Energetic and informational masking effects on Spanish vowel recognition

Mark Gibson[1], Marcel Schlechtweg[2], Judit Ayala[1], Andrea DiCiaccio[3], Xianhui Wang[3], Li Xu[3],
[1]Universidad de Navarra, [2]Carl von Ossietzky Universität Oldenburg, [3]Ohio University,
Corresponding author: mgibson@unav.es

ABSTRACT

We report results of an ongoing study examining the effects of different masking conditions (background babble, or the so-called cocktail party effect, and varying levels of noise saturation in relation to the signal, i.e. signal-to-noise ratio, henceforth SNR) on vowel recognition for two groups: native English speakers taking a test in Spanish, L1-EN, and native Spanish speakers, L1-SP, taking the same Spanish test. Results for the current study are commensurate with our earlier work showing notable confusion when discerning the rounded back vowels [o] and [u] for both masking conditions and language groups (though the L1-EN group showed less release from masking), which we surmise ensues from masking of the first formant (F1) by the third formant (F3), and a lack of visual information referencing jaw angle and/or lip aperture. The biological sex of the listener made no difference on recognition, though both groups showed a bias for the female voice targets (stimuli). Effects of language-specific biases seem likely, but the nature of the interaction of these effects is still unknown.

Keywords: Speech perception in noise, vowel recognition, language effects.

1. INTRODUCTION AND BACKGROUND

Formant frequencies and spectral shape have both been shown to be crucial cues for vowel recognition [1]. It is known that noise affects the access listeners have to these acoustic cues, but it is not altogether clear how different types of noise mask different acoustic dimensions necessary for vowel recognition (see [2] for a good explanation on how noise affects information flow in speech). On one hand, steady-state noise physically interferes with the intended signal blocking access to the acoustic cues transmitted by the sender. This has become known as *energetic masking* (see [3]), which falls out when peripheral neural activity intended to parse the transmitter's signal is overwhelmed by a noise source. A simple example of energetic masking would be where high-volume white noise cuts off access to an intended object of perception. On the other hand, *informational masking* occurs when linguistic, or other higher order dimensions of the noise source create uncertainty, or entropy, with regard to the information flow, which may lead to perceptual errors. A common example of informational masking would be where linguistic/acoustic-phonetic information, say from speakers in the background (i.e. cocktail party effect) conditions the correct interpretation of a stimulus [4].

Previous studies have shown that informational masking is reduced when similarity of the noise source and target stimulus is reduced (when they are more spatially distinct [5, 6]), or when the background speech (i.e. noise) is produced by a speaker who is the opposite biological sex of the speaker producing the stimulus [7, 8]. Additionally, effects of masking have been shown to be higher for intelligible noise (say, where the background babble is clear, and in the perceiver's native language) as compared to non-intelligible noise (say where speech has been modified, or a foreign language is used for background babble). Disentangling the effects of each type of masker is problematic due to difficulties in segregating the informational from energetic mechanisms of speech-on-speech masking.

A dearth of evidence shows that a listener's capacity to discriminate phonological contrasts is greatly reduced by exogenous factors such as the level of noise in relation to the signal (SNR) and the number of speakers in multi-speaker background babble [9-11]. Information provided by complementary signals or modalities, such as a visual input, has also been reported to have a role in parsing speech in noise [12].

Fewer studies have focused on the effects that language (both the native language, henceforth L1, of the perceiver and the language of the targets/maskers), and other endogenous factors related to the perceiver, have on speech processing in noise. However, the results from this line of research are often contradictory, and to a large extent mutually exclusive. Those studies that have addressed language-specific effects for listening in noise have mainly focused on the effects of second language, henceforth L2, deficits, and the advantages exposure in a L2 has on recognition accuracy [13], recognition accuracy increasing as a function of proficiency in the L2. At the same time, previous studies also show that release from masking is higher when background babble is unintelligible (see [5-8]). Thus, it is an intuitively straightforward observation that if unintelligibility of the background masker enhances release from masking, then exposure to that language should not increase recognition accuracy. Would not

exposure to the phonetic/acoustic dimensions of the language of exposure go part and parcel with higher morpho-syntactic/lexical awareness, which is supposedly the source of informational masking?

Thus, it is our working assumption that language-specific interactions on speech recognition in noise may not be so simplistic. We assert that L1 biases may aid in vowel recognition given the right circumstances (based on inventory size across the languages involved in the tests), and the ability to suppress informational maskers (multi-speaker background babble) may be enhanced when the masker is in a second language [5, 14, 15].

With regard to how L1 biases may enhance L2 vowel recognition 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 recognition has a long tradition in the perception literature [16], which may be pertinent to speech recognition 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 [16] 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 recognition 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 recognition in noise have reported that participants show a higher capacity to block out background babble from a foreign language than from their L1 [17]. 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 informational masking.

At the same time, previous studies have shown that recognition in noise increases as a function of exposure to a language [see 13]. The logical conclusion in this setting then 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.

In the following sections we present results for our on-going study in which we examined the effects of exogenous (energetic, SSN/SNR) and endogenous (informational) masking on vowel recognition in noise.

2. EXPERIMENT AND SPEECH MATERIALS

2.1. Hypotheses

We formulate the following hypotheses:

- H1: If L1 inventory size modulates recognition, we expect to see effects even for speakers with little to no knowledge of Spanish, and differences between the Spanish and English participants, recognition by L1-EN being generally better than recognition 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 recognition in noise. Hence, if exposure to language enhances vowel recognition in noise, we expect a trend toward better recognition as a function of proficiency in Spanish (as proficiency is a function of exposure), and differences between L1-SP and L1-EN whereby L1-SP exhibit the highest percentage of correct responses.

2.2. Experiment design and masking conditions

The perception experiment was designed using MATLAB. Two masking conditions were programmed: background babble and the signal-to-noise ratio (SNR henceforth), which is a ratio of signal power to noise power, which is expressed in decibels (dB). Background babble was generated randomly according to differing numbers of speakers: 1, 2, 4, 6, 8, 10, 12, and 16. In addition, a speech-shaped noise (SSN) was included. The SNR was set to three [0, -6, -12] dB levels. A total of 1080 stimuli (i.e., 5 vowels × 4 repetitions × 2 voices x 9 masking types × 3 SNRs) were presented randomly to each participant.

2.3. Target stimuli

Target stimuli 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]. These specific syllables were used because they exist in both English and Spanish, and do not confound possible effects of VOT, as voiceless stops would.

2.4. Testing procedure

The tests were administered at the Speech Laboratories at the Universidad de Navarra and Ohio University, in sound-proof spaces. 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 response options appeared in text boxes on the screen and the participants were instructed to select the correct syllable by left-clicking over it. Reaction times were not registered, though this option may be interesting in future rounds of testing

2.5. Participants

Ten L1-EN (5 female/5 male) and ten L1-SP (5 female/5 males) speakers (18-35 years) were randomly selected for results (we have collected for 35 L1-EN and 35 L1-SP but we only report results here for a subset). For the L1-EN group, three had an initial level of Spanish, four had an intermediate level, and three had an advanced level. Personal data related to speech and/or auditory problems, attention deficit disorder and general cognitive capacities were collected for the participants. Additional lifestyle information was solicited to isolate possible error effects due to deficient sleep, alcohol abuse, or fatigue.

3. RESULTS

Results were by and large commensurate with our earlier pilot in that both groups showed notable confusion in recognizing the rounded back vowels [o] and [u]. It is noteworthy that 0 dB SNR produced nearly perfect vowel recognition but any errors at 0 dB SNR were almost all due to [o,u] confusions. Figures 1 and 2 illustrate this confusion in matrix form, for L1-EN and L1-SP respectively, where the y-axis displays the stimuli presented to the subject and the x-axis shows responses. Numbers in the individual cells are given in percentages. Figures 3 and 4 plot recognition accuracy as a function of SNR and number of background speakers, also for L1-EN and L1-SP respectively.

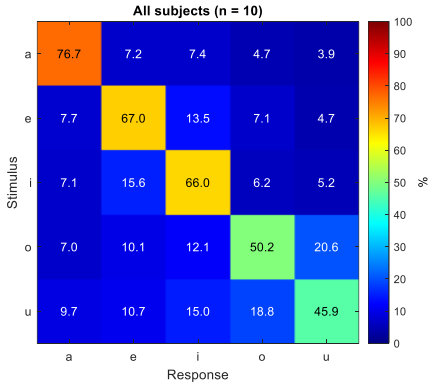

Figure 1. Confusion matrix for L1-EN. Data were pooled across all 10 L1-EN participants and across all masking conditions and SNRs.

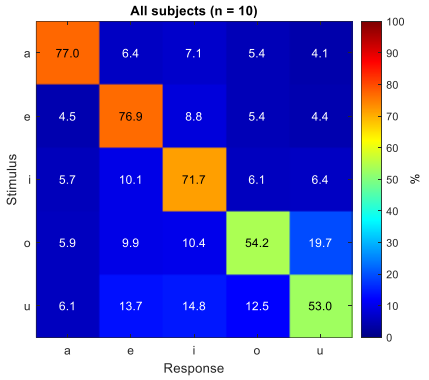

Figure 2. Confusion matrix for L1-SP. Data were pooled across all 10 L1-SP participants and across all masking conditions and SNRs.

As can be seen in the matrices, the L1-EN group showed lower ($p < 0.01$, $F = 21.67$) accuracy for [o] and [u] (50.2% accuracy for [o] and 45.9% for [u]) than the L1-SP group (54.2% accuracy for [o] and 53% for [u]).

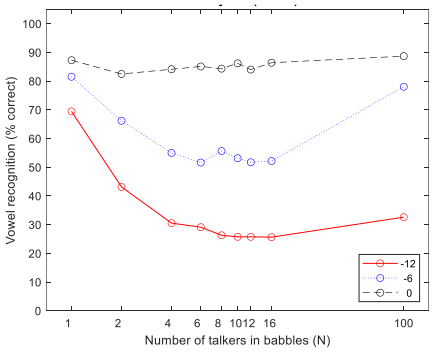

Figure 3. Vowel recognition as a function of SNR and number of background speakers for L1-EN. The results for the SSN condition were plotted arbitrarily at 100 on the abscissa.

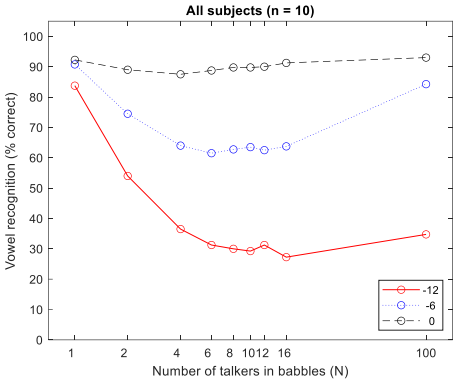

Figure 4. Vowel recognition as a function of SNR and number of background speakers for L1-SP. The results for the SSN condition were plotted arbitrarily at 100 on the abscissa.

Both groups showed relatively high release from masking at 0 dB (accuracy at between 87-92% for all numbers of background speakers), though at higher noise saturations, recognition accuracy drops off notably at around 4 speakers.

As far as responses reported as a function of the biological sex of the voice target, results were commensurate across groups. The L1-SP showed lower accuracy in recognizing vowels produced by the male talker (57.3% accuracy for male targets, 75.76% accuracy for the female voice targets). For the L1-EN group, accuracy in recognizing vowels produced by the male and female talkers was 54.9% and 67.4%, respectively. Overall accuracy was better for the L1-SP group, though significance was only reached for female voice targets. For the male voice targets, effects of group were non-significant ($p>0.05$, $F = 2.13$). There were no differences in recognition between female and male participants.

As regard accuracy scores across different levels of language proficiency, results are similar with our earlier findings in that accuracy did not improve as a function of language level past a certain point. Ceiling was reached by a L1-EN participant with a B2 level of Spanish (intermediate), at a nearly identical accuracy as for our ceiling L1-SP participant. Thus, we do not find support for the argument that L2 exposure conditions recognition, at least for the relatively small vowel inventory of Spanish.

4. CONCLUSIONS AND DISCUSSION

The results of the recognition 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. First, the confusion of the back vowels in noise is curious in that [o] is the masculine marker in Spanish, meaning a vast majority of all masculine nouns and adjectives end in [o], and most native speakers do not confuse [o] and [u] (an anecdotal observation). However, it cannot go unsaid that, across the Romance languages, alternations of [o] and [u] are quite common, even for dialects of Spanish [18], between Portuguese and Spanish [19] and Romanian [20], in addition to other non-Romance languages such as German [21].

As for the specific reason listeners may confuse [o] and [u], we surmise two scenarios. First, the confusion of [o] with [u] (and vice versa) may be due to lexical effects. 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 are inconsistent with the notion that inventory size supposes any advantage in recognizing non-native vowel contrasts. However, results for the L1-SP group are not so much better than for the L1-EN group that we can rule out competing forces that play out while recognizing phonological contrasts in noise

Nevertheless, another possibility to account for the [o]-[u] confusion could 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. This scenario finds support in a set of machine-learning models that were created [22] where it was shown that the addition of a visual variable (maximum lip aperture, obtained through motion capture images of the lips) does aid in classification of the vowels in the face of a noise source.

5. ACKNOWLEDGMENTS

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

7. REFERENCES

[1] Parikh, G., Loizou, P. 2005. The influence of noise on vowel and consonant cues. *Journal of the Acoustical Society of America*, 118, 6, 3874–3888.

[2] Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 623–656,

[3] Pollack, I. 1975. Auditory informational masking. *Journal of the Acoustical Society of America*, 57, S5.

[4] Summers, R., Roberts, B. 2020. Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *The Journal of the Acoustical Society of America* 147(2), 1113-1125.

[5] Freyman, R. L., Balakrishnan, U., Helfer, K.S. 2001. Spatial release from informational masking in speech recognition. *Journal of the Acoustical Society of America*, 109, 2112–2122.

[6] Arbogast, T.L., Mason, C.R.., Kidd, G. 2002. The effect of spatial separation on informational and energetic masking of speech. *Journal of the Acoustical Society of America*, 112, 2086–2098.

[7] Brungart, D.S., Simpson, B.D., Ericson, M.A., Scott, K.R. 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *Journal of the Acoustical Society of America*, 110, 2527–2538.

[8] Kidd, G. Jr., Mason, C.R., Swaminathan., J., Roverud, E., Clayton, K.., Best, V. 2016. Determining the energetic and informational components on speech-on-speech masking. *Journal of the Acoustical Society of America*, 140, 132–144.

[9] Liu, C., Kewley-Port, D. 2004. Formant recognition in noise for isolated vowels. *Journal of the Acoustical Society of America*, 116, 5, 3119–3129.

[10] Wang, X., Xu, L. 2020. Mandarin tone perception in multiple-talker babbles and speech-shaped noise. *Journal of the Acoustical Society of America*, 147, 4, EL307-EL313.

[11] Wang, X., & Xu, L. 2021. [Speech perception in noise: Masking and unmasking]. Journal of Otology, 16, 2, 109-119, DOI: 10.1016/j.joto.2020.12.001.

[12] Yuan, Y., Lleo, Y., Daniel, R., White, A., Oh, Y. 2021. The impact of temporarily coherent visual cues on speech perception in complex auditory environments. *Frontiers in Neuroscience*, 15, 1–7

[13] Li, M., Wang, W., Tao, S., Dong, Q., Guan, J., Liu, C. 2016. Mandarin Chinese vowel-plus-tone identification in noise: Effects of language experience. *Hearing Research*, 331, 109–118.

[14] Calandruccio, L., Dhar, S., Bradlow, A.R. 2010. Speech-on-speech masking with variable access to the linguistic content of the masker speech. *Journal of the Acoustical Society of America*, 128, 860–869.

[15] Brouwer, S., Van Engen, K. J., Calandruccio, L. Bradlow, A. R. 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *Journal of the Acoustical Society of America*, 131, 1449–1464.

[16] Toscano, J.C., McMurray, B. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34, 3, 434–464.

[17] Van Engen, K.J., Bradlow, A.R. 2007. Sentence recognition in native- and foreign-language multi-talker background noise. *Journal of the Acoustical Society of America*, 121, 1, 519–526.

[18] Hualde, J. I. 2014. Los sonidos del español. Cambridge, U.K: Cambridge University Press.

[19] Mateus, M. H. & de Andrade, E. 2000. *The Phonology of Portuguese*. Oxford, U.K: Oxford University Press.

[20] Renwick, M. E. L. 2012. Vowels of Romanian: Historical, Phonological and Phonetic Studies. Phd dissertation, Cornell University.

[21] Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R. & Fleischer, M. 2020. Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties. *Scientific Data* 7, 255.

[22] Gibson, M., González, M. & Schlechtweg, M. (accepted). A machine learning model to assess the integration of visual and auditory cues during speech perception in noise. *Proceedings of Forum Acusticum* 23, 10 Conference of the European Acoustics Associaion.