Phoneme categorization relying solely on frequencies beyond ~6 kHz

Brian B. Monson
Department of Speech and Hearing Science, College of Applied Health Sciences, University of Illinois at Urbana-Champaign

Introduction

Over phylogeny the human cochlea and brain have obtained and retained sensitivity to acoustical energy at frequencies spanning 20 Hz to 20 kHz.

Why retain sensitivity to the very high frequencies (up to 20 kHz)? We hypothesize that human sensitivity to these very high frequencies has been retained, in part, because the information provided by acoustical energy at the high frequencies is valuable for the perception of conspecific vocalizations (*i.e.*, human speech).

Some studies indicate that very high frequencies can aid in improving speech reception thresholds when target and masking speech are spatially separated. We questioned whether energy at the high frequencies *per se* could provide phonetic information.

Aim

To determine whether very high frequencies provide cues for both consonant and vowel identification.

Method

Stimuli

- CV combinations consisting of consonants /p, b, t, d, k, g, f, v, m, n, s, z, ʃ/ paired with each of the vowels /i, æ, ɑ, o, u/
- One male and one female talker
- Recorded at 16-bit, 44.1 kHz sampling rate
- 130 CV tokens (13 consonants x 5 vowels)
- Stimuli bandpass filtered with cutoff frequencies 5.7 and 20 kHz (to include 8- and 16-kHz octaves)
- Speech-shaped noise masker, low-pass filtered at 5.7 kHz

High frequency energy characteristics

High frequency energy (HFE) in speech can provide the temporal modulation information of the speech signal (Figure 1).

The long-term average spectra of voiced vs. unvoiced speech reveal HFE is dominated by unvoiced phonemes (Figure 2).

Some consonants (*e.g.*, unvoiced fricatives) exhibit distinguishing acoustic features at the high frequencies (Figure 3).

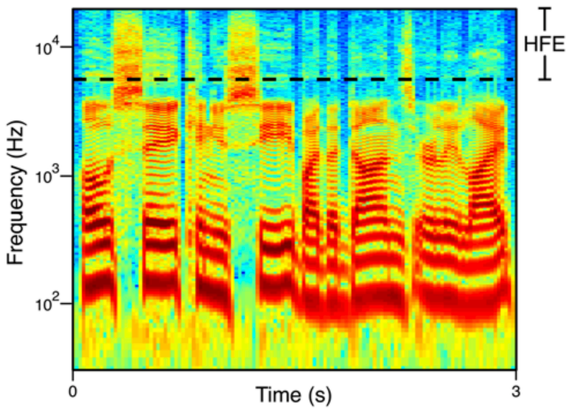

***Figure 1**: Broadband spectrogram of "Oh say can you see by the dawn's early light" uttered by a male talker. The dotted line is 5.7 kHz.*

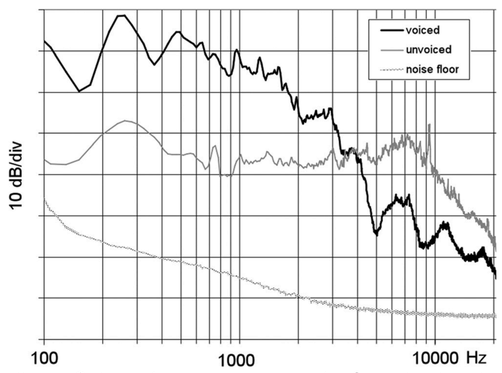

***Figure 2**: Long-term average spectra for voiced and unvoiced contributions of 60 sec of running speech produced by a female talker.*

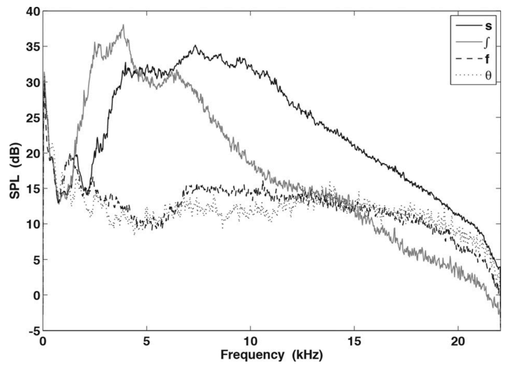

***Figure 3**: Long-term average spectra for unvoiced fricatives (mean spectrum from 15 female and male talkers).*

Method (continued)

- HFE amplitude set to 47 dB SPL
- Masker amplitude set to 62 dB SPL

Subjects

- 13 normal-hearing native English-speaking adults

Procedure

- Closed-set phoneme categorization task
- Separate blocks for consonant and vowel categorization
- Consonant response options /**p, b, t, d, k, g, f, v, m, n, s, z, ʃ**, ʒ, θ, ð, tʃ, dʒ, r, l, w, j, h/
- Vowel options /**i, æ, ɑ, o, u**, ɪ, e, ɛ, ɔ, ʊ, ʌ/
- Repeated stimulus presentation was allowed

Results

Mean scores (percent correct)

Consonant: 51.5% (chance = 4.3%, $p<0.001$)

Vowel: 15.8% (chance = 9.1%, $p<0.005$)

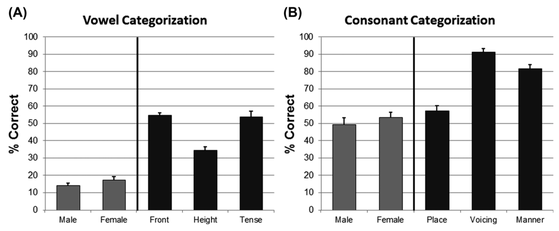

***Figure 4**: Vowel (A) and consonant (B) categorization results, separated by talker sex and articulation characteristics.*

Consonant confusion matrix

Response	Stimulus												
Phoneme	b	p	d	t	g	k	v	f	z	s	ʃ	m	n
b	**46.2**	2.3	1.5	0	3.1	0	4.6	6.9	0	0	0	15.4	6.9
p	2.3	**47.7**	0.8	0.8	0.8	4.6	0	1.5	0	0	0	0.8	1.5
d	16.9	3.1	**76.2**	4.6	61.5	1.5	25.4	0	0	0	0	0.8	4.6
t	0.8	6.9	7.7	**80.8**	1.5	35.4	3.8	0	0	0	0	0.8	0.8
g	9.2	3.8	9.2	0	**16.2**	1.5	12.3	0	0	0	0	1.5	6.2
k	0.8	17.7	0.8	0.8	2.3	**36.2**	0	0	0	0	0	1.5	0
v	0.8	0.8	0	0.8	0.8	0	**26.2**	0.8	0	0	0	2.3	0
f	0.8	1.5	1.5	0.8	0	0.8	0.8	**73.1**	0	0.8	2.3	0	0.8
z	0.8	0	0	0	0.8	0	0	1.5	**89.2**	1.5	0.8	0.8	0
s	0.8	1.5	0	0.8	0	2.3	0.8	8.5	3.8	**89.2**	75.4	0	0.8
ʃ	0	0.8	0	0	0	1.5	0	0	0.8	6.2	**20.8**	0	0
m	1.5	0	0.8	0	0.8	0	1.5	0.8	0	0	0	**21.5**	23.8
n	9.2	0	1.5	0	0.8	0	7.7	0	0	0	0	45.4	**46.2**
ʒ	0.8	0	0	0	0.8	0.8	0.8	1.5	3.1	0.8	0.8	0	0.8
dʒ	1.5	3.8	0	0.8	6.9	0.8	0	0	1.5	0	0	0.8	0
tʃ	0	3.1	0	1.5	0	6.9	0	0	0.8	0	0	0	0.8
ð	0.8	0.8	0	5.4	0.8	0	3.8	0	0	0	0	1.5	0
θ	1.5	1.5	0	1.5	2.3	2.3	3.8	4.6	0.8	0.8	0	1.5	0.8
r	0	1.5	0	0.8	0.8	3.1	0.8	0	0	0	0	0	0
l	2.3	0	0	0	0	0	2.3	0	0	0	0	3.1	1.5
w	0	0	0	0	0	0	1.5	0.8	0	0.8	0	1.5	2.3
j	1.5	2.3	0	0.8	0	2.3	3.1	0	0	0	0	0.8	1.5
h	1.5	0.8	0	0	0	0	0.8	0	0	0	0	0	0.8

***Table 1**: Consonant confusion matrix showing distribution (percentage) of responses for each phoneme.*

Results (continued)

Vowel confusion matrix

Response	Stimulus				
Phoneme	i	æ	ɑ	o	u
i	**31.4**	15.4	14.2	14.2	17.5
æ	5.6	**9.2**	10.4	5.9	5.3
ɑ	6.2	12.7	**14.8**	12.4	8.6
o	6.5	6.8	8.9	**8.6**	9.2
u	9.5	7.1	6.8	8.6	**15.1**
ɪ	7.1	8.9	3.3	7.7	8
e	5	5.3	6.5	7.7	5.9
ɛ	8	7.4	6.2	8.6	7.4
ɔ	7.1	11.2	11.5	7.4	5.6
ʊ	5.9	2.7	3.3	5.9	5.6
ʌ	7.7	13.3	14.2	13	11.8

***Table 2**: Vowel confusion matrix showing distribution of responses for each phoneme.*

Conclusions

In the absence of low-frequency acoustic cues that are the traditional focus of speech perception research (*e.g.*, the first four formants), acoustical energy beyond ~6 kHz provides phonetic information useful for both consonant and vowel identification within isolated CV tokens.

Consonant identification is much better than vowel identification, although vowel identification is still above chance.

The findings support the notion that human sensitivity to the high frequencies is beneficial for speech perception, and thus may have been retained over phylogeny for its perceptual utility within conspecific communication.

References

Vitela *et al.* (2015) *JASA* 137:EL65-EL70.
Monson *et al*. (2014) *Frontiers Psychol* 5:587.
Monson *et al.* (2012) *JASA* 132:1754-1764.
Monson *et al.* (2011) *JASA* 129:2263-2268.
Levy *et al.* (2015) *Ear Hearing* 36:e214.
Moore *et al.* (2010) *JASA* 128:360-371.

Acknowledgements

The author wishes to thank Davi Vitela, PhD, Andrew Lotto, PhD, Brad Story, PhD, Sten Ternström, PhD, and Eric Hunter, PhD for their contributions to this work. Funded in part by NIDCD Grant No. DC0105332.

ILLINOIS