Phoneme categorization relying solely on high-frequency energy

A. Davi Vitela[a)]
Department of Psychology, University of Nevada, Las Vegas, 4505 South Maryland Parkway, Las Vegas, Nevada 89154
davi.vitela@unlv.edu

Brian B. Monson
Department of Pediatric Newborn Medicine, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, Massachusetts 02115
bmonson@research.bwh.harvard.edu

Andrew J. Lotto
Speech, Language, and Hearing Sciences, University of Arizona, 1131 East Second Street, Tucson, Arizona 85721
alotto@email.arizona.edu

Abstract: Speech perception studies generally focus on the acoustic information present in the frequency regions below 6 kHz. Recent evidence suggests that there is perceptually relevant information in the higher frequencies, including information affecting speech intelligibility. This experiment examined whether listeners are able to accurately identify a subset of vowels and consonants in CV-context when only high-frequency (above 5 kHz) acoustic information is available (through high-pass filtering and masking of lower frequency energy). The findings reveal that listeners are capable of extracting information from these higher frequency regions to accurately identify certain consonants and vowels.

[AC]
Date Received: August 7, 2014 **Date Accepted:** November 30, 2014

1. Introduction

Despite the fact that the human voice produces substantial energy above 5 kHz (Moore *et al.*, 2008; Monson *et al.*, 2012b) and that listeners are sensitive to changes in these higher frequencies for human speech and song (Monson *et al.*, 2014b), much of the theory and empirical exploration in speech perception has focused on the acoustic characteristics of the speech signal below 5 kHz. This is, in part, due to the fact that normal hearing adults can achieve nearly perfect performance on speech categorization for signals limited in range to lower frequencies, as demonstrated by the communication viability of the band-limited signal on telephones. However, there has been renewed interest in high-frequency energy (HFE) (defined here as energy in the 8- and 16-kHz center frequency octave bands, or 5.7–22 kHz) and listeners' ability to extract speech information from it as hearing aids (Moore, 2012) and communication technology (Pulakka *et al.*, 2012) have attempted to better represent this frequency range (reviewed in Monson *et al.*, 2014a).

Despite early reports that very high frequencies do not change speech intelligibility scores (e.g., Pollack, 1948), more recent evidence suggests HFE can impact speech intelligibility. Lippmann (1996) presented listeners with nonsense consonant-vowel-consonant (CVC) tokens low-pass filtered at 800 Hz. With the mid-frequencies

[a)]Author to whom correspondence should be addressed.

still absent, listeners' identification of consonants improved by greater than 30 percentage points (from 44.3% correct to 76.9% correct) when HFE above 6.3 kHz was added. Apoux and Bacon (2004) showed a decrement in intelligibility performance during a speech-in-noise task when a frequency band from 3.5 to 10 kHz was removed from the signal. Similarly, Moore *et al.* (2010) found a small but significant increase in intelligibility by increasing the bandwidth of the target speech from 5 to 7.5 kHz for speech-in-noise tasks where the target speech and noise maskers were spatially separated. Further, there have been some reports of listeners' ability to recognize words and phrases using only HFE (Fullgrabe *et al.*, 2010), although this ability has not been rigorously tested.

These results motivated the experiment presented here that investigates whether there is salient information in the HFE that may affect speech intelligibility scores. To address this question, we tested listeners' ability to recognize a subset of vowels and consonants in a CV structure in tokens consisting of only HFE. Using both a male and female voice, we examined overall accuracy and then characterized the errors using vowel and consonant confusion matrices. If listeners are able to categorize these speech signals above chance, it stands to reason that there is perceptually usable phonetic information in HFE that may be relevant for speech communication.

2. Methods

2.1 Stimuli

Stimuli were recorded from both a male and female speaker in a soundproof booth, using an AKG SE 300 B microphone and the Computerized Speech Lab (model 4500) system from Kay Pentax (16-bit, 44.1 kHz sampling rate). The CV combinations consisted of the consonants /p, b, t, d, k, g, f, v, m, n, s, z, ʃ/ paired with each of the following vowels: /i, æ, ɑ, o, u/. This resulted in 130 CV stimuli (13 consonants × 5 vowels × 2 speakers). Consonants were constrained to include stops, fricatives and nasals in English because these are the most common manners of articulation resulting in straightforward calculation of information transfer. Consonants with low frequencies or those that are phonotactically inappropriate to occur at the start of a syllable in English were excluded. The vowels chosen define the vowel quadrilateral and provide maximal acoustic distinctiveness in English.

The stimuli were bandpass filtered with a digital Parks-McClellan equiripple finite impulse response filter with cutoff frequencies of 5.7 and 20 kHz using MATLAB (MathWorks, Natick, MA). A low-frequency masker was used, consisting of speech-shaped noise generated in MATLAB by filtering a white noise signal according to ANSI specifications (ANSI, 1992), and then low-pass filtering at 5.7 kHz with a 32-pole Butterworth filter. The purpose of the low-frequency masker was to ensure that listeners were using HFE, *per se*, and not distortion products that may be present in the lower frequencies. The HFE amplitude was set to 47 dB root-mean-square sound pressure level (SPL_{rms}) and the low-frequency masker was set to 62 dB SPL_{rms}, which are typical levels produced during normal speech for a speaker facing the listener (Monson *et al.*, 2012a, 2012b).

2.2 Participants

Thirteen undergraduates from the University of Arizona participated in this study for course credit. All reported normal hearing and English as their native language.

2.3 Procedure

Participants were run in groups of one to four at a time on separate computers in a soundproof booth. Stimulus presentation and data collection were controlled by the ALVIN program (Hillenbrand and Gayvert, 2005). Using this program, the experiment was split into two blocks: consonant categorization and vowel categorization. The order of the consonant and vowel categorization tasks was counter-balanced. Before each of the HFE-filtered categorization tasks, listeners participated in a short training

session that used non-filtered stimuli (different talkers than the experiment stimuli) and provided feedback. This was done to familiarize the participants with the ALVIN screen and buttons that they would be using for each task. Each button had two labels: a phonetic symbol and an example word with that phoneme in it. Only a subset of the answer choices were actually presented in the stimuli for the experiment, allowing for a more open test set and assessment of whether a specific phoneme was consistently mislabeled as another phoneme. The consonant options were **/p, b, t, d, k, g, f, v, m, n, s, z,** ʃ, ʒ, θ, ð, tʃ, dʒ, r, l, w, j, h/. The vowel options were **/i, æ, ɑ, o, u**, ɪ, e, ɛ, ɔ, ʊ, ʌ/. (The bold symbols represent the sounds that were actually produced as part of the stimulus set.) Table 1 shows the classification of the features of both the vowels and consonants. For each block, each CV stimulus was presented only once, resulting in each consonant being repeated five times (once with each vowel) and each vowel 13 times (once with each consonant) for both the female and male voice. Each stimulus was presented randomly over circumaural headphones with a reasonably flat response curve up to 15 kHz (±3 dB) and a frequency range to 25 kHz (Sennheiser HD 280). Because the empirical question was whether there is any relevant phonetic information present in HFE, we wished to make the task as easy on participants as possible. Participants were given the option to repeat the stimulus presented on each trial as many times as they wished. However, this option was not commonly used by participants. Across all participants and both blocks, 79% of the stimuli were never repeated and there were no obvious trends in the number of repeats across stimulus types. The number of repeats for the female (0.31 per trial) and male (0.29 per trial) were approximately equal. The most repeated stimulus was /v/ (0.52) and the least was /z/ (0.07). Fricatives tended to be repeated less frequently (with the exception of /v/) but not substantially less than the stops. The most repeated vowel was /æ/ (0.40) and the least repeated was /i/ (0.25).

3. Results

For the vowel categorization task, chance was 9.1%. Listeners' mean accuracy was 15.8% correct, which is significantly above chance [$t(12) = 4.38$, $p < 0.005$]. There was no significant difference in accuracy between the male and female voice stimuli (male: 14.3% and female: 17.3%; $t(12) = 2.1$, $p > 0.05$). Table 2 shows the confusion matrix for vowel categorization. From these tables, one can easily determine the amount of information transmitted at the phonetic feature level (Miller and Nicely, 1955).

Table 1. Classification of the vowel (left) and consonant (right) features. Abbreviations for the consonant place features are alveolar (Alv.), bilabial (Bilab.), labiodental (Labiod.), post-alveolar (P-alv.), voiced (V.), unvoiced (Unv.), affricate (Affr.), fricative (Fric.), and plosive (Plos.).

Vowel features				Consonant features							
Phon.	Front vs. Back	Height	Tense	Phon.	Place	Voicing	Manner	Phon.	Place	Voicing	Manner
i	Front	High	Tense	b	Bilab.	V.	Plos.	m	Bilab.	V.	Nasal
æ	Front	Low-mid	Lax	p	Bilab.	Unv.	Plos.	n	Alv.	V.	Nasal
ɑ	Back	Low	Tense	d	Alv.	V.	Plos.	ʒ	P-alv.	V.	Fric.
o	Back	High-mid	Tense	t	Alv.	Unv.	Plos.	dʒ	P-alv.	V.	Affr.
u	Back	High	Tense	g	Velar	V.	Plos.	tʃ	P-alv.	Unv.	Affr.
ɪ	Front	High-mid	Lax	k	Velar	Unv.	Plos.	ð	Dental	V.	Fric.
e	Front	High-mid	Tense	v	Labiod.	V.	Fric.	θ	Dental	Unv.	Fric.
ɛ	Front	Low-mid	Lax	f	Labiod.	Unv.	Fric.	r	Alv.	V.	Liquid
ɔ	Back	Low-mid	Tense	z	Alv.	V.	Fric.	l	Alv.	V.	Liquid
ʊ	Back	High-mid	Lax	s	Alv.	Unv.	Fric.	w	Bilab.	V.	Glide
ʌ	Back	Low-mid	Lax	ʃ	P-alv.	Unv.	Fric.	j	Palatal	V.	Glide
								h	Glottal	Unv.	Fric.

Table 2. Vowel confusion matrix that shows the percentage of times a stimulus was labeled as either the correct vowel (bold) or one of the other vowel options.

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

Listeners' accuracy was 54.9% correct in determining front versus back (chance = 50.9%), 34.5% for vowel height (chance = 23.6%), and 54.0% for tense (chance = 47.3%). Performance on each feature was significantly greater than chance ($p < 0.05$). Figure 1(A) displays the vowel results graphically.

For the consonant categorization task, chance was 4.3%. Listeners' mean accuracy was well above chance at 51.5% correct [$t(12) = 13.74$, $p < 0.001$]. There was a small but significant difference in accuracy for male (49.5% correct) versus female (53.5% correct) stimuli [$t(12) = 2.2$, $p < 0.05$]. Table 3 shows the confusion matrix for consonant categorization. Accuracy scores were 57.3% correct for place (chance = 19.7%), 91.4% correct for voicing (chance = 50.8%), and 82.0% correct for manner (chance = 28.4%). Performance on each feature was significantly greater than chance ($p < 0.05$). Figure 1(B) displays the results graphically.

4. Discussion

The purpose of the present study was to determine if there is perceptually relevant phonetic information in HFE. Listeners were presented with CVs that were high-pass filtered above 5.7 kHz with a speech noise masker present in the lower frequencies.

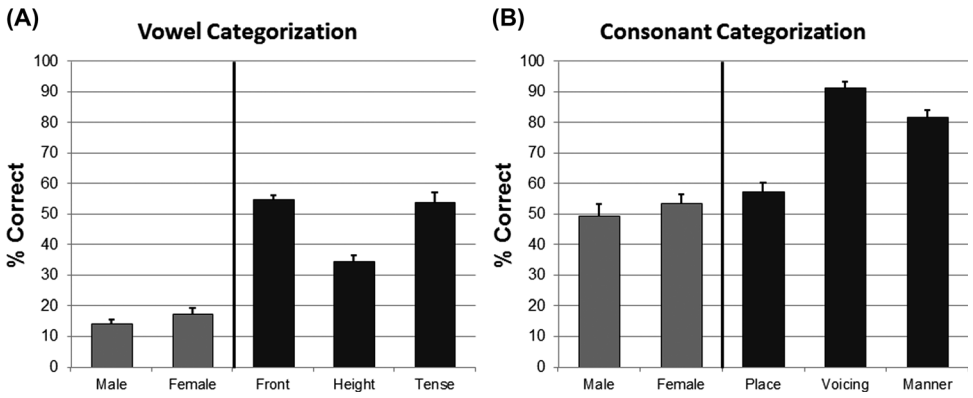

Fig. 1. (A) Vowel categorization: percent correct for male voice and female voice (left) and for vowel characteristics (right). (B) Consonant categorization: percent correct for male and female voice (left) and for consonant characteristics (right). Error bars indicate standard error of the mean.

Table 3. Consonant confusion matrix that shows the percentage of times a stimulus was labeled as either the correct consonant (bold) or one of the other consonant options.

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

Despite the absence of acoustic attributes that are the traditional focus of speech perception research (e.g., the first three formants), listeners performed above chance on both vowel and consonant categorization. As may be expected given the absence of lower frequency formant information, vowel categorization performance was poor but still significantly above chance. On the other hand, performance on consonant identification was surprisingly good (51.5% correct with chance equal to 4.3%). One might expect that this high level of performance was due to fricative categorization since fricatives have a great deal of energy in the higher frequencies. However, average performance on stop consonant categorization (50.6%) was nearly as good as average accuracy for the fricatives (59.7%). One might also predict greater performance using HFE for female speech given that female speech tends to have greater HFE overall (Monson *et al.*, 2012a). Whereas accuracy was greater for female-produced tokens, these differences were relatively modest (3 percentage points for vowels and 4 percentage points for consonants).

These results provide evidence for the presence of phonetic information in HFE, especially for consonants. Previous results had demonstrated that the presence of HFE can aid speech intelligibility (e.g., Apoux and Bacon, 2004; Moore *et al.*, 2010). Listeners could be using spectral cues, temporal cues, or a combination of both from the high frequency bands. Temporal cues are known to be useful for speech intelligibility in highly degraded signals (e.g., Shannon *et al.*, 1995), and research on amplitude envelope spectra has demonstrated that high frequency bands contain temporal cues relevant to speech intelligibility (LeGendre *et al.*, 2009). The results of the present

study indicate that substantial information about consonant voicing, place of articulation and manner is perceptually available solely from HFE even in the presence of a continuous low-frequency masker.

The presence of phonetic information in HFE may have particular relevance for speech perception in noisy environments in which the signal-to-noise ratio is poor for the lower frequencies. Because HFE is far more directional than lower frequencies in speech (Monson *et al.*, 2012b), HFE in the signal from a speaker facing the listener (presumably an interlocutor) is often not as degraded by ambient speech from other talkers as is the energy in the lower frequencies. These are considerations worth noting in the determination of appropriate bandwidths for auditory enhancement devices and telephony.

Acknowledgments

This work was supported by an F31 pre-doctoral grant from NIH-NIDCD to B.B.M. (Grant No. DC0105332) and by an NIH-NIDCD R01 to A.J.L. (Grant No. DC006859).

References and links

ANSI (**1992**). ANSI S3.42-1992, *American National Standard Testing Hearing Aids with a Broad-Band Noise Signal* (American National Standards Institute, New York).

Apoux, F., and Bacon, S. P. (**2004**). "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," J. Acoust. Soc. Am. **116**, 1671–1680.

Fullgrabe, C., Baer, T., Stone, M. A., and Moore, B. C. J. (**2010**). "Preliminary evaluation of a method for fitting hearing aids with extended bandwidth," Int. J. Audiol. **49**, 741–753.

Hillenbrand, J. M., and Gayvert, R. T. (**2005**). "Open source software for experiment design and control," J. Speech Lang. Hear. Res. **48**, 45–60.

LeGendre, S. J., Liss, J. M., and Lotto, A. J. (**2009**). "Discriminating dysarthria type and predicting intelligibility from amplitude modulation spectra," J. Acoust. Soc. Am. **125**, 2530.

Lippmann, R. P. (**1996**). "Accurate consonant perception without mid-frequency speech energy," IEEE Trans. Speech Audio Process. **4**, 66–69.

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

Monson, B. B., Hunter, E. J., Lotto, A. J., and Story, B. H. (**2014a**). "The perceptual significance of high-frequency energy in the human voice," Front. Psych. **5**, 587.

Monson, B. B., Hunter, E. J., and Story, B. H. (**2012a**). "Horizontal directivity of low- and high-frequency energy in speech and singing," J. Acoust. Soc. Am. **132**, 433–441.

Monson, B. B., Lotto, A. J., and Story, B. H. (**2012b**). "Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives," J. Acoust. Soc. Am. **132**, 1754–1764.

Monson, B. B., Lotto, A. J., and Story, B. H. (**2014b**). "Detection of high-frequency energy level changes in speech and singing," J. Acoust. Soc. Am. **135**, 400–406.

Moore, B. C. J. (**2012**). "Effects of bandwidth, compression speed, and gain at high frequencies on preferences for amplified music," Trends Amplif. **16**, 159–172.

Moore, B. C. J., Fullgrabe, C., and Stone, M. A. (**2010**). "Effect of spatial separation, extended bandwidth, and compression speed on intelligibility in a competing-speech task," J. Acoust. Soc. Am. **128**, 360–371.

Moore, B. C. J., Stone, M. A., Fullgrabe, C., Glasberg, B. R., and Puria, S. (**2008**). "Spectro-temporal characteristics of speech at high frequencies, and the potential for restoration of audibility to people with mild-to-moderate hearing loss," Ear Hear. **29**, 907–922.

Pollack, I. (**1948**). "Effects of high pass and low pass filtering on the intelligibility of speech in noise," J. Acoust. Soc. Am. **20**, 259–266.

Pulakka, H., Laaksonen, L., Yrttiaho, S., Myllyla, V., and Alku, P. (**2012**). "Conversational quality evaluation of artificial bandwidth extension of telephone speech," J. Acoust. Soc. Am. **132**, 848–861.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.