Elicitation and analysis of a corpus of robust noise-induced word misperceptions in Spanish

Maria Luisa García Lecumberri[1], *Attila Máté Tóth*[1], *Yan Tang*[1], *Martin Cooke*[2,1]

[1]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[2]Basque Foundation for Science, Bilbao, Spain

garcia.lecumberri@ehu.es

Abstract

Slips of the ear are of great relevance in the study of how listeners process speech. Our interest in speech misperceptions comes from their value as diagnostic stimuli in evaluating computational models of speech perception in noise. Previous corpora of misperceptions have largely been recorded based on reports of isolated occurrences 'in the wild', and consequently are not available for further analysis or replication. The current study involves the elicitation in the laboratory of a corpus of over one thousand robust misperceptions of Spanish words induced by stationary and non-stationary maskers. Misperceptions are analysed into single-phoneme substitutions, insertions and deletions, dual vowel/consonant changes, syllable insertions/deletions, compound reformulations and eccentric cases which defy simple explanation. A novel categorisation scheme based on the interaction between the background and foreground is introduced. The new corpus will permit the evaluation of speech perception models that make detailed predictions of listeners' responses to specific speech-in-noise tokens.
Index Terms: speech perception, misperception, corpus, noise

1. Introduction

We are far from a clear understanding of how listeners process speech in noise. While objective prediction of intelligibility in adverse conditions has improved markedly in the last decade (e.g. [1–4]), these 'macroscopic' models are limited to producing simple numeric estimates of overall intelligibility and operate at too coarse a level to provide detailed insights into the many peripheral and central processes that intervene between the reception of a noisy signal and its interpretation as speech. An alternative – termed the 'microscopic' intelligibility modelling approach [5] – aims to predict listeners' responses at a far more fine-grained level of detail (see also [6–9]). Taken to extremes, the challenge is to model responses to individual speech-in-noise tokens based on those reported by a sufficiently large group of listeners [10]. Of particular interest in this pursuit are consistent misperceptions induced by noise.

Slips of the ear [11–18] – the perception of words or utterances different from the ones intended by the speaker – have been studied in the main based on naturally and spontaneously collected examples. Misperceptions show that the listener is engaged in an active process of interpreting speech [13], making the best of the signal in order to understand it. This process often involves a reconstruction of the perceived sounds in order to make them fit possible phonemic, lexical or syntactic candidates. Slips are likely to be more frequent when there is background material to be recruited.

Misperceptions occurring in naturalistic settings are clearly most authentic, but the speech – and, equally-importantly, the misperception-inducing context – is almost never recorded at the signal level in a form suitable for further analysis and replication with other listeners. To address this limitation, some studies have provoked misperception under laboratory conditions. For example, word segmentation has been tested by presenting listeners with faint speech [14], and word frequency and neighbourhood effects have been explored using time-compressed speech [19]. Misperceptions have also been elicited under known noise conditions [10]. The use of noise has usually been regarded as a confounding factor to be avoided in experiments investigating word segmentation, for instance [14]. However, in the current study we are specifically interested in how speech and noise interact to produce misperceptions, and having access to the inducing noise context is essential for the development and evaluation of microscopic models which attempt to explain the misperception.

The current paper describes the elicitation and analysis of a large corpus of Spanish word-level confusions in the presence of stationary and fluctuating masking noise. Using efficient token pruning techniques, listeners were presented with those speech-plus-noise mixtures most likely to promote misperceptions (section 2). Section 3 presents a detailed masker-independent categorisation of confusions and compares the most frequent misperception types with those found previously. Finally, section 4 introduces a preliminary classification scheme involving foreground-background interactions.

2. Elicitation of misperceptions

2.1. Spanish spoken word corpus

A list of 3962 high frequency, 1–3 syllable Spanish words was recorded by two male and two female talkers. Talkers read words from this list, having been trained to avoid list intonation, leaving a short pause between items. Signals were downsampled to 16 kHz, manually-segmented at word boundaries and screened to identify mispronunciations or noise contamination, which led to the removal of 93 items (0.6% of the corpus).

2.2. Maskers

Table 1 lists the five maskers – chosen for their diversity – used in the current study, along with the range of SNRs at which they were mixed with individual words (values based on [10] and further refined in pilot tests). Four of the maskers are non-stationary, while two – BAB4 and BAB8 – are composed of natural speech and can be expected to result in informational as well as energetic masking [20, 21]. All maskers were constructed from material in the Spanish word corpus.

masker		SNR range (dB)
SSN	Speech-shaped noise	-4 to -7
BMN1	Speech modulated noise	-7 to -13
BMN3	3-talker babble modulated noise	-3 to -8
BAB4	4-talker babble	+1 to -3
BAB8	8-talker babble	+1 to -4

Table 1: *Maskers and SNR ranges.*

2.3. Token decision procedure

Since consistent word confusions are quite rare, even in noise [10, 22], the elicitation procedure employed adaptive token-pruning techniques to decide which speech-in-noise tokens were worth pursuing and to eliminate rapidly those unlikely to result in confusions. Tokens were marked as *active*, *discarded* or *exhausted* following the heuristics outlined below. Decisions on token state were taken after each response to that token. The most-listened-to active tokens were presented first. Tokens remained active until any of the following became true:

- L_1 listeners identified the stimuli correctly in a row in the first N presentations, or L_2 listeners identified the stimuli correctly in a row after N presentations
- the responses of the first L_3 listeners were all different
- the token had been presented N_{max} times, at which point it was marked as exhausted

In the first two cases tokens were discarded. Parameter values $L_1 = 2, L_2 = 3, L_3 = 4, N = 8, N_{max} = 15$ were chosen after pilot studies demonstrated their efficiency in maximising token 'turnover' without discarding potentially-interesting items. Deactivated tokens were replaced by newly-generated tokens. Given a target talker, masker type and SNR, new tokens were constructed by choosing a random word from the target talker and embedding it centrally in a randomly-chosen masker fragment at the required SNR, with 200 ms lead and lag time.

2.4. Listeners and procedure

Sixty nine young adults (age: $\mu = 21.9$, $\sigma = 4.6$) studying at the University of the Basque Country took part in the experiment. All passed an audiological screening and were either monolingual in Spanish or bilingual in Spanish and Basque. All were paid for their participation. Listeners identified words in blocks of 100 active tokens by typing their responses into a textfield in a custom Java applet. Within a single block, the target talker and masker type were held constant. Over the course of two non-contiguous one-hour sessions, listeners screened up to 20 blocks made up of all combinations of the 4 talkers and 5 maskers. Listening sessions took place in a sound-attenuated studio containing up to 4 participants in separate sound-isolated cubicles. Stimuli were delivered through Sennheiser HD380 pro headphones.

2.5. Elicitation outcomes

Some 124 865 responses to 23 020 different tokens were collected – a mean rate of 5.4 responses per token. Given the maximum of 15 listeners per token, this amounts to a 2.8-fold increase in efficiency compared to our earlier non-adaptive elicitation techniques [10, 22]. Listeners processed 15.4 tokens per minute. Of the tokens screened, 18 675 (81.1%) were discarded, 2805 (12.2%) survived until exhaustion, and 1540 (6.7%) remained active for further elicitation. From the exhausted tokens, a subset with a minimum listener agreement of 6 were selected, resulting in a corpus of 1248 tokens, equating to an 'interesting' token discovery rate of 9.3 per listener hour.

3. Masker-independent analysis

Misperceptions were initially analysed without regard for the masker in order to obtain a representation of typical patterns for Spanish and to compare our data with previous studies on slips of the ear (notwithstanding likely disparities due to language-specific patterns and the use of noise to induce misperceptions). Target words and corresponding confusions were converted automatically from orthographic form to phoneme sequences, from which homonyms and single phoneme confusions were identified automatically. The remaining items were manually categorised by the first author, a trained phonetician.

3.1. Top-level categories

Misperceptions were classified into a number of categories. After accounting for *homonyms*, cases involving the deletion, insertion or substitution of a *single phoneme* were identified. More complex cases followed the classification of [12, 16] into *dual phoneme* cases involving two consonants, two vowels or a vowel and a consonant, and *syllable* insertion and deletion. We added two further categories: *compounds*, in which the listener appeared to reconstruct part of the misperceived word in ways more complex than the ones listed above, and *eccentric*, which defy simple explanation. Note that although for the purposes of counting we have treated these categories as mutually exclusive, they are clearly open to interpretation. For example, a dual case affecting a vowel and a consonant, or a simple case involving a vowel may also be an instance of syllable insertion or deletion and vice versa. We have endeavoured to classify cases into the category for which they are better representatives.

Table 2 reports the frequency of each confusion type. Overall, we found similar numbers of single segment and more complex slips of the tongue, contrary to previously studies [17] in which single segment errors were far more frequent.

	Number of cases		Percentage	
Category	**corpus**	**category**	**corpus**	**category**
Homonyms	20		1.6	
Single phoneme	579		46.4	
substitution		316	25.3	54.6
insertion		92	7.4	15.9
deletion		171	13.7	29.5
Dual phoneme	225		18.0	
two vowels		4	0.3	1.8
two consonants		103	8.3	45.8
vowel & consonant		118	9.5	52.4
Syllable	40		3.2	
insertion		17	1.4	42.5
deletion		23	1.8	57.5
Compounds	273		21.9	
Eccentric	111		8.9	

Table 2: *Top-level categories.*

3.2. Single phoneme confusions

Single phoneme confusions consist of insertions (e.g., /falsa/ $\mapsto$ /falsas/, a frequent inflectional insertion), deletions (e.g., /tropa/ $\mapsto$ /ropa/, or the inflectional /estabas/ $\mapsto$ /estaba/) and substitutions (e.g., /kiso/ $\mapsto$ /piso/ or /eskutʃe/ $\mapsto$ /eskutʃa/, the latter vowel substitution being a frequent morphological inflectional change). Single segment cases were also annotated

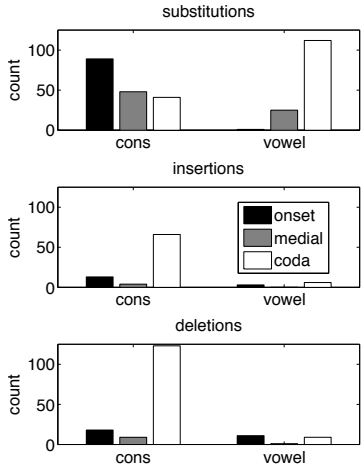

Figure 1: *Breakdown of single-phoneme confusions.*

	phoneme heard																										
	del	a	e	i	o	u	p	b	t	d	k	g	s	f	θ	j	ʎ	l	ɾ	r	m	n	ɲ	sum
ins	.	3	.	.	6	.	4	2	2	2	3	.	45	.	2	.	.	1	18	.	1	3	.	92
a	11	.	9	1	69	1	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	92
e	4	13	.	1	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	28
i	1	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	6
o	3	22	4	.	.	3	.	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	33
u	2	.	.	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	3
p	5	.	.	.	.	.	.	4	1	.	8	.	.	1	.	1	.	1	.	.	1	1	.	23
b	.	.	.	.	.	.	8	.	5	3	.	1	.	3	3	.	.	3	1	.	3	1	.	31
t	9	.	.	.	.	.	1	2	.	3	3	.	.	.	.	.	.	2	.	.	.	.	.	20
d	2	.	.	.	.	.	.	.	4	.	.	.	.	1	1	.	1	1	1	.	.	.	.	11
k	.	.	.	.	.	.	5	.	.	.	.	.	.	1	1	.	.	.	.	1	.	1	.	9
g	2	.	.	.	.	.	.	1	1	1	2	.	.	.	1	.	.	.	.	.	2	.	.	10
s	49	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	50
f	1	.	.	.	.	.	3	.	.	.	1	1	.	.	2	.	.	.	.	.	.	1	.	9
θ	.	.	.	.	.	.	.	.	2	.	.	.	.	.	3	.	.	21	.	.	.	.	.	26
j	.	.	.	.	.	.	3	.	.	.	4	3	.	.	.	.	.	.	.	.	.	.	.	10
ʎ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	.	1
l	1	.	.	.	.	.	.	.	.	1	.	2	.	.	.	.	.	.	2	.	1	.	1	8
ɾ	5	.	.	.	.	.	.	.	.	2	.	.	1	.	.	.	.	2	.	.	.	2	.	12
r	2	.	.	.	.	.	.	1	.	.	1	1	.	.	.	.	.	.	.	.	.	1	.	6
m	1	.	.	.	.	.	.	.	.	.	.	.	2	.	.	.	.	1	.	.	.	3	.	7
n	73	.	.	.	.	.	.	.	1	.	.	.	11	.	.	.	.	2	2	.	1	.	2	92
ɲ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	0
sum	171	42	13	3	85	4	24	10	16	12	22	8	59	6	8	6	1	13	47	1	9	16	3	579

Table 3: *Single phoneme insertions, deletions and substitutions.*

by whether they involve a consonant or a vowel and for position within the word: onset, medial or coda (figure 1). Our data agrees with previous studies [12, 16, 17] in finding that consonant errors are much more frequent than vowel errors (71% vs. 29%). Overall, segments in final position are more likely to be affected (62% vs. 23% initial and 15% medial) contrary to [17]. Additionally, as in other studies, we found that consonants in initial position are more likely to be substituted than inserted or deleted, whereas consonant deletion is frequent in word final position [12, 16]. Our data also suggests that vowels are much more likely to be substituted than inserted or deleted. Such substitutions occur mainly word-finally and rarely initially. This is partly due to the lower frequency of vowel-initial words, and, more interestingly, because in Spanish it is more likely for word final vowels in polysyllabic words to be unstressed, since penultimate stress is the most common pattern. Thus, many of these vowel changes occurred in unstressed syllables, as has been found previously [12, 16].

Table 3 shows counts of the phonemes involved in insertions, deletion and substitutions, and reveals some interesting tendencies. It is notable that /n/ is the sound most often deleted, consistent with [23] who showed that nasals are the most likely consonants to be perceptually deleted in noise, followed by laterals and voiced stops, whereas voiceless stops and sibilant fricatives were more resistant. This explanation is complemented by the fact that final /n/ in Spanish is a verb inflectional marker (third person plural) so that its omission results in possible lexical candidates. Inflectional morphology is also in part responsible for the frequency of /s/ insertions and deletions (plural and third person singular markers), /ɾ/ insertions (infinitive marker), /a,o/ alternations (gender markers) and tense changes (/e/ vs. /a/ or /o/).

3.3. Dual phoneme confusions

Two consonant cases can involve different combinations of the single segment processes, such as in the double insertion /lios/ ↦ /libɾos/ or in the double substitution /poste/ ↦ /boske/ or the insertion and substitution in /lesion/ reported as /presion/. Two vowel errors such as /beber/ ↦/bibiɾ/ are very rare. Vowel and consonant errors (e.g., /todas/ ↦ /todo/, /etʃo/ ↦ /mutʃo/, /nombrar/ ↦ /sembrar/) also include different combinations of single segment processes.

Single segment errors are more than twice as frequent as two segment errors (579 vs. 225). In [12] the proportion is more balanced (200 vs. 177). However, this discrepancy could be due to the fact that we interpreted the 'vowel and consonant' category as consisting of one of each, and thus relegated more complex errors to the compound category. Considering compound and dual errors together, our proportions are quite similar to those reported in [12].

3.4. Simple syllable insertions/deletions

Syllable insertions and deletion made up just over 3% of our corpus, far less than the 15% reported in [12]. This divergence could be due to the smaller degree of syllable weakening found in Spanish compared to English. Examples of insertions include the interesting word-medial case /daras/ ↦ /dexaras/, word-initial /se/ ↦ /pense/ and word-final /tan/ ↦ /taŋgo/. Syllable deletions had a similar frequency, and include word-initial /deten/ ↦ /ten/ and word-final /aθeɾka/ ↦ /aθeɾ/.

3.5. Compound reformulations

Compound processes are those that are more complex than the dual or syllable cases but which can nevertheless be reconstructed (e.g., metatheses such as /kreemos/ ↦ /keremos/ and /medio/ ↦ /miedo/). Other examples include the 3-consonant substitution /pagado/ ↦ /θapato/ (a good demonstration of the superior robustness of vowels in noise) and /suenan/ ↦ /sueɲo/, where the changes involve one consonant substitution, one vowel substitution and a consonant deletion. The case /komun/ ↦ /mundo/ involves both an initial syllable deletion and a final syllable insertion, while /armario/ ↦ /manos/ has initial-syllable deletion and in the final-syllable consonant substitution, vowel deletion and consonant insertion.

3.6. Eccentric cases

Unlike compound reformulations, eccentric examples defy explanation in terms of simple phonetic reconstruction. Eccentric examples in the corpus include /fila/ ↦ /entɾamos/, /nuebe/ ↦ /dutʃas/, /dato/ ↦ /kaktus/ and /komiendo/ ↦ /fumaɾ/. A full analysis of eccentric cases cannot be carried out without

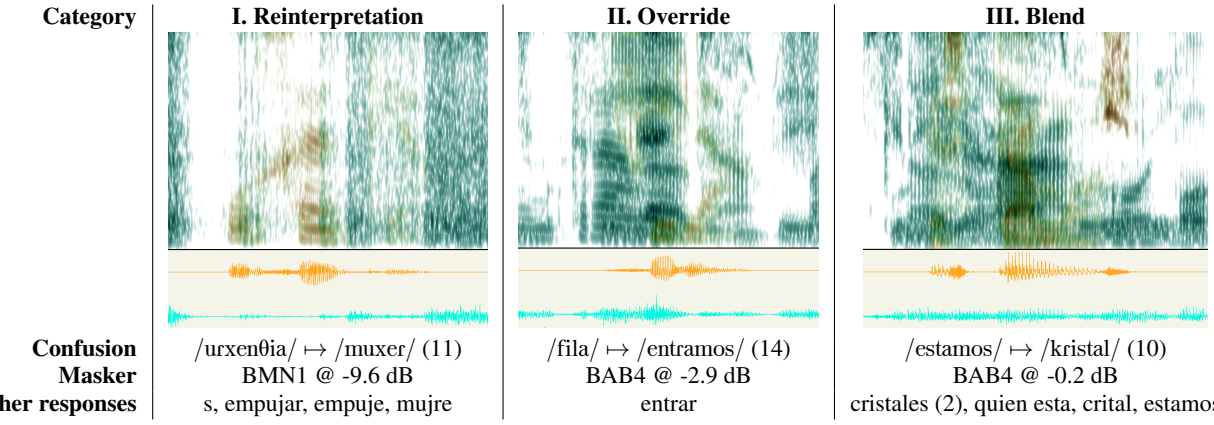

Figure 2: *Examples of target-masker interactions. In all cases the waveform and associated spectrogram components for the presented word are in yellow [colour online] and those for the masker are in blue. The confusion (and number of times reported), masker, SNR and other responses are shown. See the text for further details of each example.*

consideration of the masker, to which we now turn.

4. Masker-dependent analysis

The effect of a masker on speech is traditionally explained in terms of its energetic and informational components [20, 21]. However, this dichotomy is too coarse to be useful in categorising the masker-induced misperceptions in our corpus. Instead we prefer the following tripartite taxonomy based on the degree to which information from the masker appears to contribute to the misperception. Note that this scheme is in part orthogonal to that of section 3 in the sense that these target-masker interactions can, in principle, give rise to many of the observed types of misperception (i.e., single, dual, eccentric, etc.).

I **Reinterpretation:** the reported word is based solely on those components of the target word which escaped energetic masking, using none of the masker components. Here, listeners are forced to reinterpret the audible fragments as a word.

II **Override:** a word contained within the masker is reported in its entirety.

III **Blend:** The reported word is composed of parts of both the target word and the masker. Blends can make use of elements of various types, ranging from subphonemic cues to segments or entire words. They can also incorporate prosodic information from the masker.

An example of each type is given in figure 2. Eleven listeners **reinterpreted** the word 'urgencia' [urgency] as 'mujer' [woman] in the presence of speech-modulated noise. Here, while the start of the word suffers little masking, both the weak fricative /θ/ and unstressed final syllable are masked. Even though there is no evidence of an initial nasal, 'mujer' was chosen as the most likely candidate. Other examples include simple truncations such as /termino/ ↦ /no/ and more complex reinterpretations (e.g., /sanos/ ↦ /sangre/, /kuantas/ ↦ /kuatro/).

For 14 of the 15 listeners, 'entramos' [we enter] from the BAB4 masker **overrode** the target word 'fila' [rank]. Although the SNR was not particularly adverse, the critical stressed vowel /i/ of the target was masked. Further, the start of the reported word happened to coincide with that of the masker.

Twelve listeners reported either 'cristal' [glass] or its plural form in place of the target 'estamos' [we are], a transformation which appears to require the **blending** of the target word sequence /sta/ with elements from the masker. The corpus contains many examples of apparent blending ranging in complexity from /ojos/ ↦ /pojos/ and /rampa/ ↦ /trampa/ to /primos/ ↦ /tʃikos/ and /arde/ ↦ /arboles/.

5. Discussion and further work

To date we have focused on a phonetic analysis of noise-induced misperceptions to facilitate comparison with earlier corpora. In the next phase the target/masker-based classification scheme will be elaborated based on close auditory screening and the incidence of each category assessed in the corpus and related to factors such as masker type and SNR level.

The corpus will be further stratified to provide subsets of tokens for modelling purposes. In the current study only the most frequent confusions with an above-threshold minimum listener agreement were analysed, but there is also useful information in the entire response sequence. For example, around 4% of tokens produced a strongly-bimodal response, which will be of interest in microscopic modelling as these exemplars demonstrate a clear preference for a narrow response set which ought to be reflected in models' word likelihoods.

Follow-up listening tests – along the lines of [10] – will be undertaken to determine what properties of the target and masker combine to give rise to the misperception. For example, adjustment of SNR, modified temporal alignment relative to the masker, or resynthesis with changes in fundamental frequency can be expected to weaken or eliminate the misperception. Finding the point at which this occurs should provide insights into the way salient information in speech is both represented and integrated during word identification.

Finally, microscopic models are in their infancy and it is surely premature to expect them to respond to slips of the ear in the same way as listeners. Nevertheless, we believe that the existence of a sizeable corpus of confusions is an essential component of an evaluation framework which will guide the development of future speech perception models.

Acknowledgements. The research leading to these results was partly funded from the European Community 7th Framework Programme Marie Curie INSPIRE ITN, the Language and Speech project of the Basque Government and the Spanish Government DIACEX grant FFI 2012-31597.

6. References

[1] K. S. Rhebergen and N. J. Versfeld, "Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[2] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Comm.*, vol. 52, no. 7-8, pp. 678–692, 2010.

[3] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2224–2237, 2005.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

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

[6] T. Jürgens, T. Brand, and B. Kollmeier, "Modelling the human-machine gap in speech reception: microscopic speech intelligibility prediction for normal-hearing subjects with an auditory model," in *Proc. Interspeech*, 2007, pp. 410–413.

[7] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1220–1233, 2008.

[8] 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.

[9] J.-P. Ramirez, H. Ketabdar, and A. Raake, "Intelligibility predictions for speech against fluctuating masker," in *Proc. Interspeech*, 2010, pp. 2486–2489.

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

[11] 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.

[12] ——, "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.

[13] A. Cutler, "The reliability of speech error data," *Linguistics*, vol. 19, pp. 561–582, 1981.

[14] 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.

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

[16] ——, "Slips of the ear," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Oxford: Blackwell, 2005, pp. 290–310.

[17] 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 Linguistic, 2004.

[18] 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*, 2012.

[19] 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.

[20] R. Carhart, T. Tillman, and E. Greetis, "Perceptual masking in multiple sound backgrounds," *J. Acoust. Soc. Am.*, vol. 45, pp. 694–703, 1969.

[21] D. Brungart, B. Simpson, M. Ericson, and K. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 100, pp. 2527–2538, 2001.

[22] M. Cooke, J. Barker, and M. L. Garcia Lecumberri, "Crowdsourcing in speech perception," in *Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment*, M. Eskenazi, G.-A. Levow, H. Meng, G. Parent, and D. Suendermann, Eds. John Wiley, 2013, pp. 141–176.

[23] J. Benki, "Analysis of English nonsense syllable recognition in noise," *Phonetica*, vol. 60, pp. 129–157, 2003.