A corpus of noise-induced word misperceptions for Spanish

Máté Attila Tóth and María Luisa García Lecumberri
Language and Speech Laboratory, Universidad del País Vasco, 01006 Vitoria, Spain
a.m.toth@laslab.org, garcia.lecumberri@ehu.es

Yan Tang
School of Computing, Science and Engineering, University of Salford, Salford, United Kingdom
y.tang@salford.ac.uk

Martin Cooke
Ikerbasque (Basque Science Foundation), Bilbao, Spain
m.cooke@ikerbasque.org

Abstract: Word misperceptions are valuable in designing and evaluating detailed computational models of speech perception, especially when a number of listeners agree on the misperceived word. The current paper describes the elicitation of a corpus of Spanish word misperceptions induced by different types of noise. Stimuli were presented using an adaptive procedure designed to promote the rapid discovery of misperceptions. The final corpus contains 3235 misperceptions along with speech and masker waveforms, permitting further experimental and modeling studies into the origin of each misperception. The corpus is available online as an open resource.

[DOS]
Date Received: August 21, 2014 **Date Accepted:** December 22, 2014

1. Introduction

Progress in objective speech intelligibility models has led to increasingly accurate predictions of mean word identification scores under a variety of masker conditions (e.g., Christiansen *et al.,* 2010; Taal *et al.,* 2010). However, such "macroscopic" models are not designed to predict individual word confusions and consequently say little about the detail of how listeners process speech in noise. Recently, a number of "microscopic" models have been proposed (e.g., Holube and Kollmeier, 1996; Cooke, 2006; Jürgens and Brand, 2009) that output word or phoneme hypotheses and hence can be evaluated with respect to behavioral data at a fine-grained level. What is lacking is a significant collection of word misperceptions in noise to support the development and refinement of microscopic models of speech perception. The current article describes the elicitation of such a corpus for Spanish.

The value of speech error corpora has long been recognized in speech perception research. Garnes and Bond (1980) collected more than 900 spontaneously occurring speech misperceptions in everyday conversations, while Tang and Nevins (2013) gathered more than 3000 misperceptions in similar conditions. Misperceptions have been elicited in the laboratory too using fast speech (Vitevich, 2002) or faint speech (Cutler and Butterfield, 1992). The current study elicited misperceptions in the presence of noise, with a focus both on consistency—misperceptions have to be reported by several listeners—and completeness—the noise-inducing signal is recorded for future studies. A pilot study in English (Cooke, 2009) demonstrated the feasibility of finding consistent noise-induced misperceptions and estimated their rate of occurrence. In the present study, the discovery rate was increased using an adaptive method for early pruning of speech-in-noise tokens that were deemed unlikely to result in a misperception. A preliminary report on a smaller version of the corpus was given in Garcia Lecumberri *et al.* (2013).

2. Corpus elicitation

2.1 Speech material

Four talkers, two male and two female, were recorded reading a word list containing 3968 of the most frequent Spanish words of up to three syllables. Talkers were trained to avoid list intonation. Recordings took place in a sound-attenuated studio using an AKG 4500 microphone and RME Fireface 800 analog-to-digital interface. The resulting recordings were manually segmented and downsampled to 16 kHz. Some 15 753 items remained after removal of 119 mispronounced or noise-contaminated items.

2.2 Maskers

With the goal of promoting word misperceptions due to both energetic and informational masking, five maskers (Table 1) were generated using the recorded speech material. One masker was stationary (SSN) while the rest differed in their depth of temporal modulation. In two cases (BAB4 and BAB8), maskers were composed of natural speech material, while for BMN1 and BMN3 the envelope of competing speech and three-talker babble was used to modulate a speech-shaped noise carrier. Speech-plus-noise stimuli were presented to listeners at SNRs within the masker-specific ranges shown in the table. These values were chosen based on previous work (Cooke, 2009) and pilot tests as likely to elicit consistent misperceptions.

2.3 Participants

A total of 172 listeners provided responses to words in noise. Listeners were native monolingual Spanish or bilingual Spanish-Basque speakers studying at the University of the Basque Country in Vitoria, Spain (mean age 22 yr, s.d. 4.8). Apart from three listeners from Spanish-speaking countries in South America, all participants were born in the Basque Country. Listeners gave written consent for anonymous use of their responses and were paid for their participation.

2.4 Adaptive stimulus pruning

Corpus elicitation made use of a heuristics-based pruning technique designed to remove stimuli deemed unlikely to result in consistent misperceptions. The number of identical misperceptions for each stimulus was monitored online and a decision made following presentation as to whether to remove the stimulus from further consideration. Stimuli were pruned if any of the following conditions held: (a) They were heard correctly by two consecutive listeners prior to the first eight presentations or by three consecutive listeners after the first eight presentations; (b) the first four responses were different; (c) the ratio of correct to incorrect responses exceeded 1.5; or (d) the token had been presented at least 15 times. The first three criteria aimed to remove, as rapidly as possible, stimuli that were unlikely to be misperceived in a consistent fashion. The final criterion is a simple stopping condition: Any stimulus surviving to this point

Table 1. Maskers used in the experiment. The column headed "Speech?" indicates those maskers containing natural speech signals.

	Masker	Speech?	Stationary?	SNR range (dB)
SSN	Speech shaped noise	×	×	−7 to −4
BMN1	Speech modulated noise	×	✓	−13 to −7
BMN3	Three-talker babble modulated noise	×	✓	−8 to −3
BAB4	Four-talker natural babble	✓	✓	−3 to +1
BAB8	Eight-talker natural babble	✓	✓	−4 to +1

was marked as a potentially interesting misperception. For each pruned stimulus, a replacement was generated online using the same SNR, masker type, and speaker as the pruned stimulus. The replacement word and masker fragment were chosen at random. Note that while the pruning procedure might inadvertently remove a potential misperception some of the time, these losses are outweighed by the efficiency gains in discovering misperceptions. Indeed, subsequent analysis indicated that adaptive pruning enabled a near-tripling of the rate of discovery of consistent misperceptions.

2.5 Procedure

Over the course of two non-contiguous sessions lasting approximately 1 h each, listeners identified blocks of 100 words in each of 20 conditions made up from all combinations of the four talkers and five maskers. Within each block, words were mixed with noise in a descending sequence of SNRs. For the first five stimuli, the SNR decreased linearly from +30 dB to the upper SNR value shown in Table 1 to accustom the listener to the target talker and masker type. For the remaining 95 stimuli, the SNR was set randomly in the ranges corresponding to the masker type and presented in decreasing order of SNR, the goal being to explore a range of SNRs without large jumps between stimuli. As a consequence of pruning, the sequence of stimuli presented to each participant was assembled online and hence differed from listener to listener. The masker led and lagged the speech by 200 ms, and 20 ms linear ramps were applied to the mixed token prior to presentation. Participants heard stimuli through Sennheiser HD 380 pro headphones at 75 ± 1.5 dB(A) while seated in a sound-attenuated booth. Listeners were instructed to type a single word in response to each stimulus although on a small proportion of occasions listeners typed more than one word.

2.6 Postprocessing

Listeners' responses were subject to a number of post-processing steps designed to maximize the number of useful misperceptions. First, because on many occasions participants omitted stress marks or the diacritic in ñ, such words were identified and replaced whenever unambiguously possible [e.g., "máximo" (maximum) for "maximo," "baño" (bath) for "bano"]. Second, words with orthographic errors (e.g., "abestruz") but which resulted in a phoneme sequence identical to a unique existing word [e.g., "avestruz" (ostrich)] were corrected. Finally, homophones [e.g., "hola" (hello) and "ola" (wave)] were replaced by the most frequent form. These steps were performed automatically using a combination of the GNU spell checker *aspell,* a rule-based Spanish semi-phonemic transcriber *HAPLO,* and the *CREA* Spanish word frequency list published by the Spanish Royal Language Academy (REAL, 2008). Semi-phonemic (i.e., intermediate between broad and narrow) transcriptions were used because Spanish plosive realizations differ in a largely systematic manner according to the phonetic context. Further, words contrasting in the lateral versus central approximants ("ʎ," "j") were treated as homophones because most Spanish speakers do not distinguish them.

To complement semi-phonemic transcriptions, syllable boundaries and stress were obtained using *TIP* (Hernández-Figueroa *et al.,* 2012), a Spanish word syllabification tool based on morphological analysis. Syllable boundaries are marked with a period, while "!" denotes the start of the stressed syllable. In addition, the phoneme alignment between target and misperception was computed using dynamic programming with a constraint that enforced alignment of consonants to consonants and vowels to vowels. Alignment costs for insertions and deletions were set to 7, while substitutions had a cost of 10.

3. Corpus description

Some 308 152 responses were collected during the elicitation process. In all, 53 039 individual speech-in-noise tokens were generated, of which 9288 survived pruning and were heard by at least 15 listeners. Of these, a minimal level of listener agreement of

Table 2. Counts of consistent misperceptions per test condition.

		Masker						
Speaker	Gender	SSN	BMN1	BMN3	BAB4	BAB8	Totals	Percentage
S1	M	177	224	223	128	119	871	26.92
S2	F	143	201	195	196	153	888	27.45
S3	F	122	171	162	137	64	656	20.28
S4	M	187	201	171	150	111	820	25.35
Totals		629	797	751	611	447	3235	100.00
Percentage		19.44	24.64	23.21	18.89	13.82	100.00	

six listeners was applied to produce the final corpus. Some 3235 misperceptions meet this criterion and jointly make up the corpus of consistent misperceptions.

Table 2 summarizes the number of misperceptions obtained in each test condition. All speaker/masker combinations contributed substantial numbers of misperceptions to the corpus with somewhat more resulting from the two babble modulated noise conditions.

Figure 1 visualizes counts of misperceptions as a function of phoneme alignment distance and consistency (expressed as the proportion of listeners reporting the same misperception), plotted separately for each of the five masker types. This plot suggests that while simple misalignments are frequent—45% of misperceptions involve the insertion, deletion, or substitution of a single phoneme—more complex confusions are also present. Highly consistent complex misperceptions are less common. However, for the two natural babble maskers, and especially for the four-talker case, such misperceptions exist, possibly due to the recruitment of phonetic material from speech-based maskers.

The corpus contains a substantial number of near-bimodal cases: On 189 occasions, the number of listeners reporting the majority misperception differs from the second most frequent response (which might be the target) by two, 110 differ by one, while in 26 cases they are equally consistent.

4. Open corpus resource

The released corpus consists of four components: (a) A tabular representation of misperceptions, each row of which has the structure shown in Table 3; (b) speech, masker, and speech-plus-masker waveforms for each misperception; (c) the masker waveforms from which individual masker fragments were chosen; and (d) for the maskers composed of natural speech (i.e., BAB4 and BAB8), the transcriptions indicating which words were present, and when. The corpus is released under the Creative Commons CC BY license.

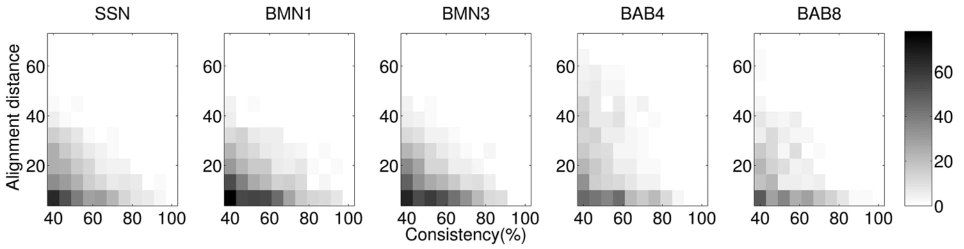

Fig. 1. Counts of misperceptions as a function of consistency and phoneme alignment distance for each masker.

Table 3. Example corpus entry for the word "baño" [bath] in four-talker babble noise, misperceived by 10 of 15 listeners as "España" [Spain].

Field	Description	Example
ID	Integer used to identify the speech waveform corresponding to the entry	35877
Length	Speech signal length in samples	8003
Masker	One of [SSN, BMN1, BMN3, BAB4, BAB8]	BAB4
Onset	Starting location of the masker fragment within the masker waveform; along with the Length field this can be used to extract the masker waveform	562883
SNR	Signal-to-noise ratio in dB	−0.545
Speaker	One of [s1, s2, s3, s4]	s2
Target	Orthographic representation of target word	baño
Raw	Raw responses prior to post-processing, one per listener	Espana|baño|espana|espana|baño|bano|espana|españa…
Responses	Responses following post-processing, collected into groups; nonwords are identified with an asterisk; the first entry is the majority misperception	España|baño|espainia*|baino*
N-Listeners	Number of listeners who heard the token	15
Counts	For each processed response, in decreasing order	10 3 1 1
Confusion	Most frequently reported response	España
Consistency	Number of listeners reporting majority misperception	10
Target-X-Sampa	Sequence of phonemes corresponding to the target in X-SAMPA notation with syllable boundaries and stress marked	!b a. J o
Target-IPA	Sequence of phonemes corresponding to the target in IPA notation with syllable boundaries and stress marked	!b a. ɲ o
Target-frequency	Normalized frequency (number of occurrences per 10^6 word-forms) of target word according to word-frequency list CREA (REAL, 2008)	44.64
Confusion-X-Sampa	As for Target-X-Sampa	e s! p a. J a
Confusion-IPA	As for Target-IPA	e s! p a. ɲ a
Confusion-frequency	As for Target frequency	525.66
Phoneme-distance	Alignment distance computed using dynamic programming string alignment	34

5. Conclusions

A large collection of consistent misperceptions of Spanish words in noise is available for unrestricted use. Unlike most previous word misperception corpora, waveforms corresponding to the speech and masker stimuli responsible for generating each misperception are available. Consequently, the corpus is suitable for the evaluation of microscopic computational models of speech perception as well as behavioral studies of speech processing. Near-bimodal stimuli in particular lend themselves to studies that investigate the effect of signal or masker manipulation on the reported misperception. The corpus can be found at http://laslab.org/resources/confusions.

Acknowledgments

Corpus collection was funded by the European Community 7th Framework Programme Marie Curie Initial Training Network INSPIRE (Grant Agreement No. FP7-PEOPLE-2011-290000). We thank Dr. Jon Barker for valuable discussions.

References and links

Christiansen, C., Pedersen, M. S., and Dau, T. (**2010**). "Prediction of speech intelligibility based on an auditory preprocessing model," Speech Commun. **52**, 678–692.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**(3), 1562–1573.

Cooke, M. (**2009**). "Discovering consistent word confusions in noise," in *Proceedings of Interspeech*, pp. 1887–1890.

Cutler, A., and Butterfield, S. (**1992**). "Rhythmic cues to speech segmentation: Evidence from juncture misperception," J. Mem. Lang. **31**, 218–236.

Garcia Lecumberri, M. L., Toth, A. M., Tang, Y., and Cooke, M. (**2013**). "Elicitation and analysis of a corpus of robust noise-induced word misperceptions in Spanish," in *Proceedings of Interspeech*, pp. 2807–2811.

Garnes, S., and Bond, Z. S. (**1980**). "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*, edited by A. Fromkin (Academic, New York).

Hernández-Figueroa, Z., Rodríguez-Rodríguez, G., and Carreras-Riudavets, F. (**2012**). Separador de sílabas del español—Silabeador TIP (Separator of Spanish syllables—Syllabifier TIP), http://tip.dis.ulpgc.es/en/syllabification/ (Last viewed September 26, 2014).

Holube, I., and Kollmeier, B. (**1996**). "Speech intelligibility prediction in hearing impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**(3), 1703–1716.

Jürgens, T., and Brand, T. (**2009**). "Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model," J. Acoust. Soc. Am. **126**(5), 2635–2648.

REAL (**2008**). Corpus de referencia del español actual (A reference corpus of modern day Spanish), http://corpus.rae.es/creanet.html (Last viewed July 15, 2014).

Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2010**). "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of ICASSP*, pp. 4214–4217.

Tang, K., and Nevins, A. (**2013**). "Naturalistic speech misperception—a computational corpus-based study," in *Proceedings of the 43rd Meeting of the North East Linguistic Society*, New York.

Vitevich, M. S. (**2002**). "Naturalistic and experimental analyses of word frequency and neighborhood density effects in slips of the ear," Lang. Speech **45**, 407–434.