



## Language-specific interactions of vowel discrimination in noise

Mark Gibson<sup>1</sup>, Marcel Schlechtweg<sup>2</sup>, Beatriz Blecia Falgueras<sup>3</sup>, Judit Ayala Alcalde<sup>1</sup>

<sup>1</sup>Universidad de Navarra

<sup>2</sup>Carl von Ossietzky Universität Oldenburg

<sup>3</sup>Universitat de Girona

mgibson@unav.es

### Abstract

Our pilot vowel discrimination experiment addresses the competition between attentional focus and language exposure in two noise conditions using two groups of participants (L1 English-speakers (L1-EN) taking a perception test in Spanish and L1 Spanish (L1-SP) speakers taking a perception test in Spanish). Our noise conditions included three signal-to-noise ratio (SNR) conditions (-12, -6 and 0 decibels (dB)) and conditions using automatically generated multi-speaker background babble for 1-12 speakers. Our results show notable confusion by both groups in discriminating back round vowels [o] and [u] regardless of L1 or language exposure. We attribute this confusion to the fact that tongue height, detectable through F1, is obfuscated by F3 (lip rounding). In the absence of a visual input by which a listener can discriminate mid and high vowels by a control parameter such as lip aperture (or jaw angle), listeners experience notable difficulty in discerning vowel categories regardless of L1 or exposure to a target L2. Our results are consistent with the notion that both attentional focus and language exposure may provide advantages to vowel discrimination in noise, but compete in bottom-up/top-down protocols.

**Index Terms:** speech recognition in noise, vowel discrimination, linguistic interaction with noise

### 1. Introduction and background

Crucial to Shannon's [1] seminal work on the mathematical basis of communication is the idea that a noise source acts on the signal, which necessarily challenges information flow. As much holds for computationally derived messages as for human speech. An information source encodes a message that a subsequent transmitter converts to a source signal by way of some channel, in our case an acoustic channel. The receiver unpacks the information encoded in the channel(s) in order to decode the message intended by the transmitter. At the same time, noise is intrinsically related to entropy (uncertainty or randomness) in that noise degrades the information content encoded in the message. It is not improbable to speculate that noise, and the entropy engendered by it, in this context may lead to variation in the representation of discrete sound units a speaker/listener codifies to memory and account for some individual variation in continuous speech sounds.

A number of studies have addressed the effects noise has on information flow for human speech. For example, it has been shown that a listener's capacity to discriminate vowels is greatly reduced both by the level of noise in relation to the signal (SNR) and the number of speakers in computer-generated multi-speaker background babble [2]. At the same time, information provided by complementary signals or

modalities, such as a visual input, has also been shown to have a complementary role in interpreting speech in noise [3].

Fewer studies have focused on the effects that language (both the languages that the subjects speak and the language of the test) has on speech processing in noise. Those studies that have addressed language-specific effects for listening in noise have mainly focused on the effects of L2 deficits, and the advantages exposure in a L2 has on discrimination accuracy [4]. However, it is our assertion in the present study that language-specific interactions in speech processing in noise may not be so simplistic. We assert that L1 biases may aid in vowel discrimination given the right circumstances (based on inventory size across the languages involved in the tests), and the ability to suppress multi-speaker background babble may be enhanced when babble is in a second language because speakers can more easily block out noise that is semantically off-limits to them.

With regard to how L1 biases may enhance L2 vowel discrimination in noise, a speaker's native phonological inventory (size) may affect the subsequent weighting of acoustically salient cues when perceiving speech, conditioning the attention of the listener to hear specific contrasts. Acoustic cue-weighting as a function of its reliability in speech discrimination has a long tradition in the perception literature [5], which may be pertinent to speech discrimination in noise. The more reliable a cue is in signaling a phonological contrast, the more perceptual weight that cue receives, and hence directs a speaker's attentional focus to that cue (see [5] for a good review). Accordingly, a speaker from a language with a large vowel inventory (with high inter-categorical overlap and high intra-categorical variation) like English may have an advantage over a speaker with a relatively small vowel inventory (with low inter-categorical overlap and low intra-categorical variation), such as Spanish, in vowel discrimination in noise because their attention is attuned to finer acoustic cues.

In addition to any advantage L1 vowel inventory would provide the L1-EN group, previous studies addressing the language of background babble itself in speech discrimination in noise have reported that participants show a higher capacity to block out background babble from a foreign language than from their L1 [6]. For the current study, this would provide a further advantage for the L1-EN group, since not only would their native cue weights help them focus on the fine phonetic differences of the vowel stimuli, but the language of the background babble makes them more impervious to linguistic interference from informational masking effects.

At the same time, previous studies have shown that discrimination in noise increases as a function of exposure to a language [5,6]. Most of the studies addressing the effects of exposure do so in the context of L2 learning. Nevertheless, the

logical conclusion in this setting is that L1 speakers will generally outperform even advanced L2 speakers given the groups are matched for age, and the L1 speakers reside in a place where their L1 is the dominant language. Hence, in this case, our L1-SP speakers will have an advantage over the L1-EN group, even if the L1-EN speakers have heightened attention to different acoustic cues, due to their increased exposure to Spanish. In this paradigm, we expect competition between attentional focus and exposure.

However, any advantages L1 vowel inventory size and exposure may have in vowel discrimination in noise may largely be restricted to contexts where bottom-up processing is required (where the subject has no knowledge about which vowels they should be preparing to discriminate). In top-down designs (which most forced-choice discrimination tests are), the subject has knowledge about which sounds they should be listening for, and can therefore condition their focus in an anticipatory manner. In such a top-down design, we might expect a native speaker to have an upper hand over a non-native speaker, even if the non-native speaker's vowel inventory is larger because there is a more direct, unfiltered mapping between the continuous and discrete levels of speech.

In the following article we present results for a pilot in which we asked whether L1 vowel inventory (size) and language exposure suppose any special advantage in discriminating non-native vowel categories in noise.

## 2. Speech materials and experiment

### 2.1 Hypotheses

We formulate the following hypotheses:

H1: If L1 inventory size modulates discrimination, we expect to see effects even for speakers with little to no knowledge of Spanish, and differences between the Spanish and English participants, discrimination by L1-EN being generally better than discrimination by L1-SP speakers. This is buttressed by the idea that non-native speakers can more efficiently suppress non-L1 multi-speaker background babble [6].

H2: Past research has shown that exposure to a language enhances discrimination in noise. Hence, if exposure to language enhances vowel discrimination in noise, we expect a trend toward better discrimination as a function of proficiency in Spanish (as proficiency is a function of exposure), and differences between L1-SP and English speakers whereby L1 SP exhibit the highest percentage of correct responses.

### 2.2 Perception experiment and noise conditions

The perception experiment was designed using MatLab. Stimuli were presented randomly to the subject. Two noise conditions were manipulated for the experiment: background babble and the signal-to-noise ratio (SNR henceforth), which is a ratio of signal power to noise power, which is normally expressed in decibels (dB). Background babble was generated randomly according to differing numbers of speakers to 1, 2, 3, 4, 6, 8, 10 and 12. Additionally, SNR was set to three [0, -6, -12] dBs, where [0] represents a 1:1 ratio of signal to noise and [-6, -12] have more noise than signal. A total of 1080 stimuli were presented to the participants.

### 2.3 Stimuli

Stimuli for the test were recorded in a sound-proof recording booth at the Speech Laboratory of the Universidad de Navarra. One female and one male read isolated syllables /da, de, di, do, du/, which appear in both stressed and unstressed positions in Spanish.

### 2.4 Testing procedure

The test was administered at the Speech Laboratory at the Universidad de Navarra, which is a semi sound-proof space with minimal background noise. The participants heard the stimuli using Audio-Technica ATH-R70X studio headphones. Volume was set to a comfortable listening level, which could be changed following an initial trial session programmed into the MatLab-based test. Participants were instructed that they would hear an isolated syllable /da, de, di, do, du/ in different noise conditions and that their task was to listen and identify the syllable they heard. The five options appeared in text boxes on the computer screen and the participants were instructed to select the correct syllable by left-clicking over it, whereupon the next syllable would be presented.

### 2.5 Participants

Five native English (2 female/3 male) and five Spanish (3 female/2 males) speakers between the ages of 18 and 35 were selected for the test. Background information was collected for the speakers by way of self-report surveys soliciting information related to speech and/or auditory problems, attention deficit disorder and general cognitive capacities. Additional lifestyle information was solicited to isolate possible error effects due to deficient sleep, alcohol abuse or fatigue.

## 3. Results

As regards the ability to correctly discriminate vowels in different noise contexts, results were generally consistent across and within language groups. Percentages for the correct discrimination of [a], [e] and [i] were largely commensurate. For both groups, the back rounded vowels [o] and [u] exhibited the highest uncertainty. The following Tables 1 (for native L1-SP) and 2 (for native L1-EN) show percentages for stimuli presented (y-axis) versus actual response (x-axis) by vowel.

Table 1. Percentages for L1-SP of stimuli presented versus response by vowel.

a	76,67	5,37	8,15	5,37	4,44
e	4,63	75,56	9,54	5,46	4,81
i	5,93	8,89	73,15	5,65	6,39
o	6,48	8,80	11,94	55,28	17,50
u	5,93	12,41	17,31	13,43	50,93
	a	e	i	o	u

Table 2. Percentages for L1-EN of stimuli presented versus response by vowel.

a	82,56	5,48	5,02	3,16	3,78
e	6,10	75,39	9,95	3,78	4,78
i	6,40	11,65	68,67	5,56	7,72
o	6,56	9,49	7,72	46,53	29,71
u	6,87	11,19	11,81	10,88	59,26
	a	e	i	o	u

The L1-SP speakers identified [a] correctly in 76.7% percent of the trials versus a mean percentage of 82.6% discrimination by the L1-EN speakers. For [e], both groups showed similar accuracy (L1-EN = 75.4%, L1-SP = 75.6%), though the difference was larger for [i] (L1-EN = 68.7%, L1-SP 73.2%), both groups also exhibited some confusion with the high-back vowel [u] (L1-EN = 11.8%, L1-SP = 17.3%). For the back rounded vowels [o] and [u] (represented in the four bottom right-hand corner cells in Tables 1 and 2), accuracy for both groups was generally no better than chance. For [o], L1-EN identified [o] as [o] in 46.5% of the cases, whereas [o] was correctly discriminated in 55.3% of the trials by L1-SP. In both cases, confusion with [u] was the greatest source of uncertainty. Both groups exhibited similar behavior vis-à-vis discrimination of [u] where L1-EN correctly identified the high back vowel in 59.3% of the trials and L1-SP was capable of discriminating the vowel in 50.9% of the cases. Nevertheless, confusion with [o] was apparent for both groups (notice that [u] was confused with [o] in 17.5% of the trials by L1-SP and in 29.7% of the trials by L1-EN). A representative confusion matrix from each group is offered in Figure 1 (L1-SP) and Figure 2 (L1-EN), where the y-axis shows stimuli that were presented and the x-axis shows the responses by the participants.

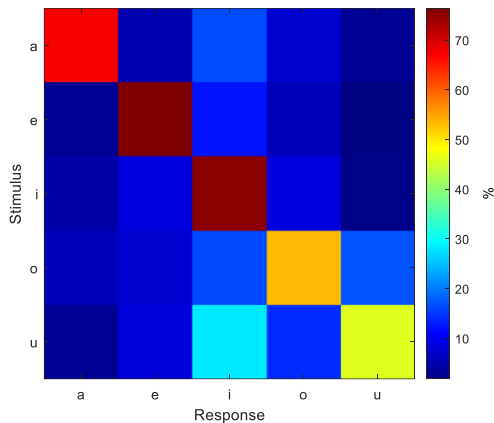


Figure 1. Confusion matrix showing the varying level of confusion by vowel by the L1-SP group. The y-axis shows the stimuli presented while the x-axis represents the actual response by the subject.

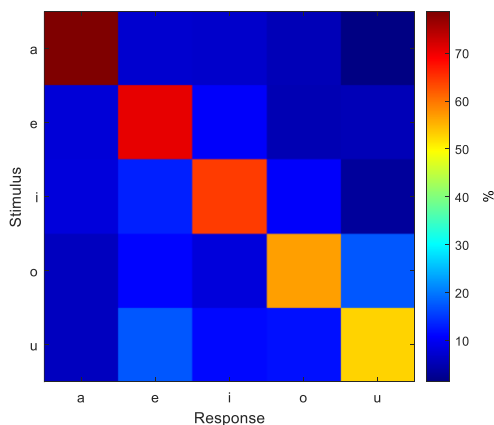


Figure 2. Confusion matrix showing the varying level of confusion by vowel by the L1-EN group. The y-axis shows the stimuli presented while the x-axis represents the actual response by the subject.

stimuli presented while the x-axis represents the actual response by the subject.

As per within-group variation, which is only pertinent to our English group, and predicted to be modulated as a function of L2 proficiency in that it relates to L2 exposure, results are equally non-conclusive. Our participants' levels of Spanish ranged from A1 (initial level - European Common Framework for References of Languages) to C1 (intermediate advanced – European Common Framework). Results for all speakers are commensurate in that all showed very high accuracy for [a], [e] and [i] (within the 70%-80% accuracy range attested for native and non-native listeners) and diminished capacity to discriminate [o] and [u] (within the 50%-60% accuracy range for all listeners). However, there is no progressive trend toward a more accurate discrimination as a function of proficiency (and/or exposure). Our C1 level participant showed an accuracy in discrimination on par with the A1 participants. Notice the commensurate results plotted in the confusion graphs for the C1 participant (Figure 3) versus the A1 participant (Figure 4).

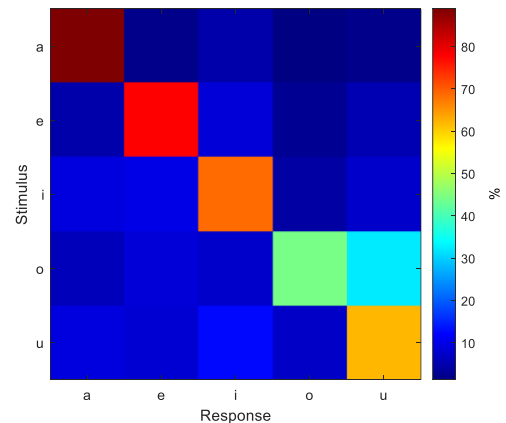


Figure 3. Confusion matrix for (L1-EN) C1 participant showing the varying level of confusion by vowel by the L1-SP group. The y-axis shows the stimuli presented while the x-axis represents the actual response by the subject.

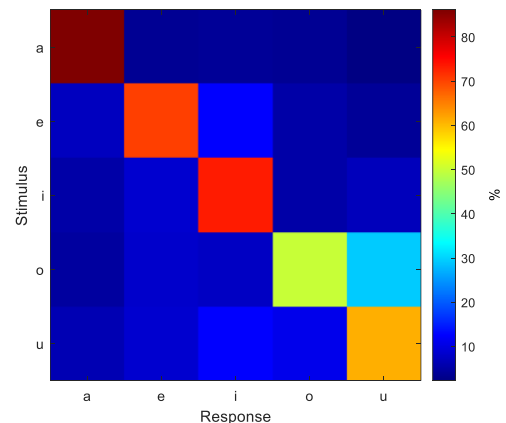


Figure 4. Confusion matrix for (L1-EN) A1 participant showing the varying level of confusion by vowel by the L1-SP group. The y-axis shows the stimuli presented while the x-axis represents the actual response by the subject.

As for results for the different noise conditions, accuracy diminished in relation to the number of background speakers

across groups as well, with no appreciable differences between groups. Accuracy in discrimination with background babble across all number of speakers and [0] SNR was stably at ceiling (between 90%-100%) for most participants, except for two English speakers that exhibited accuracy in the 80%-90% range. Nevertheless, as noise increased in relation to signal, a dramatic drop-off occurred for all speakers at between 4 and 6 speakers, where minima were reached in most cases at around 8-10 speakers. This pattern is true for both the [-6] and [-12] conditions. Representative line graphs showing accuracy of responses (y-axis) in relation to the number of speakers in background babble (x-axis) in the different SNR conditions (see legend) are offered in the following Figure 5 (for L1-EN) and Figure 6 (for L1-SP).

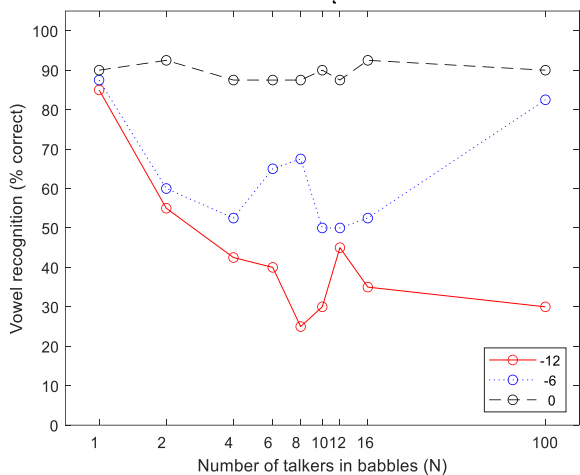


Figure 5. Line graph for L1-EN, A1 (beginner) Spanish participant showing discrimination accuracy (y-axis) in relation to number of speakers in babble (x-axis) and the different SNR conditions (lines, see legend).

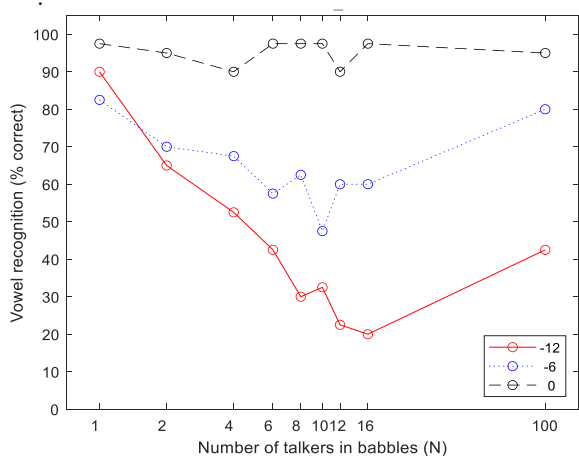


Figure 6. Line graph for L1-SP participant showing discrimination accuracy (y-axis) in relation to number of speakers in babble (x-axis) and the different SNR conditions (lines, see legend).

#### 4. Discussion and conclusion

The results of the pilot discrimination experiments suggest notable confusion in discerning the back vowels [o] and [u], which increases with noise saturation and the number of speakers, but not as a function of L1 or language exposure.

These results lead to a couple of interesting insights that will inform our work in moving forward.

First, the confusion of the back vowels in noise is curious in that it is not normally attested outside of the experimental setting. In fact, [o] is the masculine marker in Spanish, meaning a vast majority of all masculine nouns and adjectives end in [o]. We surmise two scenarios with which to address this confusion.

It may be that tongue height (which is what the listeners are finding difficult to discern), as expressed acoustically as F1, is obfuscated by the rounding of the lips (expressed by F3). In normal speech conditions, a speaker not only has access to the auditory signal, but also has access to visual information that is used for information gain (i.e., reducing entropy or uncertainty). This means that in noisy conditions, the extra visual input may aid the listener in discerning the back vowels. In future studies, we intend to test this conjecture empirically, by providing our participants with visual and audio stimuli, and computationally, using long-short-term memory models trained on acoustic and visual signals to see if the visual input increases accuracy.

Nevertheless, another possibility to explain the general confusion with [o] and [u] may be that lexical effects are skewing discrimination. The stimuli [da], [de] and [di] are all words in Spanish, while [do] and [du] are not. There is a possibility that this imbalance in the stimuli is skewing responses. However, if that were the case, we should expect to see differences across language and proficiency groups since this would only influence native and advanced speakers of Spanish, which we did not find evidence for in our results.

Returning to the idea that L1 vowel inventory size may suppose an advantage for the L1-EN group, our results here, even though non-conclusive, are remarkable in the sense that there was no difference between L1-EN and L1-SP in discriminating Spanish vowels. So, while they did not do better than the L1-SP group, it is noteworthy that they did just as well. Hence, the idea that L1 vowel inventory size provides an advantage in vowel discrimination cannot be discarded. It may very well suppose an advantage that was counterbalanced by competition with language exposure (where L1-SP had an advantage)

On a final note, anecdotal evidence collected from post-test interviews indicate an interesting finding as it pertains to compensation strategies to increase information gain (or decrease in entropy). The highest scorers from both groups (L1 SP = 70% mean accuracy across trials and L1 EN= 69% mean accuracy across trials) reported that they had learned the time that spanned from the screen change for a particular trial to the stimulus onset (about 200 ms) and used this information to prepare for the stimulus and block out background noise. This provides anecdotal evidence for the claim that not only do subjects utilize multiple alternate modalities (like visual input) to recover information in the face of noise, but also use alternate dimensions (temporal) from which to extract information for perceptual recovery.

#### Acknowledgements

This work was financed by a grant [ref. PID2019-105929GA-I00] by the Ministry of Science and Innovation (Spain).

## References

- [1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal* vol. 27, pp. 379–423, 623–656, Jul. 1948.
- [2] C. Liu and D. Kewley-Port, "Formant discrimination in noise for isolated vowels," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3119–3129, Nov. 2004.
- [3] Y. Yuan, Y. Lleo, R. Daniel, A. White, and Y. Oh, "The impact of temporarily coherent visual cues on speech perception in complex auditory environments," *Frontiers in Neuroscience*, vol. 15, pp. 1–7, Jun. 2021.
- [4] M. Li, W. Wang, S. Tao, Q. Dong, J. Guan, and C. Liu, "Mandarin Chinese vowel-plus-tone identification in noise: Effects of language experience," *Hearing Research*, vol. 331, pp. 109–118, Jan. 2016.
- [5] J. C. Toscano and B. McMurray, "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," *Cognitive Science*, vol. 34, no. 3, pp. 434–464, Apr. 2010.
- [6] K. J. Van Engen and A. R. Bradlow, "Sentence recognition in native- and foreign-language multi-talker background noise," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 519–526, Jan. 2007.