A corpus of noise-induced word misperceptions for English

Ricard Marxer,[1,a)] **Jon Barker,**[1] **Martin Cooke,**[2]
and Maria Luisa Garcia Lecumberri[3]
[1]*Department of Computer Science, University of Sheffield, Sheffield, United Kingdom*
[2]*Ikerbasque (Basque Science Foundation), Bilbao, Spain*
[3]*Language and Speech Laboratory, Universidad del Pais Vasco, 01006 Vitoria, Spain*
r.marxer@sheffield.ac.uk, j.p.barker@sheffield.ac.uk, m.cooke@ikerbasque.org,
garcia.lecumberri@ehu.es

Abstract: Words spoken against a noise background often form an ambiguous percept. However, in certain conditions, a listener will mishear a noisy word but report hearing the same incorrect word as reported by other listeners. These *consistent* hearing errors are valuable as tests of detailed models of speech perception. This paper describes the collection of a corpus of consistent speech misperceptions for English. The mishearings were elicited using a large scale listening study involving 212 participants and over 300 000 token presentations. The study led to the identification of 3207 consistent misperceptions. For each of these, the corpus records the speech and masker waveforms that generated the error, the set of responses made by the listeners, and phonemic transcriptions of the target word and the response. The corpus is freely available online.

[RS]
Date Received: March 15, 2016 **Date Accepted:** August 11, 2016

1. Introduction

Speech intelligibility modeling has traditionally been concerned with designing objective intelligibility measures that are able to estimate average word identification scores in broadly stated noise conditions (Taghia and Martin, 2014; Taal *et al.*, 2010; Rhebergen and Versfeld, 2005). In contrast, there has been a growing interest in "microscopic" intelligibility models that are able to make predictions of listener responses to specific noisy speech stimuli (Cooke, 2006; Jürgens and Brand, 2009; Geravanchizadeh and Fallah, 2015). These models not only predict whether a word will be heard correctly, but also attempt to predict the specific errors that will occur in the case that a word is misheard.

Evaluation of microscopic models requires data that characterizes listeners' misperceptions. One methodology for capturing such data is to present noise-corrupted words to a group of listeners and to ask them to report the word that is heard, e.g., by typing a response. In certain noise conditions it can happen that a group of listeners will consistently mishear a word in the same way. These so-called "consistent confusions" form a valuable diagnostic tool: a good microscopic model should agree with listeners in situations where listeners agree amongst themselves.

The value of speech listening errors, and of consistent confusions in particular, has motivated the publication of several collections of misheard speech stimuli. The largest of these is a recent Spanish study and contains 3235 examples of misperceptions elicited through a large-scale listening experiment (Toth *et al.*, 2015). The current paper follows the same methodology as Toth *et al.* (2015) to collect an equivalent corpus for English confusions.

The motivation for the new confusion corpus is not only that an English corpus will be a valuable resource in its own right, but also that by mirroring the Spanish collection it will facilitate the study of intelligibility from a cross-language perspective. Spanish and English have phonological differences that make such a study potentially fruitful. For example, English is stress-timed rather than syllable-timed (Abercrombie, 1967); consequently, stress placement has important consequences for the segmental make up of syllables, since unstressed syllables typically have weak vowels. English has a much larger set of vowels and is less phonotactically restricted in its use of consonants and consonant clusters. Pellegrino *et al.* (2011) compares languages in terms of

[a)]Author to whom correspondence should be addressed.

the information density of an average syllable—based on the number of syllables needed to represent the same semantic content—and finds that the density of English is much higher than Spanish. Thus English, on average, needs fewer syllables than Spanish to convey the same information. Considering these differences it can be expected that Spanish and English have very different vulnerabilities to the effects of noise masking.

2. Methods

The corpus is constructed from the responses of listeners to common English words mixed with random noise maskers. Responses are collected for each noisy token [i.e., a word-masker combination at a specific signal-to-noise ratio (SNR)] from 15 different listeners. Following Toth *et al.* (2015), noisy tokens are added to the consistent confusion collection if they are misheard in the same way by at least 6 of the 15 listeners. Responses in the same homophone set (e.g., "hear" and "here") are treated as being the same.

2.1 Speech material

Recordings of English words were provided by four native British English talkers all of whom were clear speakers without strong regional accents, two male (S1 and S2) and two female (S3 and S4). Speakers read a word list containing 3134 of the most frequent English words of up to three syllables, resulting in an average of 1.5 syllables/4.26 phonemes per word. Word frequencies were taken from the CELEX English Lexical Database (Baayen *et al.*, 1995) which is based on an analysis of written texts. Talkers were trained to avoid list intonation. Recordings took place in an IAC single-walled acoustically-isolated booth (IAC Acoustics, North Aurora, IL) using a Bruel & Kjaer (B & K) type 4190 1/2-in. microphone (Bruel & Kjaer, Denmark) placed approximately 30 cm in front of the talker. The signal was preamplified by a B & K Nexus model 2690 conditioning amplifier prior to digitization by a MOTU 8pre analogue to digital interface (MOTU Inc., Cambridge, MA). The resulting recordings were manually segmented into words and downsampled to 16 kHz. A total of 12 489 items remained after removal of mispronounced or noise-contaminated items comprising 3126, 3125, 3109, and 3129 words for talkers S1 to S4, respectively.

2.2 Maskers

In order to induce misperceptions, three different types of noise masker were generated: stationary speech shaped noise (SSN); four-talker babble (BAB4); and three-talker babble modulated noise (BMN3). The BAB4 signal was generated by first concatenating randomly selected words from the recorded speech materials to form prolonged streams of speech and then summing four such streams. The BMN3 masker was generated by estimating the envelope of a three-talker babble signal and then using this envelope to modulate a SSN carrier. Speech-plus-noise stimuli were generated by randomly selecting segments of the masker stream and mixing them with the target speech at SNRs within masker-specific ranges, namely [−7, −4], [−8, −3], and [−3, +1] dB for the SSN, BMN3, and BAB4 maskers, respectively. These were chosen based on those reported in Toth *et al.* (2015), after confirming their effectiveness in a series in pilot listening tests. Note that the BMN3 masker shares non-stationary properties with BAB4 (and so will produce a similar pattern of energetic masking), while being similar to SSN in that it does not contain recognizable speech.

2.3 Participants

A cohort of 212 listeners provided responses to the stimuli presented. Participants were students recruited at the University of Edinburgh (mean age 23.9; standard deviation 6.5). They reported to be native English speakers with normal hearing. Participants provided written consent to use their responses anonymously and were paid to perform the task.

2.4 Procedure

A total of 12 listening conditions were formed by combining all pairings of the four speakers with the three masker noise types. Stimuli were presented in single condition blocks of 50 tokens, randomly selected for each listener. During a 1 h session a participant listened to as many 50-token blocks as time allowed. Each block began with the presentation of 5 control stimuli, starting with an SNR of 30 dB and ramping down linearly until reaching the upper bound SNR value for the condition. For the remaining 45 stimuli, the SNR was set randomly within the range of SNRs specified for that masker type, but ordered such that SNR reduced throughout the block.

Table 1. Counts of consistent misperceptions collected per speaker/noise condition.

		Masker				
Speaker	Gender	BAB4	BMN3	SSN	Totals	Percentage
S1	Male	197	248	174	619	19.30
S2	Male	227	294	217	738	23.01
S3	Female	335	355	296	986	30.75
S4	Female	312	303	249	864	26.94
Totals		1071	1200	936	3207	100.0
Percentage		33.40	37.42	29.19	100.0	

Responses were collected using a simple software interface that allowed listeners to type the word they heard into a text box. Accepted responses were restricted to words available in a British English dictionary. If the input did not match any word in the dictionary, the listener was prompted to correct the spelling or skip to the next stimulus. Participants completed up to four 1-h sessions in total, but with no more than one session per day.

2.5 *Dynamic stimulus generation*

Tokens were generated and formed into blocks in a dynamic manner under the control of software that was previously used for the collection of the Spanish confusion corpus (Toth *et al.*, 2015). This software allows consistent confusions to be collected efficiently using an online token selection process. In brief, the software generates and maintains a pool of noisy tokens with which it populates the listening experiments. Tokens remain in the pool until they have been heard by 15 listeners.[1] However, a token can be removed from the pool and discarded before being heard by all 15 listeners (i.e., "pruned") if the developing pattern of responses makes it seem unlikely that it will be consistently misheard. This pruning can greatly improve efficiency, i.e., the number of consistent confusions discovered per token presentation. Stimuli were pruned if the tokens were heard correctly by three consecutive listeners or the ratio of correct to incorrect responses exceeded 1.5.

2.6 *Post-processing*

Listening experiments were conducted over two separate one-week periods. At the end of this time, responses were collected for all tokens that had been heard the required 15 times. Of these, only responses to tokens that had been consistently confused were retained and recorded in the published dataset.

For most entries in the dataset there are 15 separate lexical responses from listeners. However, it was noticed during analysis that 9 of the 212 listeners did not meet the enrollment criteria (native English with self-reported normal hearing). These listeners were discarded leaving a number of tokens with one or two fewer responses.

For each entry in the corpus, Arpabet (Shoup, 1980) and IPA transcriptions of the target word and the most consistent confusion have been provided. Arpabet transcriptions were taken from the BEEP dictionary (BEEP, 1997). Where BEEP lists multiple pronunciations of the target word, a listener with phonetic training selected the pronunciation that best matches what was spoken. For the consistent confusions—which were responses typed by the listener—the pronunciation has been chosen as that which has the smallest edit distance to the target phone sequence. The IPA

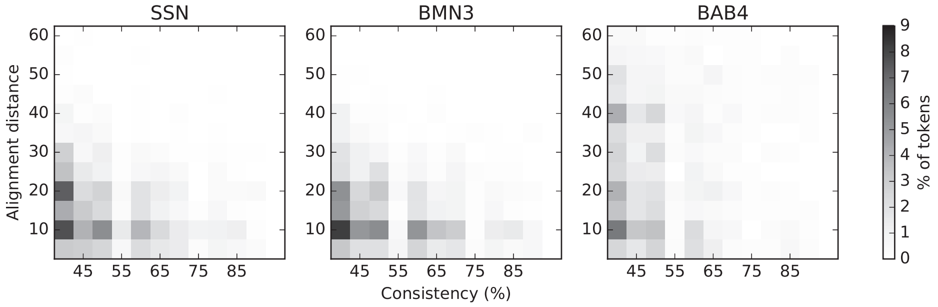

Fig. 1. Misperception density with relation to alignment distance and consistency.

[http://dx.doi.org/10.1121/1.4967185]

Published Online 23 November 2016

Table 2. Example corpus entry for the word “shrewd” in SSN, misperceived by 6 out of 15 listeners as “intrude.”

Field	Description	Example
ID	Integer used to identify the speech waveform corresponding to the entry	20 123
Length	Speech signal length in samples	14 720
Masker	One of [SSN, BMN3, BAB4]	SSN
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	129 438
SNR	SNR in dB	−6.262
Speaker	One of [s1, s2, s3, s4]	s3
Target	Orthographic representation of target word	shrewd
Raw	Raw responses prior to post-processing, one per listener	intrude/shrew/shrewd/rude/intrude/shrewd/intrude/intrude/...
Responses	Responses following post-processing, collected into groups; presented in decreasing order of response count	intrude/shrewd/rude/shrew/truth
N-Listeners	Number of listeners who heard the token	15
Counts	Number of listeners reporting each response	6 5 2 1 1
Confusion	Most frequently reported response	intrude
Consistency	Number of listeners reporting majority misperception	6
Target-Arpabet	Sequence of phonemes corresponding to the target in Arpabet notation with syllable boundaries and stress marked	! SH R UW D
Target-IPA	Sequence of phonemes corresponding to the target in IPA notation with syllable boundaries and stress marked	! ʃ ɹ u d
Target-frequency	Normalized frequency on Zipf scale (log10 occurrences per 10^9 word-forms) for target word according to word-frequency list SUBTLEX-UK (van Heuven *et al.*, 2014)	1.17
Confusion-Arpabet	Same as Target-Arpabet for the Confusion	IH N! T R UW D
Confusion-IPA	Same as Target-IPA for the Confusion	ɪ n! t ɹ u d
Confusion-frequency	Same as Target-frequency for the Confusion	2.92
Phoneme-distance	Alignment distance between Target and Confusion phoneme sequences	24

transcriptions are based on a 1-to-1 translation of Arpabet symbols to equivalent IPA symbols. Syllable boundaries have been marked with a period, while "!" denotes the start of the stressed syllable. Syllable boundaries and stress patterns are taken from the CELEX database (Baayen *et al.*, 1995). Again, ambiguous cases were resolved by a listener.

3. Results

A total of 301 696 responses were collected from 41 437 unique stimuli presented. Some 9725 survived pruning and received responses by at least 15 listeners. Of these, exactly 3207 passed the condition of minimal consistency where at least 6 of the listeners reported the same incorrect response.

Table 1 summarizes the number of consistent confusions elicited in each speaker and noise-type subcondition. As expected, all noise conditions produced significant numbers of confusions. As with the Spanish study (Toth *et al.*, 2015), the modulated BMN3 masker produced more confusions than the stationary SSN masker. There is variability across speakers with speaker S3 producing 59% more confusions than speaker S1 (986 versus 619). It is possible that acoustic factors such as spectral tilt played a role, alongside acoustic-phonetic properties underlying speech clarity; we also cannot rule out an effect of mismatch between the accent of the talkers and those of the listeners. These are matters for further investigation.

Figure 1 reproduces an analysis presented in Toth *et al.* (2015). For each token, the target word was phonetically aligned with the majority confusion and the alignment distance was measured. Alignment between target and confusion was computed using a dynamic programming algorithm with costs for insertions, deletions, and substitutions of 7, 7, and 10, respectively. The figure displays the distribution of alignment distances as a function of consistency (the percentage of listeners reporting the same misperception) for each of the different noise maskers. It can be seen that small alignment distances (i.e., confusions that are close to the correct response) are more common and tend to have the highest consistency scores. However, many confusions have large distances suggesting that they are phonetically far removed from the target word. This is particularly the case for the BAB4 masker, the only masker that contains competing phonetic material. The pattern of distributions very closely follows those observed in the Spanish study.

4. Corpus description

The corpus is being made freely available for download under a Creative Commons Attribution 4.0 International license.[2] The download consists of several components: (a) waveforms for the speech and masker pairs that led to each of the 3207 consistent misperceptions; (b) transcriptions for the target speech and the BAB4 maskers; (c) the BAB4 and BMN3 signals from which masker segments were drawn (SSN maskers were generated for each token individually); and (d) a tabulated description of the 3207 consistent confusions where each entry has the structure shown in Table 2.

5. Summary

The paper has presented a corpus of noise-induced speech misperceptions for English that is freely available online. The corpus has been collected using a listening study in which over 300 000 dynamically-generated tokens were presented to 212 listeners to result in over 3000 consistent mishearings. The corpus is comparable in scale and design to the Spanish corpus of Toth *et al.* (2015) allowing for the study of language-dependent effects in microscopic computational models of speech perception.

Acknowledgments

The corpus collection was funded by the EC 7th Framework Programme Marie Curie ITN INSPIRE (Grant Agreement No. FP7-PEOPLE-2011-290000). Thanks to Yan Tang and Mate Toth for contributing to the development of the corpus collection software.

References and links

[1] Some listeners participated in parallel, leading to some tokens with more than 15 listens.

[2] The corpus can be downloaded from http://spandh.dcs.shef.ac.uk/ECCC.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Edinburgh University Press, Scotland).

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (**1995**). "The CELEX Lexical Database" (Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995), (CD-ROM).

BEEP (**1997**). "The British English Example Pronunciation (BEEP) dictionary," http://svr-www.eng.cam.ac.uk/comp.-speech/Section1/Lexical/beep.html (Last viewed February 2016).

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

Geravanchizadeh, M., and Fallah, A. (**2015**). "Microscopic prediction of speech intelligibility in spatially distributed speech-shaped noise for normal-hearing listeners," J. Acoust. Soc. Am. **138**, 4004–4015.

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.

Pellegrino, F., Coupé, C., and Marsico, E. (**2011**). "A cross-language perspective on speech information rate," Language **87**(3), 539–558.

Rhebergen, K. S., and Versfeld, N. J. (**2005**). "Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**(4), 2181–2192.

Shoup, J. E. (**1980**). "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, edited by W. A. Lea (Prentice-Hall, Englewood Cliffs, NJ), Chap. 6, pp. 125–139.

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.

Taghia, J., and Martin, R. (**2014**). "Objective intelligibility measures based on mutual information for speech subjected to speech enhancement processing," IEEE Trans. Audio, Speech, Lang. Process. **22**(1), 6–16.

Toth, M. A., Garcia Lecumberri, M. L., Tang, Y., and Cooke, M. (**2015**). "A corpus of noise-induced word misperceptions for Spanish," J. Acoust. Soc. Am. **137**(2), EL184–EL189.

van Heuven, W. J., Mandera, P., Keuleers, E., and Brysbaert, M. (**2014**). "SUBTLEX-UK: A new and improved word frequency database for British English," Q. J. Exp. Psychol. **67**(6), 1176–1190.