INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Language effects in noise-induced word misperceptions

Maria Luisa Garcia Lecumberri[1], *Jon Barker*[2], *Ricard Marxer*[2], *Martin Cooke*[3,1]

[1]University of the Basque Country, Vitoria, Spain
[2]University of Sheffield, UK
[3]Ikerbasque (Basque Science Foundation), Bilbao, Spain

garcia.lecumberri@ehu.es

Abstract

Speech misperceptions provide a window into the processes underlying spoken language comprehension. One approach shown to catalyse robust misperceptions is to embed words in noise. However, the use of masking noise makes it difficult to measure the relative contributions of low-level auditory processing and higher-level factors which involve the deployment of linguistic experience. The current study addresses this confound by comparing noise-induced misperceptions in two languages, Spanish and English, which display marked phonological differences in properties such as consonant-vowel ratio, rhythm and syllable structure. An analysis of over 5000 word-level misperceptions generated using a common experimental framework in the two languages reveals some striking similarities: the proportion of confusions generated by three distinct types of masker are almost identical for the two languages, as are the proportions of phonemic and syllabic insertions, deletions and substitutions. The biggest difference is seen for babble noise, which tends to induce relatively complex confusions in English and simpler confusions in Spanish. We speculate that the inflectional morphology of Spanish lends itself to more easily recruit single elements from a babble masker into valid word hypotheses.

Index Terms: speech perception, confusions, noise, Spanish, English

1. Introduction

It is a common occurrence for listeners to mishear an intended message, especially in natural settings characterised by conversational speaking styles and the presence of competing acoustic sources. It has been argued that the resulting 'slips of the ear' have great potential in dissecting the processes that underlie speech perception [1, 2, 3, 4, 5, 6, 7]. The value of such misperceptions is amplified when (i) the intended speech and confusion-inducing context (e.g., background noise) are recorded for further analysis; (ii) a sufficient number of listeners report the same confusion; and (iii) a large corpus of confusions is available. Such corpora can then be used to evaluate microscopic models of speech perception i.e., computational models that aim to predict the confusion that results from each individual stimulus e.g., [8, 9].

Previous corpora of misperceptions have either been collected on the basis of anecdotal reports [3, 7], or induced in the laboratory by simulating challenging listening conditions such as time-compressed speech [4], faint speech [10], or presentation in noise [11, 12]. Although listener reports provide the most genuine real-world confusions, they are usually single-person confusions and lack the inducing context, precluding further analysis. Lab-induced confusions permit subsequent manipulation of speech and/or masker signals to further explore the factors causing the confusion [11]. However, one potential confound in examining lab-based confusions is the degree to which the confusion is influenced by expectations based on prior experience of the structures of the target language, or by the acoustic manipulations employed to catalyse the generation of confusions. In the case of confusions elicited in noise, we are interested in finding out whether the pattern of misperceptions is dominated by the masker, or by the target language.

Two large-scale word confusion corpora have been collected recently for Peninsular Spanish [13] and British English [14], both using a similar noise-induced protocol. The availability of corpora for two different languages provides an opportunity to examine further the possible contributions of acoustic and linguistic factors in the presence of misperceptions. English and Spanish have clear phonological differences which might be expected to influence the form of misperceptions. In terms of phonemic inventory, Spanish has just five simple vowels whereas Standard British English has 12. English has a preference for closed rather than open syllables (60% vs. 40% respectively; [15]) whereas Spanish, like other Romance languages, has more open syllables (approximately 70% of all syllables; [16, 15]). English has a higher consonant:vowel (C:V) ratio [15] which is a product of both its preference for closed syllables and its higher tolerance for consonant clusters both in onset and in coda position.

An additional key difference between the two languages is the English tendency to stress-timed rhythm, which encourages weakening of unstressed syllables. As a consequence, in unstressed syllables, weak central vowels like schwa predominate, and there are even vowel-less syllables through the presence of syllabic consonants e.g., "button": [bʌtən] ↦ [bʌtn̩]. Spanish unstressed vowels, however, differ little in quality and quantity from their stressed vowel counterparts [17]. Thus, English unstressed syllables are considerably weaker than Spanish ones and, we hypothesise, more vulnerable to masking.

The differences between the two languages lead us to predict that English is more likely to suffer consonant coda substitutions in noise due to the larger set of choices in that position, and that Spanish will have more consonant deletions due to its preference for open syllables. We also predict consonant deletions and insertions as well as vowel substitutions stemming from the morphological flexibility of Spanish. A further hypothesis is that masking of English unstressed syllables will surface as syllable deletions or reconstructions. Apart from these inter-language differences, it is also to be expected that some misperception phenomena, such as the robustness of stressed syllables and stress location will emerge as commonalities between the two languages in the face of masking noise.

http://dx.doi.org/10.21437/Interspeech.2016-330

2. Corpora

The Spanish Confusion Corpus [13] is a collection of 3235 robust word misperceptions induced in the presence of five types of masker. Since word misperceptions that are consistently reported (i.e., by several listeners) are not frequent, an adaptive stimulus presentation protocol was used in which word-plus-masker tokens were generated dynamically and identified by listeners; based on simple heuristics, only those tokens deemed likely to lead to confusions were preserved and presented to other listeners on subsequent trials. Only those stimuli with at least six listeners agreeing on the misperception were retained for the corpus, with an upper limit of 15 presentations. The English Confusions Corpus [14] used an identical protocol, generating 3207 confusions.

The current study made use of the entire English Confusions Corpus and a comparable subset of the Spanish Confusions Corpus. The English corpus used only three of the five maskers employed in the Spanish corpus, namely four-talker babble (BAB4), speech-shaped noise modulated by the temporal envelope of three-talker babble (BMN3), and speech-shaped noise (SSN). For Spanish, we use the subset of 1943 confusions induced by these maskers.

Table 1 compares the target words – i.e., the words presented in noise to listeners – of the Spanish (SP) and English (EN) confusions corpora with respect to a range of phonemic, syllabic and other metrics. A comparison of the data in Table 1 shows that the corpora reflect the main characteristics of each language described in the Introduction. Overall word duration is similar in the two languages, but we encounter differences in the number of syllables per word. In Spanish we see a bias towards bisyllables, whereas in English there is an equal presence of one and two syllable words. Although the number of syllables was a selection criterion in the two corpora [13, 14], the outcome reflects the tendencies of the two languages [18].

Table 1: *Corpus target word statistics.*

	SP	EN
Target words	1943	3207
Word duration (ms)	635	685
Monophthongs	5 (38%)	12 (26%)
Diphthongs	11 (4%)	8 (9%)
Consonants	22* (58%)	24 (65%)
C:V ratio	1.4	1.9
Phonemes per word	5.3	4.3
Syllables per word	2.3	1.5
Monosyllables	1.9 %	49.5 %
Bisyllables	71.7 %	50.0 %
Trisyllables	26.4 %	0.5 %
Open syllables	61.0 %	31.8 %
Open final syllables	24.0 %	13.7 %
Stressed syllables	44.5 %	66.2 %

* includes 3 allophones (see text)

Table 2 provides a breakdown of the Spanish and English phoneme inventories as employed in the current study. For better comparability with English, we treat Spanish vowel pairs within the same syllable as diphthongs. We follow common practice in Spanish corpora [19, 20] by including the approximant allophones /β, ð, ɣ/, which are systematically distinct from plosive realisations.

Table 2: *Phonemic content of the corpus.*

	SP	EN
Vowel	a, e, i, o, u	ɑ, æ, ʌ, ɔ, ə, e ɜ, ɪ, i, o, ʊ, u
Diphthong	ia, ie, io, ua, ue, ui uo, ai, au, ei, oi	aɪ, aʊ, ɪə, eɪ, oʊ, ɔɪ ɛə, ʊə
Plosive	p, t, k, b, d, g	p, t, k, b, d, g
Nasal	m, n, ɲ	m, n, ŋ
Fric/affric	f, s, θ, tʃ, x, ɟ, ð	f, s, θ, ʃ, v, z, ð, ʒ, tʃ, dʒ, h
Liq/approx	l, r, ɾ, β̞, ð̞, ɣ̞	l, ɹ, j, w

3. Analysis

In the following, statistical comparisons are based on differences of proportions between Spanish and English, with Bonferroni corrections applied for multiple comparisons.

3.1. Maskers

Table 3 shows that the proportion of confusions for each masker type in English and Spanish are very similar; indeed, they are statistically-equivalent [$p \geq .54$]. In both languages significantly more confusions are generated in the presence of the BMN3 masker than for SSN or BAB4 [$p < .001$].

Table 3: *Confusions generated by each masker (%).*

	SSN	BMN3	BAB4
SP	32	38	31
EN	33	37	29

3.2. Confusion classes

We extend a classification scheme for slips of the ear introduced by [2] based on the number of segment differences between the target and confusion. Target words and confusions were aligned using dynamic programming string alignment, with a constraint to match consonants to consonants and vowels to vowels, and with deletion, substitution and insertion costs of 7, 10, 7 respectively. *Single* cases involve a single insertion, deletion or substitution, while *Duals* require two such changes. The remainder is split into three further categories based on the number of positions in the alignment with matching segments: *Reformulations* are cases where the target and confusion match in two or more locations; *Complex* cases are those where they match in a single segment; *Eccentric* cases have no matching segments. While a target/confusion pair might enter into more than one of these categories (for example, a three-segment word with one error could be regarded as a Single or Reformulated case), the Reformulation, Complex and Eccentric classes are applied after excluding the Single and Dual cases, so in practice there is no ambiguity as to the class of any confusion. Table 4 provides some examples from the corpora for both languages.

Figure 1 shows the distribution of confusion classes in the two corpora. Approximately 40% of confusions involve one change, 27% two changes, and 30-35% more than two changes. Spanish has a higher proportion of Single and Reformulation cases [$p < .001$]; for English the proportions of Complex and Eccentric cases are higher [$p < .001$].

Different maskers might be expected to lead to differing

Table 4: *Example confusions by class.*

	SP	EN
Single	preciosa ↦precioso	pleasant ↦peasant
	vistas ↦pistas	toll ↦tall
	cuánto ↦cuantos	grows ↦growth
Dual	noté ↦maté	parting ↦party
	cerebral ↦celebrar	junk ↦jump
	fijar ↦dejar	statesman ↦statement
Reformulation	tarta ↦montar	doctrine ↦doctor
	sorda ↦sol	trial ↦final
	roba ↦droga	stopping ↦tricky
Complex	locura ↦leche	winter ↦hatred
	antes ↦alcohol	likely ↦white
	sabrá ↦choca	shelf ↦shout
Eccentric	guardan ↦pozo	lounge ↦adapt
	iré ↦ropa	pool ↦design
	creó ↦duchas	modern ↦suggest

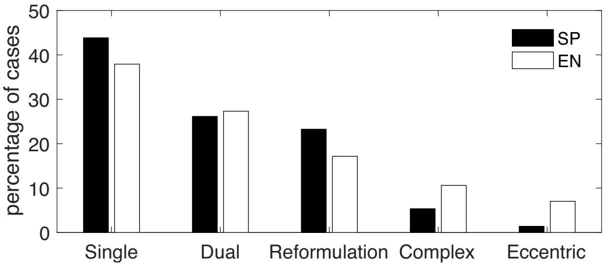

Figure 1: *Distribution of confusion classes.*

proportions of each confusion class. Figure 2 compares the two languages with respect to masker type and confusion class. The two languages are highly correlated [$r = 0.9, p < .001$]. Nevertheless, there is a clear language difference for the BAB4 masker; indeed, excluding BAB4 leads to a correlation of 0.99 [$p < .001$]. For English, BAB4 produces similar proportions of confusions for each confusion class, while for Spanish the proportion falls with confusion complexity.

3.3. Single change cases

Figure 3 (top) shows how the Single change cases pattern into substitutions, insertions and deletions. The proportions of error types are statistically-equivalent for English and Spanish [$p \geq .49$]. The lower part of the figure provides a more detailed breakdown. English has proportionally more errors in initial position, while Spanish has more in final position [both $p < .001$]. Consonant errors are more common than vowel errors in both languages [$p < .001$]. For English, 88% of such changes involve consonants, while for Spanish the comparable figure is 72% [$p < .001$]. We argue in the Discussion that the inflectional morphology of Spanish and the stress and vowel weakening patterns of English go a long way to explain these disparities.

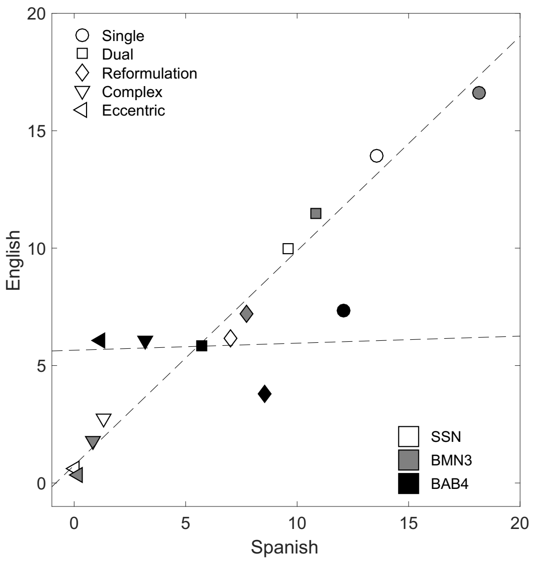

Figure 2: *Percentage of tokens split by confusion class and masker type. Separate linear fits for (SSN, BMN3) and BAB4 are also shown.*

	sub	ins	del
SP	58	14	28
EN	60	12	28

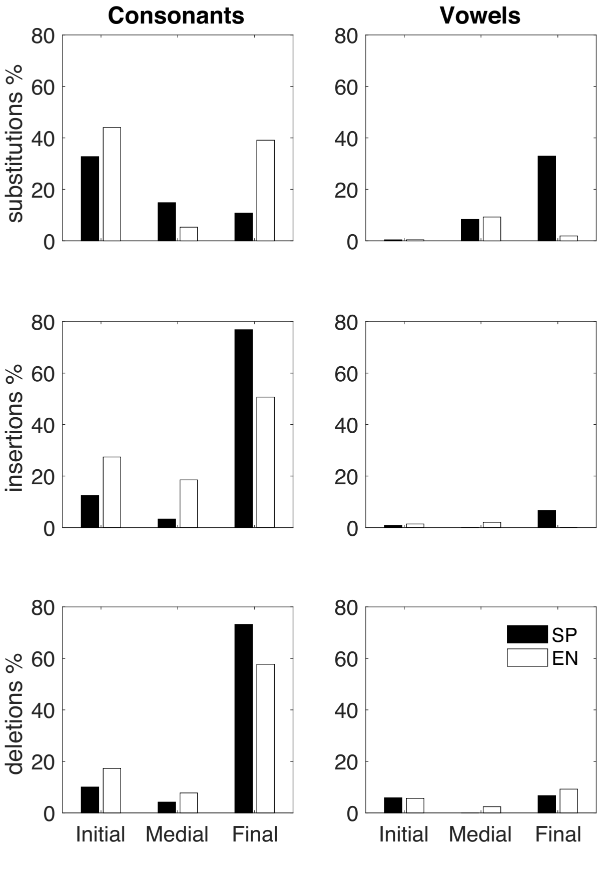

Figure 3: *Locus and segment type for Single cases.*

3.4. Syllabic processes

Figure 4 (top) shows the percentage of each corpus as a function of the difference in syllable counts between target and confusion for the two languages. Around 22% of all confusions involve syllable deletion or insertion, with a ratio of about 2:1 in favour of deletion. The distribution is remarkably similar for the two languages [$p \geq 0.57$]. Syllable insertions and deletions nearly always involve just a single syllable: only 18 of the 5150 confusions across both languages have more than one syllable insertion or deletion. The breakdown across masker of syllable substitutions, insertions and deletions is also depicted in Figure 4. The pattern is again similar for the two languages [all $p \geq .21$ except deletions for BAB4, where $p < .01$]. Both languages show a very clear preference for syllable insertions in the BAB4 masker.

	sub	ins	del
SP	78	8	15
EN	78	7	15

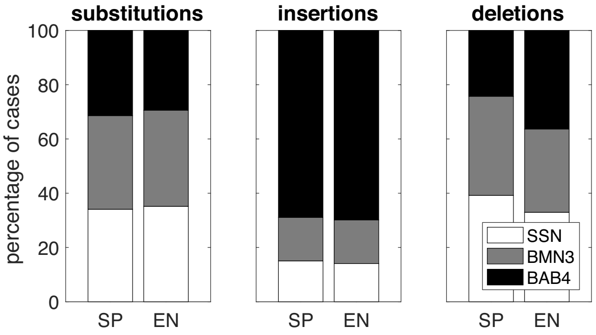

Figure 4: *Top: percentage of syllable change types. Bottom: breakdown across maskers.*

4. Discussion

The current study demonstrates a clear similarity in the effect of masking noise on the pattern of confusions in Spanish and English, evidenced in a number of measures. The distribution of confusions over the three masker types is statistically-equivalent, as are the rates of substitutions, insertions and deletions at both the segmental and syllabic level. For some factors these similarities are also present at a more fine-grained level. The effect of masker types on syllable changes shows clear parallels across the two languages, with babble noise generating near-identical proportions of insertions (Figure 4).

Similarity at the syllabic level is particularly intriguing, given the differences in percentages of syllable counts, open syllables and stressed versus unstressed syllables shown in Table 1. Assuming that stressed vowels are generally more resistant to masking [2], at first sight the higher proportion of unstressed syllables in Spanish might be expected to lead to a difference in the rate of deletion across the two languages. However, the tendency for Spanish unstressed syllables to be less weak than their English counterparts might to some extent compensate for their higher frequency of occurrence.

There are some cross-language differences in the effect of masking, particularly in the location at which single segment changes tend to occur in the word, and their consonantal or vocalic dominance (Figure 3). The existence of more consonant errors in English might reflect differences in C:V ratios. The higher number of consonant coda substitutions in English may be due to the fact that English has many more choices for coda consonants. The presence in Spanish of more Single case insertions and deletions in coda position seems likely to be due to morphology. As a highly inflected language, Spanish possesses base forms with numerous morphological variants, which typically differ segmentally via the presence or absence of a single phoneme e.g., “cantas, canta, cantan, cantar, cantad, cantase” are all forms of the verb “to sing”. Thus, a single segment alteration occurring word-finally is quite likely to result in a legitimate word. For the same reason it is understandable that there are more vowel substitutions in coda in Spanish e.g., “canta, canto, cante” are all possible words. English coda vowels may be less likely to change if they are in a monosyllable and thus stressed. Further, since English has a tendency for initial stress, bisyllabic words have few choices for the final vowel due to weakening of unstressed syllables.

A less predictable difference is seen in the distribution of confusion classes, with English showing proportionally more Complex and Eccentric cases, and fewer instances of the Single and Reformulation categories (Figure 1). However, the difference is almost entirely due to the babble masker (Figure 2): in English, babble induces similar rates of each confusion class, while for Spanish in general it favours the generation of simpler confusion types. Again, the inflectional nature of Spanish may be partly responsible for this, since single segment changes often lead to viable words in which the base stressed syllable is preserved. In particular, vowel insertion by reconstruction or recruitment from the background may be fostered by the fact that the five Spanish vowels can occur practically anywhere and therefore they are easier to recruit and fit into a new genuine word. Spanish has more word-final vowel insertions, which also result in syllable insertions. English, on the other hand, has a more restricted vowel distribution: not all vowels can appear in open syllables or followed by all consonants and there is a limited set of vowels that may appear in unstressed syllables. It also has fewer morphological variants. Consequently, insertion of material recruited from babble noise may require more top-down processing to produce a viable word. In addition, since that are proportionally more monosyllabic word forms in English, any noise sufficiently intense to mask the (single) vowel nucleus will inevitably require recruitment of an alternative candidate from the background, leading to more complex and eccentric cases. Spanish, having far fewer monosyllabic words, appears less likely to suffer from a total loss of syllable nuclei.

5. Conclusions

Robust word misperceptions in English and Spanish produced in the presence of masking noise show remarkably similar proportions of phonemic and syllabic changes in spite of clear differences in phonology between the two languages. Some language-based differences were observed, mainly in the locus of single segment alterations and in the effect of a babble masker on the complexity of listener confusions. Both the inflectional morphology of Spanish and the dissimilar configuration of unstressed syllables in the two languages appear to influence the differential confusion patterns.

Acknowledgements: This work was supported by the Basque Government Consolidado grant to LASLAB, the Spanish Ministry MICIN Diacex project, and the EU Marie Curie INSPIRE Network. We thank Mate Toth for his contribution to the Spanish Confusions Corpus.

6. References

[1] S. Garnes and Z. S. Bond, “Slips of the ear: Errors in perception of casual speech,” in *In Papers from the Eleventh Regional Meeting, Chicago Linguistic Society*, 1975.

[2] ——, “A slip of the ear? a snip of the ear? a slip of the year?” in *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand*, V. A. Fromkin, Ed. New York: New York: Academic Press, 1980.

[3] Z. Bond, “Morphological errors in casual conversation,” *Brain and Language*, vol. 68, no. 12, pp. 144–150, 1999.

[4] M. S. Vitevich, “Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear,” *Language and Speech*, vol. 45, pp. 407–434, 2002.

[5] A. Cutler and C. Henton, “There’s many a slip ’twixt the cup and the lip,” in *On Speech and Language: Studies for Sieb G. Nooteboom*, H. Quené and V. van Heuven, Eds. Netherlands Graduate School of Linguistics, 2004.

[6] Z. Bond, “Slips of the ear,” in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Oxford: Blackwell, 2005, pp. 290–310.

[7] K. Tang and A. Nevins, “Naturalistic speech misperception – a computational corpus-based study,” in *Proceedings of the 43rd Meeting of the North East Linguistic Society*, 2013.

[8] M. Cooke, “A glimpsing model of speech perception in noise,” *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[9] T. Jurgens and T. Brand, “Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model,” *J. Acoust. Soc. Am.*, vol. 126, no. 5, pp. 2635–2648, 2009.

[10] A. Cutler and S. Butterfield, “Rhythmic cues to speech segmentation: evidence from juncture misperception,” *Journal of Memory and Language*, vol. 31, pp. 218–236, 1992.

[11] M. Cooke, “Discovering consistent word confusions in noise,” in *Proc. Interspeech*, 2009, pp. 1887–1890.

[12] M. L. Garcia Lecumberri, A. M. Toth, Y. Tang, and M. Cooke, “Elicitation and analysis of a corpus of robust noise-induced word misperceptions in Spanish,” in *Proc. Interspeech*, 2013, pp. 2807–2811.

[13] M. A. Tóth, M. L. García Lecumberri, Y. Tang, and M. Cooke, “A corpus of noise-induced word misperceptions for Spanish,” *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. EL184–EL189, 2015.

[14] R. Marxer, J. Barker, M. Cooke, and M. L. García Lecumberri, “A corpus of noise-induced word misperceptions for English,” *J. Acoust. Soc. Am. EL*, submitted.

[15] P. Delattre and C. Olsen, “Syllabic features and phonic impression in English, German, French and Spanish,” *Lingua*, vol. 22, pp. 160–175, 1969.

[16] A. Moreno Sandoval, D. Torre, N. Curto, and R. Torre, “Inventario de frecuencias fonémicas y silábicas del castellano espontáneo y escrito,” in *IV Jornadas en Tecnología del Habla*, 2006, pp. 77–81.

[17] A. Quilis and M. Esgueva, “Frecuencia de fonemas en el español hablado,” *Lingüística Española Actual*, vol. 2, pp. 1–25, 1980.

[18] A. Quilis, *Tratado de fonología y fonética españolas*. Madrid: Gredos, 1993.

[19] A. Duchon, M. Perea, N. Sebastián-Gallés, A. Martí, and M. Carreiras, “EsPal: One-stop shopping for Spanish word properties,” *Behavior Research Methods*, vol. 45, no. 4, pp. 1246–1258, 2013.

[20] E. Martínez-Celdrán, A. Fernández, and J. Carrera, “Castilian Spanish: Illustrations of the IPA,” *Journal of the International Phonetic Association*, vol. 33, no. 2, pp. 255–259, 2003.