ACOUSTIC SALIENCE AND NATURALNESS IN VOWEL RECALL

Maya L. Barzilai

Georgetown University
mlb290@georgetown.edu

ABSTRACT

This paper presents the results of an Immediate Serial Recall (ISR) study in which sequences of syllables were heard and repeated by native American English speakers. Stimulus sequences varied in vowel duration, naturalness, and overall stimulus duration. The results show that natural-sounding syllables were easier to recall than unnatural-sounding syllables. Additionally, and perhaps surprisingly, shorter vowels were marginally easier to recall than longer ones.

These findings have several implications for speech perception and salience. First, they show that natural-sounding stimuli have an advantage over unnatural-sounding stimuli in this type of speech processing. Second, the results show that there is not a linear relationship between vowel duration and salience: longer vowels are in fact harder, not easier, to recall than shorter vowels. Finally, these results suggest that stimuli in psycholinguistic experiments should be as natural-sounding as possible, as unnatural-sounding stimuli may produce results that are less representative of natural speech processing.

Keywords: vowels, acoustic salience, duration, recall

1. INTRODUCTION

Several different properties of a speech sound can impact the way it is processed. For example, sounds with longer durations and higher intensities are considered more salient than those that are shorter or quieter [3, 4]. However, speech sound processing is influenced by factors other than the acoustics of the sounds. For example, adults distinguish between sounds they believe to be non-linguistic better than those they believe to be part of natural speech [8]. The way these different properties of sound stimuli interact to influence performance on certain psycholinguistic tasks is not yet clear.

This study aims to investigate the relative effects of several properties of speech stimuli on vowel recall. The experiment conducted here builds upon the surprising results of a previous study, which showed that shorter vowels were easier to recall than longer vowels [1]. These results contradict predictions made by the supposedly higher acoustic salience of longer vowels. It is possible that this unexpected result was not due to the relative durations of the vowels, but rather to the unnatural sound of the stimuli used. To this end, the present study manipulates vowel duration, stimulus naturalness, and syllable spacing to determine which of these factors has the strongest influence on the rates at which syllable sequences are accurately recalled and reproduced.

2. BACKGROUND

Though there is no single acoustic correlate to salience, there are certain acoustic properties of a sound that can contribute to its overall perceptual salience. For example, vowels are longer, louder, and have an acoustic steady state that is not characteristic of most consonants; due to these acoustic properties, vowels are said to be more acoustically salient than consonants [3, 4]. This high acoustic salience of vowels impacts their recall in ISR studies. Vowels have been showed to be remembered more accurately than consonants, both when presented visually [5] and auditorily [3, 7, 1]. However, this result does not hold across speakers of all languages. Speakers of languages that exhibit templatic morphology, in which the lexical root is comprised solely of consonants, recall consonants and vowels with equal accuracy [7, 1]. For these speakers, the effect of a language's morphophonology is apparently strong enough to outweigh the effect of acoustic salience in recall.

Given that the relatively long duration of vowels has been claimed to contribute to their overall salience, it would be reasonable to assume that lengthening vowels would make them more acoustically salient and therefore easier to recall. If this were the case, the effect of morphophonology on ISR, which has been shown to be stronger than the effect of acoustic salience [7, 1], would in turn be counteracted by an increase of the vowel salience. The result would be that speakers of all languages, even those with templatic morphology, would remember longer vowels better than they did consonants. However, this hypothesis has been disproven.

When vowel duration was manipulated in an ISR task, it was the shortened vowels that were remembered more accurately than lengthened vowels instead of the longer vowels being easiest to recall. This result held for speakers of English, Arabic, and Amharic, showing that the morphophonology of a language did not have an impact on the recall of long versus short vowels [1].

It could be the case that it was not vowel duration but rather another property of the stimuli that led to these surprising ISR results. For instance, it is possible that these digitally-manipulated stimuli were perceived as unnatural or even non-speech stimuli. The perception of a stimulus as linguistic or non-linguistic has been shown to impact results in speech processing tasks. Adults discriminate phonemes categorically, but when they are asked to discriminate non-linguistic stimuli, or are coached to attend to differences in consonants that may not be contrastive in their native language, their discrimination abilities become more fine-graned [11, 8]. The substitution of non-linguistic for linguistic stimuli also improves discrimination by adults perceiving non-native differences in stress [6] and tone [9]. These results show that the extent to which a given sound is perceived as natural language impacts the way it is processed by the listener. Therefore, the perception of the manipulated stimuli in previous studies [1] as somehow unnatural or non-linguistic could have influenced the ISR results more than the mere duration of the vowels.

Following these surprising findings, the purpose of this study is to investigate the effects of several acoustic properties of stimuli on recall accuracy. This study follows the assumption used in previous studies [5, 3, 7, 1] that recall correlates with salience, such that higher recall scores can be interpreted as the result of higher acoustic salience of the sounds being recalled. Vowel duration and stimulus naturalness are both manipulated, in order to determine whether the relationship between long vowels and short vowels is the same regardless of stimulus naturalness. In addition, overall stimulus duration is manipulated in this study, with some stimuli containing syllables that are dispersed in time and others containing syllables that are more condensed. This factor is included as it may be that longer vowels are easier to remember, but only if the total duration of the stimulus is below a certain threshold for recall. Determining the relative effects of these manipulations has implications for recall and speech sound processing, as well as for stimulus design in future studies that aim to investigate the perception of natural speech.

3. METHODS

3.1. Participants

Twenty-four native speakers of American English participated in this study. All were students at Georgetown University, ages 18-21 (mean age = 18.75). Participation in this experiment was completed in exchange for course credit.

3.2. Materials

The stimuli in this study were sequences of six CV syllables that varied in vowel duration (long or short), acoustic naturalness (natural or unnatural), and syllable spacing (dispersed or condensed). The eight stimulus types are schematized in Table 1. The six syllables in each sequence had the same consonant but different vowels (e.g., “ma mi mu mu ma mi”).

Table 1: Schematization of stimulus types differing in vowel duration, naturalness, and syllable spacing.

		Long Vowels
Dispersed	Natural	— — — — — —
	Unnatural	═ ═ ═ ═ ═ ═
Condensed	Natural	— — — — — —
	Unnatural	═ ═ ═ ═ ═ ═
		Short Vowels
Dispersed	Natural	- - - - - -
	Unnatural	= = = = = =
Condensed	Natural	- - - - - -
	Unnatural	= = = = = =

Each of the nine possible syllables generated from inventory /m z k i u a/ was recorded by an American English speaker. In each stimulus sequence, all of the vowels were either lengthened or shortened from their original recordings. Vowel duration manipulations were conducted in one of two ways. In the unnatural manipulation, the four consecutive glottal pulses of each syllable with the highest intensities were identified. These glottal pulses were then either deleted, for unnatural shortened vowels, or doubled, for unnatural lengthened vowels. The syllables resulting from this manipulation differed in both duration and overall intensity, making them distinct in salience along two acoustic dimensions. In the natural manipulation, the duration tier in Praat [2] was used to increase or decrease the overall duration of the vowel portion of each syllable. Intensity was not altered in this manipulation, and therefore the resulting natural-sounding syllables only differed from each other in duration.

All short vowels and all long vowels had approximately the same duration, regardless of naturalness

manipulation and segments. Native English speakers confirmed that the unnaturally-manipulated syllables were perceptibly digitized and that the naturally-manipulated syllables sounded more like natural speech.

Stimuli also differed in spacing, as a means to control for effect of overall stimulus length; some stimuli sequences were dispersed to last for a total of about seven seconds, whereas the syllables in the other sequences were more condensed, with shorter periods of silence in between. For the condensed sequences, a period of silence was added at the beginning, such that the syllables in all sequences were aligned to the end of the recording, as shown in Table 1.

Filler sequences were made up of syllables with the same vowel but different consonants (e..g, "ma ka za ka za ma"). In the distractor sequences, the syllables were presented as they were originally recorded, with no acoustic manipulations.

3.3. Procedure

The study was conducted in a sound-attenuated booth. Stimulus sequences were presented auditorily on a laptop computer via PsychoPy [10]. The stimuli were presented in four pseudo-randomized blocks, and participants were permitted to take short breaks between each of the blocks. In each trial, the laptop screen was gray as each syllable sequence played. The participants were instructed to repeat the stimulus aloud when the screen turned blue, 1500ms after the end of the stimulus recording.

Responses were recorded and coded for accuracy. Each stimulus received one point if it was repeated accurately and zero points if any of the repeated syllables was incorrect. If fewer than six syllables were produced in the response, zero points were awarded. If more than six syllables were produced, the first six syllables were scored for accuracy and all syllables beyond the sixth were ignored. In these cases, the sequence received a point if the first six syllables produced matched the six syllables in the stimulus sequence; otherwise, the sequence received zero points.

4. RESULTS

The mean score of all 24 participants was 0.39 out of a possible mean of 1.0, with a standard deviation of 0.16. The mean scores of nine participants were more than one standard deviation away from this overall mean, and therefore the data from these participants were removed. Of these removed participants, four had a mean score significantly below the group mean and five had a mean score significantly above the group mean. Excluding these participants therefore eliminated the presence of floor or ceiling effects. The remaining 15 participants had a mean score of 0.38 with a standard deviation of 0.09. This subset of the data is reported on below.

A mixed effects logistic regression model was run to predict the mean scores. The fixed effects were vowel duration, naturalness, and syllable spacing. The random effect was participant. The results of the regression model are in Table 2.

Table 2: Mixed Effects Logistic Regression Results.

Factor	P-value
Vowel Duration	0.063 .
Naturalness	0.045 *
Spacing	0.163

Naturalness was the only factor that achieved statistical significance (p=0.045), though vowel duration was marginally significant (p=0.063). There was no significant effect of stimulus spacing. The mean scores are shown in Figure 1; differences in stimulus spacing are collapsed here, as this factor did not significantly impact the scores.

Figure 1: Mean scores by vowel duration and naturalness.

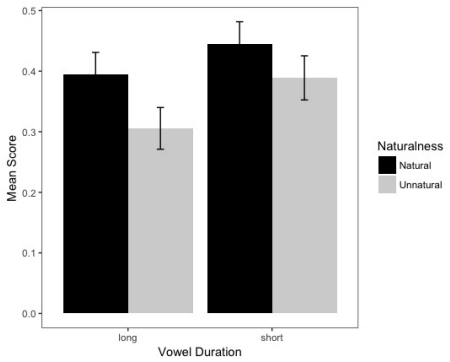

Figure 1 reveals that the natural stimuli were significantly easier to remember than the unnatural stimuli. Though the effect of vowel duration did not reach significance, the plot also shows that short vowels were on the whole more easily remembered than long vowels.

5. DISCUSSION

The results of this experiment show that natural-sounding stimuli were easier to remember than unnatural-sounding stimuli. This finding adds an important piece to the literature on the perception

of linguistic or natural versus non-linguistic or unnatural stimuli. As discussed above, it has been shown that non-linguistic perception is more accurate than phonological processing in discrimination tasks. The results here suggest that in a different task, one that requires perceiving, remembering, and repeating syllables, it is the more natural stimuli that are remembered better. In other words, while non-linguistic stimuli are easier to discriminate, stimuli most similar to natural speech seem to be easier to recall.

This finding not only sheds more light on the question of linguistic versus non-linguistic processing, but also has implications for stimulus design in psycholinguistic experiments. Sounds resembling natural language are processed more easily than those that sound digitally manipulated. Therefore, studies making claims about the processing of natural speech should be designed to use stimuli that resemble natural speech as closely as possible. If, conversely, studies use non-linguistic or otherwise non-natural stimuli, it may be the case that findings from those studies do not perfectly extend to the processing of natural speech processing. Studies using natural stimuli produce results that more closely approximate the perception of natural speech.

Though previous findings on recall of lengthened versus shortened vowels employed unnatural-sounding stimuli [1], the results of this study do not suggest that those findings, or findings from any study using audibly manipulated speech, are incorrect or wholly unreliable. In fact, the unexpected result that shorter vowels were easier to remember than longer vowels [1] was replicated in the present study, even when naturalness was controlled.

Though the effect of vowel duration did not achieve statistical significance, Figure 1 reveals a trend in which short vowels are remembered more accurately than long vowels. It is possible that with a larger sample size than the one examined here, this trend would have reached statistical significance. This pattern in the results of the present study is in line with previous findings [1], but remains surprising. If the relatively long durations of vowels is what makes them more salient, given the assumption that salience predicts accuracy in ISR tasks, then it should be the case that longer vowels are more salient and therefore more accurately remembered. The fact that the results results show the opposite trend suggests that there is not a linear relationship between vowel duration and salience in recall. Rather, it may be the case that there is an acoustic 'sweet spot,' such that longer vowels are more salient than shorter vowels, as has been widely suggested in the literature, but only up to a certain point. Past this ideal duration range, increasing the duration of vowels may not result in higher recall. Future work will determine whether this range of optimal duration exists, and if so, past what threshold recall rates stop improving with vowel duration increases.

The notion of an acoustic sweet spot with respect to salience is logical. In addition to duration, intensity has also been claimed to correlate with salience, with louder sounds being easier to perceive than quieter ones. However, it is easy to imagine that past a certain intensity threshold, sounds are no longer easier to perceive. In fact, a sound that is too loud is probably more difficult to process than one that is simply loud relative to its surrounding sounds. Therefore, as this study reveals that the relationship between duration and salience may not be linear, replicating previous findings, future research should investigate other acoustic properties of speech sounds and their relationship to acoustic salience.

6. CONCLUSION

This paper has presented the results of an ISR study in which speakers were asked to recall sequences of CV syllables. The results show that natural-sounding stimuli were easier to recall than unnatural-sounding stimuli. Though not a statistically significant difference, shorter vowels trended towards being easier to remember than longer vowels, replicating results from a previous study. Overall, the results imply that there is a meaningful difference in the processing of natural and unnatural speech, such that natural stimuli have the advantage in recall. This finding also implies that stimuli should be designed to imitate natural speech in order for experimental results to most accurately reflect speech processing. Finally, the tendency of shorter vowels to be remembered more accurately than longer vowels suggests that duration is not directly correlated with acoustic salience, but rather that there may simply be a range within which segment duration is optimized for speech processing.

7. ACKNOWLEDGEMENTS

Special thanks to Emily Schluper for her hard work on data coding and analysis for this project. Thanks also to Georgetown University's PhonLab for feedback on previous versions of this work.

8. REFERENCES

[1] Barzilai, M. L. 2019. Effects of templatic morphology on segmental recall. Paper presented at the Linguistics Society of America annual meeting. https://drive.google.com/file/d/1FaYbtt0tLMPPcxhEGdohgpn0QwlAQzsV/view.

[2] Boersma, P., Weenink, D. 2017. Praat: doing phonetics by computer. *http://www. praat. org/*.

[3] Crowder, R. G. 1971. The sound of vowels and consonants in immediate memory. *Journal of Verbal Learning and Verbal Behavior* 10(6), 587–596.

[4] Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., Van Ooijen, B. 2000. Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & cognition* 28(5), 746–755.

[5] Drewnowski, A. 1980. Memory functions for vowels and consonants: A reinterpretation of acoustic similarity effects. *Journal of Verbal Learning and Verbal Behavior* 19(2), 176–193.

[6] Dupoux, E., Pallier, C., Sebastian, N., Mehler, J. 1997. A destressing "deafness" in french? *Journal of Memory and Language* 36(3), 406–421.

[7] Kissling, E. M. 2012. Cross-linguistic differences in the immediate serial recall of consonants versus vowels. *Applied Psycholinguistics* 33(3), 605–621.

[8] Mann, V. A., Liberman, A. M. 1983. Some differences between phonetic and auditory modes of perception. *Cognition* 14(2), 211–235.

[9] Mattock, K., Burnham, D. 2006. Chinese and english infants' tone perception: Evidence for perceptual reorganization. *Infancy* 10(3), 241–265.

[10] Peirce, J. W. 2007. Psychopy - psychophysics software in python. *Journal of neuroscience methods* 162(1-2), 8–13.

[11] Werker, J. F., Tees, R. C. 1984. Phonemic and phonetic factors in adult cross-language speech perception. *The Journal of the Acoustical Society of America* 75(6), 1866–1878.