	liquid	/l r/
	glide	/j w/
	nasal	/m n ŋ/
		/p f b v m w/
Consonant place	labial	/p f b v m w/
	dental	/θ ð/
	alveolar	/t s d z n l/
	palatal	/ʃ tʃ ʒ dʒ r/
	velar	/k g ŋ/
Consonant voicing	glottal	/h/
	voiced voiceless	/b d g v d z ʒ dʒ j m n ŋ l r w/
Vowel height	voiced voiceless	/p t k f θ s ʃ tʃ h/
	high	/i I U u/
	mid	/e I E v o U ə/
	low	/æ a ɔ A/
Vowel backness	front	/i I e I ε ae/
	central	/ə/
	back	/a ɔ A o U u/
Vowel tenseness	tense	/i e I c o U u ə/
	lax	/I E æ a A U/

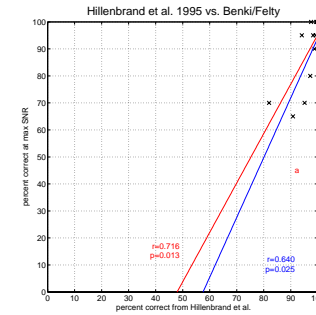


Figure 4. Comparison of hVd results from Hillenbrand et al. 1995 and the current study, using the results at the max S/N ratio.

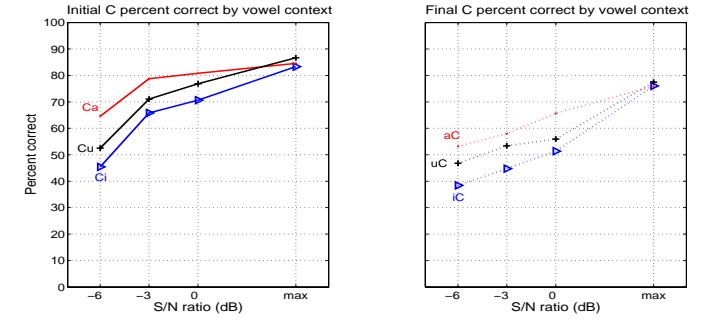


Figure 5. Percent correct according to vowel context, for initial and final consonants, for each S/N ratio.

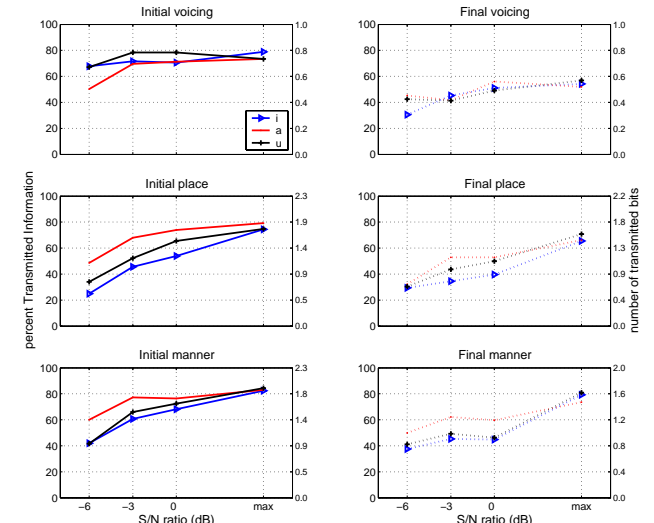


Figure 6. TI featural analysis for each vowel context, pooled across talkers and listeners.