TYPE Brief Research Report
PUBLISHED 31 August 2022
DOI 10.3389/fcomm.2022.865587

OPEN ACCESS

EDITED BY
Mikkel Wallentin,
Aarhus University, Denmark

REVIEWED BY
Andreas Højlund,
Aarhus University, Denmark
Angarai Ganesan Ramakrishnan,
Indian Institute of Science (IISc), India

*CORRESPONDENCE
Richard Granger
richard.granger@dartmouth.edu

[†]These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 30 January 2022
ACCEPTED 04 August 2022
PUBLISHED 31 August 2022

CITATION
Brown AC, Childers E, Bowen EFW, Zuckerberg GA and Granger R (2022) Phonemes in continuous speech are better recognized in context than in isolation. *Front. Commun.* 7:865587. doi: 10.3389/fcomm.2022.865587

Phonemes in continuous speech are better recognized in context than in isolation

Annemarie C. Brown[†], Eva Childers[†], Elijah F. W. Bowen, Gabriel A. Zuckerberg and Richard Granger*

Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, United States

The contribution of context to phoneme perception is a subject of extensive study. In recent years, while the perception of phonemes in and out of context has become characterized as well-understood, new studies have emerged to challenge prevailing wisdom. Findings derived from rigorously controlled stimuli have failed to hold up when tested against continuous or more naturalistic speech, and vowels produced in isolation have been shown to possess different frequencies than vowels in spontaneous speech. In the present study, we examine the effect of context on vowel recognition, via stimuli taken directly from natural continuous speech in an audiobook. All tested vowel sounds, except /EH/, were better recognized with surrounding context than in isolation, affirming the resilience of findings from past studies.

KEYWORDS

vowel, phoneme identification, consonantal context, vowel recognition, continuous speech, speech perception

Introduction

Although context has been shown to aid in sentence perception (Martin and Bunnell, 1981; Tanenhaus et al., 1995; Fowler and Brown, 2000; Chambers et al., 2004), recent evidence has called into question the direction of the relationship between context and phoneme perception. Put simply, which process comes first in perceiving words from a stream of sounds? Do we take a purely bottom-up approach, identifying each phoneme first in isolation? In this case, consonants and enclosed vowel sounds may be parsed first from continuous sound, then additively combined to create syllables and words (e.g., /P/ + /EH1/ + /T/ = "PET"). Or does context allow phonemes to be perceived in a manner that is at least partially top-down, such that perceived syllables and words may influence the perception of phonemes themselves (e.g., perception of "PET" leads to identification of the enclosed vowels as /EH1/)?

In support for the latter top-down hypothesis, consonantal context has been shown to improve vowel identification in human listeners using a tightly controlled set of stimuli. When speakers produced consonant-vowel-consonant (CVC) syllables, and vowels-only tokens separately, vowel identification error rates were lower when presented in a CVC context compared to a lone vowel (Bischoff, 1976; Strange et al., 1976; Gottfried and Strange, 1980; Gottfried et al., 1985; Reinisch and Sjerps, 2013). This effect was also observed for similar tokens produced synthetically

(Millar and Ainsworth, 1972; Bischoff, 1976). However, because these studies used vowels produced in isolation, it is unclear whether these findings generalize to an externally valid model of language perception.

In fact, vowels produced in isolation have been consistently shown to have fundamental frequencies which differ from the frequencies of vowels produced in natural speech, as in a full text of sentences read aloud or in natural, spontaneous speech (Fitch, 1990; Murry et al., 1995; Moon et al., 2012; Iwarsson et al., 2020). This distinction is non-trivial as vowel frequency is demonstrably an important component of vowel quality and identification (Assmann and Summerfield, 1989; Hirahara and Kato, 1992; Owsianny, 2019). Due to the difference in frequency of vowels produced in isolation, these stimuli are not consistent with natural connected speech. Poorer identification performance of these vowels generated in isolation may be responsible for the contextual advantage observed found in previous studies.

Van Son and Pols (1999) identified a slight difference between the whole CVC token and the central, steady state portion of the vowel using Dutch tokens extracted from connected read speech. Firstly, studies have shown that the steady state portion of the vowel, even when extended, does not contain sufficient information for subjects to identify the vowel (Bond, 1976; Shankweiler et al., 1978). Van Son and Pols (1999) also removed all consonant information by removing the outer 20 ms of the vowel. The authors concluded that most errors occurred on similar sounding vowels but adding more context confused subjects. In our study, we provide subjects with as much vowel information as natural speech will allow. We also select similar sounding English vowels and find that context improves the identification for all vowels. Lastly, while both studies used connected read speech, our study uses an audiobook where the speaker reads with a conversational style, using a variety of inflections, tones, and pitches.

Here, we reexamine the question of whether consonant context provides an advantage in human vowel categorization of CVC syllables, relative to lone vowels presented without context using stimuli extracted from connected read speech. We aim to demonstrate that context provides an advantage in the identification of vowels in the enriched environment of natural speech, in which vowel frequency matches natural language use. Thus, we compare these naturally produced CVC tokens against vowel-only tokens extracted from the same CVC syllables, removing only the plosive and transitional silence. In the vowel-only condition, transitional coarticulatory regions are preserved to the extent that subjects anecdotally report hearing full syllables. If the use of vowels produced in isolation (not consistent with connected speech) are responsible for the contextual advantage in vowel identified, then we may observe that there is little to no contextual advantage when stimuli tokens from connected read speech are used. In this case, we cannot conclude that the previously demonstrated contextual advantage is a true attribute of speech perception; instead, we are left to consider that this finding may be an artifact of experimental conditions using artificial stimuli, as previously posited (Diehl et al., 1980; Macchi, 1980; Assmann, 1982). If, however, this contextual advantage remains in this native language space, then we may conclude that some influence of surrounding vowel context does assist in the categorization of enclosed vowels. This could suggest the use of top-down information from context influences speech perception at the phoneme level.

Methods and materials

Participants

Thirty-nine native speakers of American English with intact hearing (22 females; 18–33 years old; mean = 22.3 ± 4.08) participated in this study. All participants gave informed written consent in accordance with guidelines set by the Committee for the Protection of Human Subjects at Dartmouth College.

Stimuli

To assess whether the consonantal context advantage persists using stimuli from connected read speech, we compare responses to CVC stimuli with responses to the same stimuli without consonants. If the consonantal context advantage is a byproduct of vowels being produced separately from CVC tokens, then there may be no significant difference between the CVC and isolated vowel condition. If consonantal context is a true integral part of vowel identification, then we should observe superior performance in the CVC condition.

Stimuli were extracted from audiobooks, selected to emulate natural speech (the complete Harry Potter audiobook series). All the audiobooks were annotated using Speechmatics [Computer program]. This large corpus yielded a high volume of unique tokens across categories of phoneme trigrams. Words that did not match the CMU Pronouncing Dictionary version 0.7b and/or did not contain a vowel of interest were excluded. Silence was added to the start and end of the word. One hundred and eight unique stimuli were extracted from words in the corpus and normalized for loudness. These stimuli were selected from combinations comprising 3 consonantal contexts and 6 monophthong stressed vowels: /B-vow-K/, /P-vow-T/, and /T-vow-K/ with target vowels /EH1/, /IH1/, /AA1/, /AH1/, /UH1/, and /AE1/) to create a total of 108 full consonant-vowel-consonant (CVC) stimuli. The state-of-the-art, Montreal Forced Aligner (MFA), which transcribes phonemes to ARPAbet was used to determine the segmental boundaries between the consonants and vowels (McAuliffe et al., 2017). Plosive-vowel phoneme transitions were selected for this dataset due to their relative simplicity to segment. Each stimulus token was

TABLE 1 Conversion table from ARPAbet to phonetic spelling used in paradigm.

ARPAbet	IPA	Phonetic label
AA	[ɑ]	AH
AH	[ʌ]	UH
AE	[æ]	AA
EH	[ɛ]	EH
IH	[I]	IH
UH	[ʊ]	EUH

presented once as a full CVC phoneme trigram, and once as an isolated vowel, creating a total of 216 trials. The number of trials were matched for both context frequency (e.g., % of trials equal across /B-vow-K/, /P-vow-T/, and /T-vow-K/ and vowel frequency (e.g., % of trials equal across /BIHK/, /BEHK/, /BAAK/).

The duration of vowels, as segmented, are nominal for connected speech (Crystal and House, 1988) and necessarily identical in both conditions. However, CVC stimuli contain extra phonemes, making them inherently longer in duration (Supplementary Table 1). The difference in duration is unlikely to be a confounding variable in this paradigm since the only additional vowel-discriminative content in the CVC condition comes under the moniker of coarticulation. Coarticulation is a necessary component of a naturalistic phoneme trigram and therefore, an integral and primary effect in this study. Uninformative content is also present in the surrounding consonants and might cue vowel processing. We therefore provided visual cues (see Procedure) that a stimulus was impending.

Additionally, while most of the CVC trigrams used in this study are valid words, a majority of these (14/18) have word usage rate of <5% according to the Corpus of Contemporary American English (Supplementary Table 2). Any CVC extracted from real words will contain implicit circumstantial clues; we use those clues to understand vowels. The present study, in concord with previous studies, provides evidence that context helps and does not impair vowel identification.

Three human raters performed quality control for each stimulus token by listening for length of vowel, overall length of stimulus, and similarity of vowel sounds. The stimuli were not controlled for timbre, pitch or character accent emulated by the speaker. Any vowel not judged to be prototypical of its phoneme category by all three raters was removed. Vowel selection buttons were labeled using the phonetic spelling (e.g. /AA/ sounds like "ahh" and was provided the phonetic label of "AH"), see Table 1 for the conversions from ARPAbet to the phonetic spelling. Stimuli were presented via PsychoPy (Peirce et al., 2019) using headphones with a frequency response of 15 Hz–20 kHz.

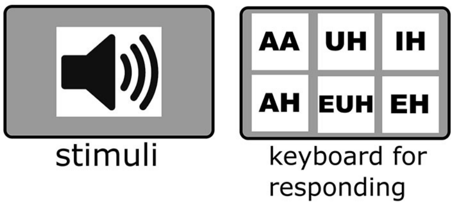


FIGURE 1
Experimental paradigm. Participants listen to randomized audio clips containing either CVC or vowel-only sounds then respond via a keyboard press to indicate which vowel sound they heard.

Procedure

An a priori power analysis was performed for sample size estimation using the R pwr package (v. 1.3-0) based on effect size from Jenkins et al. (1983). This analysis tested the difference between two independent group means using a two-tailed test, a medium effect size ($d = 0.78$) and an alpha of 0.05. Results showed that a sample of 34 participants was required to achieve a power of 0.80.

During the experiment, participants were seated in a dimly lit room $\sim$22$''$ from the monitor. Each participant was asked to adjust headphones to the individual's optimal volume before beginning the experiment; this setting was then held constant across all trials for the duration of the experiment.

To familiarize participants with the task of matching auditory stimuli to selected vowel identity, participants were presented with a training clip for each selection button presented in the paradigm (see Figure 1). Each clip contained a vowel sound and an example CVC word that contained the vowel (e.g., "AA" (/AE/) as in bat). A different-gendered speaker from the experimental trials was used for these training trials. To verify adequate training on each vowel's identity, participants were asked to verbally report each vowel sound before commencing experimental trials. A shorthand guide was also provided to the participant throughout the experiment.

Participants were instructed to identify the vowel present in the stimuli containing either a CVC word or isolated vowel. Participants were shown a visual indication of when the stimuli were about to play (speaker icon showed as playing). The participants could indicate their categorical choice via a specially labeled keyboard press (see Figure 1). The trials were fully randomized for each participant and there were 4 evenly spaced breaks.

Results

Trials resulting in correctly identified vowels were tallied for each condition (total correct CVC trials, total correct

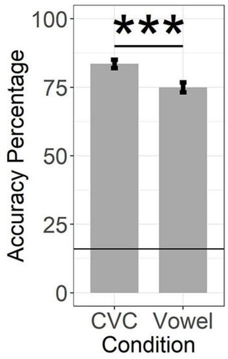
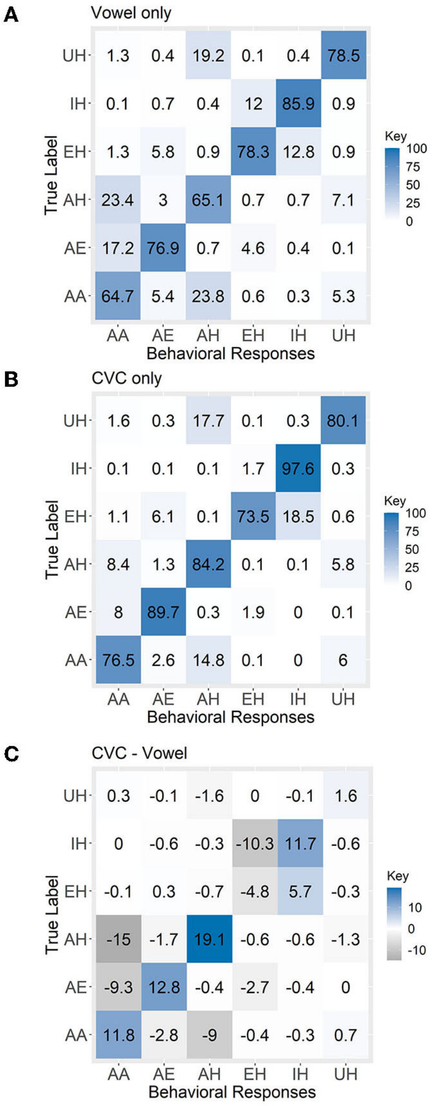


FIGURE 2
Average percent accuracy in identifying vowels in CVC vs. vowel-only conditions. Subjects were significantly more accurate in identifying vowels in the CVC condition. The value of chance was 0.16 (represented by the black horizontal line). Error bars represent standard error, $p < 0.0001$, Wilcoxon signed-rank test. ***$p < 0.0001$.

vowel-only), converted to a percentage (correct/total trials), and averaged across participants for each condition (Figure 2). We used a Wilcoxon signed-rank test to examine whether consonantal context (CVC) facilitated vowel identification, causing higher rates of accurate vowel identification than vowels presented in isolation. We observed a statistically significant ($z = 5.01$, $p < 0.0001$) advantage, by 8.59%, in average classification accuracy for vowels presented in consonantal context (CVC: $M = 83.57\%$, $SD = 9.37\%$) over isolated vowels (vowel-only: $M = 74.98\%$, $SD = 11.39\%$) (see Figure 2).

Confusion matrices display the nature of classification errors made in each condition. For each participant, we summed the total number of times each vowel was given as a response (e.g., participant responded with /EH/ on 20 trials) and then categorized based on the vowel identity for those trials (e.g., 10 trials were /EH/, 5 trials were /AH/, 5 trials were /IH/). Wherever the percentage across the diagonal is not 100% (e.g., the participant did not accurately identify all true /EH/ trials as EH), the rest of the column indicates what the participant reported instead. For each vowel sound except /EH/, we found that the percentage of accurately classified vowels was higher in the CVC condition compared to the vowel-only. Errors of both commission (Type I false positives) and omission (Type II false negatives) were more common for vowels presented in isolation (Figure 3) relative to CVC trials.

We calculated the mean subject's ability to detect a certain vowel (d') while treating all other vowels as distractors. Sensitivity (d') was calculated by measuring the difference between the z-scores of the hit rate and false alarm rate. The sensitivity was higher for each vowel across CVC trials than vowels presented in isolation (Figure 4), indicating that without exception, vowel identification was improved when vowels were



Vowel only

	AA	AE	AH	EH	IH	UH
UH	1.3	0.4	19.2	0.1	0.4	78.5
IH	0.1	0.7	0.4	12	85.9	0.9
EH	1.3	5.8	0.9	78.3	12.8	0.9
AH	23.4	3	65.1	0.7	0.7	7.1
AE	17.2	76.9	0.7	4.6	0.4	0.1
AA	64.7	5.4	23.8	0.6	0.3	5.3

CVC only

	AA	AE	AH	EH	IH	UH
UH	1.6	0.3	17.7	0.1	0.3	80.1
IH	0.1	0.1	0.1	1.7	97.6	0.3
EH	1.1	6.1	0.1	73.5	18.5	0.6
AH	8.4	1.3	84.2	0.1	0.1	5.8
AE	8	89.7	0.3	1.9	0	0.1
AA	76.5	2.6	14.8	0.1	0	6

CVC - Vowel

	AA	AE	AH	EH	IH	UH
UH	0.3	-0.1	-1.6	0	-0.1	1.6
IH	0	-0.6	-0.3	-10.3	11.7	-0.6
EH	-0.1	0.3	-0.7	-4.8	5.7	-0.3
AH	-15	-1.7	19.1	-0.6	-0.6	-1.3
AE	-9.3	12.8	-0.4	-2.7	-0.4	0
AA	11.8	-2.8	-9	-0.4	-0.3	0.7

FIGURE 3
Confusion matrix for **(A)** vowel-only condition, **(B)** CVC condition and **(C)** CVC—vowel only. Each cell of the confusion matrix was divided by the number of trials per vowel (e.g., 18 trials) and converted to a percentage. Along the diagonal are the percent of trials accurately classified as each vowel when vowel was present (i.e., responded /EH/ on true /EH/ trial). Down the vertical is each category of incorrect behavioral response broken down by true vowel identity (i.e., percent of trials identified as /EH/ which contained a true /IH/, a true /AA/, etc.; showing which vowels participants were incorrectly categorizing by vowel category/rate of Type I false positives/errors of commission). Across the horizontal are inaccurate participant responses from each vowel category broken down according to true vowel identity (i.e., true /EH/ trials incorrectly identified as /IH/, as /AH/, etc.; showing which vowels participants were failing to categorizing as a particular vowel category by incorrect label given, aka rate of Type II false negatives/errors of omission). All vowel labels are in ARPAbet. /EH/ was the only vowel that had a better recognition accuracy, by 4.8%, when presented as an isolated vowel than when it is presented in context.

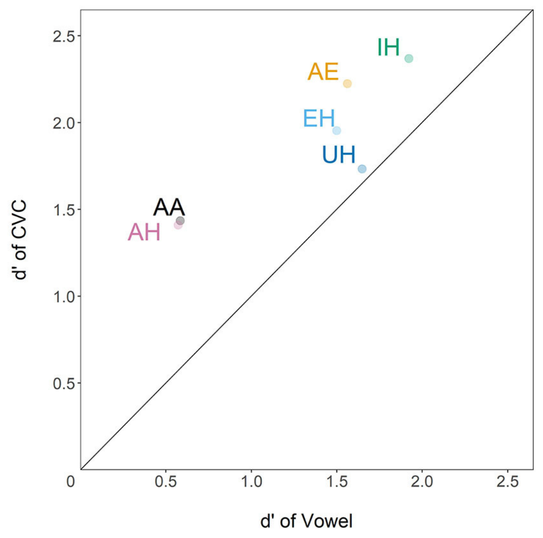


FIGURE 4
d' (*Zscore* (*hit*) − *Zscore* (*false alarm*)) for each vowel in vowel-only condition (X-axis) vs. d' of same vowels in CVC condition (Y-axis). All vowel labels are in ARPAbet. Every vowel sound in the CVC condition was above the center line indicating that the d' was higher than in the vowel-only condition.

facilitated by neighboring consonants. The numerical values of d' can be found in Supplementary Table 3.

Discussion

Many studies have examined the question of how context contributes to speech perception. There has been strong evidence that for vowel perception and identification, consonantal context is an important contributor (Bischoff, 1976; Strange et al., 1976; Gottfried and Strange, 1980; Gottfried et al., 1985). All these studies utilized vowels that were produced in isolation. Vowels produced in isolation have recently been shown to have frequencies different from vowels in spontaneous speech (Fitch, 1990; Iwarsson et al., 2020). Therefore, these previous results are challenged by a potential frequency confound. We tested a simple instance of vowels from real continuous speech, clipped either to isolation (removing only the plosive and transitional silence), or surrounded by consonants. Our results demonstrated that the consonantal context advantage in vowel identification persisted for stimuli tokens extracted from read continuous speech. For all vowels except /EH/, vowel sounds were better recognized with surrounding consonant context than in isolation; this effect is in contrast to Van Son and Pols (1999) (Figure 3). However, the sensitivity to each vowel sound was increased when presented in context vs. as an isolated vowel (Figure 4, Supplementary Table 3).

Further qualifications to this effect are worthy of consideration. When examining classification sensitivity in context, we observed that /EH/ was comparable to the other vowels examined (Figure 4), despite a lack of improvement in raw hit rate across conditions. This effect is propelled by a marked reduction in false alarms. When presented in context, the rate at which alternative vowels were mistaken for /EH/ was unmatched, achieving the lowest rate of false positives of any vowel condition. That is, participants observing a true /EH/ in context may mislabel the instance as an alternative vowel, but it was exceedingly uncommon to mistake another vowel for /EH/. This observation could suggest that the base rate for marking a consonant-encased vowel as /EH/ was markedly low, below chance, relative to all other vowels whether alone or in context. Overall, when presented in context, participants appeared to most commonly mislabel /EH/ as /IH/. This is consistent with a previously established index of vowel confusability, where /IH/ and /EH/ have shown the highest rates of confusability compared to a wider selection of vowels (Weber and Smits, 2003).

Nonetheless, it is worth considering how this might differ were we to make use of all possible vowels standard to American English. Our use of a pre-existing non-academic corpus was essential to our primary aim. However, it imposed limitations on the full range of possible vowels in context. This restricted the use of CVC combinations to ones with sufficient instances to ensure equal stimuli frequency; thus, not all possible English vowel sounds were included. Limitations on a complete set of possible sounds were overwhelmingly necessary for our purposes, in service of careful attention to the nuances of potential confounds and natural quirks of the English language. We discuss each of these limitation cases in turn.

Case 1: Diphthongs were intentionally excluded. Monophthong vowels are largely characterized by a steady state, holding a primarily stable formant pattern. On the other hand, diphthongs, by definition, transition between the formant patterns of two distinct vowel sounds, a confounding hallmark of a distinct class of vowel. This places diphthongs beyond the scope of this study, in which we removed transitions between separate sounds. This intentional exclusion is consistent with prior study on the CVC advantage (Strange et al., 1976, 1979; McMurray et al., 2013), and thus important to take into consideration as we challenge and extend these findings.

Case 2: Vowel durations were limited to a similar approximate range, as perceived by independent raters, which skewed shorter in length. Control of duration was essential to ensure that findings were not confounded by varying exposure time between vowel categories. Therefore, our stimuli are shorter to accommodate the lowest common denominator in length of prototypical vowel category.

Case 3: A within-subjects design allowed us to eliminate person confounds entirely, a crucial boon to such a delicate task. The differences between categories and conditions were quite subtle, so a round-robin design in which each participant would be exposed only to a limited set of all vowel and CVC types in question would introduce a great deal more noise to our finding's signal, when our goal was to refute the influence of extraneous confounds from prior work on this topic. Thus, it remains untenable to study a full range of all possible CVC types potentially relevant to this effect in a session whose length does not extend beyond the attentional resources of the participant.

Case 4: Vowels occurring *naturally* in isolation played no part in this study. It may be tempting to highlight this absence as a weakness of the study's design, but it lays bare an important puzzle of natural language use that extends well beyond instances in our corpus of choice. Vowels in isolation simply do not appear as often as in words in typical American English, or even at all for many types, except for modifiers such as "a" (when pronounced colloquially as /AH/), or non-lexical utterances (as in fillers /EH/, /AH/).

Our selection criteria allowed for a thoughtful balance of internal and external validity, which offered a firm foundation to demonstrate our ultimate intention: to show that context aided in vowel perception in connected speech, confirming that this advantage was not an artifact of prior work, but a true attribute of speech perception. Nonetheless, we recognize that there is ample room for future extensions. Future studies could push the boundaries that we held here, beginning with making use of a larger corpus with further variation in vowel types, or truly broadening horizons of this effect by exploiting the flexibility of a multilingual sample. In the latter extension, a language which supports more regular vowels in isolation may be tested alongside this effect in American English. Or, perhaps, we might build toward a stronger case for our final tenet put forth in our initial hypotheses: implications for a broader consideration of top-down influences in speech perception.

As originally posited, our findings suggest that use of top-down information from context influences speech perception at the phoneme level. We must acknowledge that while our findings suggest this broader interpretation, they do not insist upon it. Consider, for example, an alternative explanation: What if vowels aren't specific, immovable categories, but rather groups with inherent variation based on context? Perhaps an /AE/ alone is not perceived as the same sound as /B/+/AE/, implying that vowel identities may be better described as context-specific allophones. In this case, our study is hindered by its main asset, in which we specifically derived our isolated vowel from an existing context frame to keep vowels at a conversational frequency. In eliminating easily identified features of coarticulation, we dropped features of plosion and transitional silence; perhaps the /AE/ inside of /B//AE//T/ is thus lost from a real piece of its basic identity. But if /AE/ (as featured in "bat" and "cat") is cognitively categorized not as a single whole, but rather stored as two unique tokens, "ba-" or "ca-," this increased level of complexity must be observed and factored into perceptual categorization, as further subtleties of ground-level information must be observed and accounted for. This seems particularly taxing when considering that a vowel may be expressed among a great number of unique combinations of consonants, and each identity might vary accordingly. Further, the type of information available to consonants surrounding a vowel vary considerably (features of a sibilant including a period of turbulent noise, for example, vs. a plosive consonant with an observable release burst), and context both precedes and follows the placement of the enveloped vowel.

On the one hand, this offers a possible "bottom-up" interpretation wherein each unit of sound is processed individually and additively combined to yield meaning from the sounds of speech; that is, we may be organizing "speech sounds" along more complex combinations of multiple phonemes, but nonetheless deriving meaning from purely additive sequences of sound units. If true, this pushes against our current characterization of vowel phonemes. A true allophone requires that subtle variations in sound do not cross a line such that a minimal pair would change a phoneme's identity. While a native speaker would not acknowledge a new identity of vowel between "bat" and "cat," perhaps this conscious delineation does not fully encompass a greater complexity of sub-phoneme variation which we nonetheless computationally differentiate. Although we are unable to rule out this account of how we mentally organize speech sounds, it complicates the mental resources necessary to support such a system.

A more parsimonious explanation simply puts forth that English-speaking perceivers expect phonemes to occur in groups, and struggle to understand them in isolation. This explanation coincides with how we utilize experience to cumulatively facilitate automation of future categorizations. Humans are constantly exposed to an overwhelming volume of information input and benefit greatly from such a probabilistic framework that allows one to rapidly categorize, and thus react appropriately to, new encounters. Such a "top-down" interpretation neatly combines prior research demonstrating the advantage of context holds over artificially generated vowels, therefore fully absent of any acoustic influence of consonants, with our results toward a logical assumption of how the human mind typically endeavors to make sense of information. Although it is not yet possible to declare this interpretation as true, the results herein may be used in concert with future work to further close the gap between this explanation's likelihood, and perhaps, its certainty. Regardless, whether our effect is explained by a more intricate granularity of bottom-up input properties, or a simple top-down context effect, each possibility intriguingly suggests that English phonemes, though parse-able to a native English speaker as different sounds, may not exist in the brain as auditory categories, but rather as secondary features of syllables (as allophones, whose acoustic characteristics change subtly with context) or patterns (as frequently encountered patterns of speech sounds become more readily recognizable, providing additive cues to the identity of enclosed sounds). As such, this simple effect is worthy of consideration as we continue to refine a model of how phonemes interact to aid in speech perception.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary material. Available online at: https://drive.google.com/drive/folders/1mDyteHkPITgKE6h7iG-BMuxJQdS-5zr-?usp=sharing.

Ethics statement

The studies involving human participants were reviewed and approved by Committee for the Protection of Human Subjects, Dartmouth College. The patients/participants provided their written informed consent to participate in this study.

Author contributions

AB and RG contributed to the design of the study. EB wrote the code that generated the stimuli used in this study. AB and EC selected the specific stimuli clips to be used in the experiment. AB, EC, and GZ performed quality assurance on all the stimuli. EC led the development of the experimental paradigm and testing parameters with assistance from GZ. EC led the data collection with assistance from GZ. EC and AB contributed to the statistical analyses. EC and AB wrote the manuscript and integrated revisions from AB, EB, and GZ. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by grants from the Office of Naval Research.

Acknowledgments

We would like to thank Joachim Lohn-Jaramillo, Khari-Elijah Jarrett, and Jack Burgess who all provided insightful discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.865587/full#supplementary-material

References

Assmann, P. F. (1982). Vowel identification: orthographic, perceptual, and acoustic aspects. *J. Acoust. Soc. Am.* 71, 975–989.

Assmann, P. F., and Summerfield, Q. (1989). Modeling the perception of concurrent vowels: vowels with the same fundamental frequency. *J. Acoust. Soc. Am.* 85, 327–338. doi: 10.1121/1.397684

Bischoff, D. M. (1976). Secondary acoustic characteristics and vowel identification. *J. Acoust. Soc. Am.* 60, S90–S90. doi: 10.1121/1.2003598

Bond, Z. B. (1976). Identification of vowels excerpted from/l/and/r/contexts. *J. Acoust. Soc. Am.* 60, 906–910.

Chambers, C. G., Tanenhaus, M. K., and Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *J. Exp. Psychol. Learn. Memory Cognit.* 30, 687–696. doi: 10.1037/0278-7393.30.3.687

Crystal, T. H., and House, A. S. (1988). The duration of American-English vowels: an overview. *J. Phonetics* 16, 263–284. doi: 10.1016/S0095-4470(19)30500-5

Diehl, R. L., McCusker, S. B., and Chapman, L. S. (1980). Perceiving vowels in isolation and in consonantal context. *J. Acoust. Soc. Am.* 69, 239–248. doi: 10.1121/1.385344

Fitch, J. (1990). Consistency of fundamental frequency and perturbation in repeated phonations of sustained vowels. *J. Speech Hear. Disord.* 55, 360–363.

Fowler, C. A., and Brown, J. M. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Percept. Psychophys.* 62, 21–32. doi: 10.3758/BF03212058

Gottfried, T. L., Jenkins, J. J., and Strange, W. (1985). Categorial discrimination of vowels produced in syllable context and in isolation. *Bull. Psychon. Soc.* 23, 101–104. doi: 10.3758/BF03329794

Gottfried, T. L., and Strange, W. (1980). Identification of coarticulated vowels. *J. Acoust. Soc. Am.* 68, 1626–1635. doi: 10.1121/1.385218

Hirahara, T., and Kato, H. (1992). "Effect of F0 on vowel identification," in *Speech Perception, Production and Linguistic Structure*, 89–112.

Iwarsson, J., Hollen Nielsen, R., and Næs, J. (2020). Mean fundamental frequency in connected speech and sustained vowel with and without a sentence-frame. *Logoped. Phoniat. Vocol.* 45, 91–96. doi: 10.1080/14015439.2019.1637455

Jenkins, J., Strange, W., and Edman, T. R. (1983). Identification of vowels in "vowelless" syllables. *Percept. Psychophys.* 34, 441–450. doi: 10.3758/BF03203059

Macchi, M. J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *J. Acoust. Soc. Am.* 68, 1636–1642. doi: 10.1121/1.385219

Martin, J. G., and Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *J. Acoust. Soc. Am.* 69, 559–567.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderregger, M. (2017). *Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi*. Available online at: https://montrealcorpustools.github.io/MontrealForcedAligner/images/MFA_paper_Interspeech2017.pdf

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., and McEchron, W. (2013). Infant directed speech and the development of speech perception: enhancing development or an unintended consequence? *Cognition* 129, 362–378. doi: 10.1016/j.cognition.2013.07.015

Millar, J. B., and Ainsworth, W. A. (1972). Identification of synthetic isolated vowels and vowels in H-D context. *Acustica* 27, 278–282.

Moon, K. R., Chung, S. M., Park, H. S., and Kim, H. S. (2012). Materials of acoustic analysis: sustained vowel versus sentence. *J. Voice* 26, 563–565. doi: 10.1016/j.jvoice.2011.09.007

Murry, T., Brown, W. S., and Morris, R. J. (1995). Patterns of fundamental frequency for three types of voice samples. *J. Voice* 9, 282–289. doi: 10.1016/S0892-1997(05)80235-8

Owsianny, M. (2019). Perceptual identification of Polish vowels due to F0 changes. *Arch. Acoust.* 44, 13–26. doi: 10.24425/aoa.2019.126348

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203. doi: 10.3758/s13428-018-01193-y

Reinisch, E., and Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *J. Phonetics* 41, 101–116. doi: 10.1016/j.wocn.2013.01.002

Shankweiler, D., Verbrugge, R. R., and Studdert-Kennedy, M. (1978). Insufficiency of the target for vowel perception. *J. Acoust. Soc. Am.* 63, S4

Strange, W., Edman, T. R., and Jenkins, J. J. (1979). Acoustic and phonological factors in vowel identification. *J. Exp. Child Psychol.* 5, 643–656. doi: 10.1037/0096-1523.5.4.643

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (1976). Consonant environment specifies vowel identity. *J. Acoust. Soc. Am.* 60, 213–224. doi: 10.1121/1.381066

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863

Van Son, R. J. J. H., and Pols, L. C. W. (1999). Perisegmental speech improves consonant and vowel identification. *Speech Commun.* 29, 1–22. doi: 10.1016/S0167-6393(99)00024-2

Weber, A., and Smits, R. (2003). "Consonant and vowel confusion patterns by american english listeners," in *Proceedings of the 15th International Congress of Phonetic Sciences*, eds M. J. Sole, D. Recasens, and J. Romero (Barcelona). Available online at: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_1437.pdf