The English Consistent Confusion Corpus (v1.1)

Ricard Marxer, Jon Barker and Martin Cooke

February 4, 2016

1 Methods

The corpus is constructed from the responses of listeners to common English words mixed with a random noise maskers. In particular, responses are collected for each noisy token (i.e. a word-masker combination) from 15 different listeners. 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. Details are provided in the sections that follow.

1.1 Speech material

Recordings of common English words were provided by four talkers, two male (S1 and S2) and two female (S3 and S4), who read a word list containing 3134 of the most frequent English words of up to three syllables. Talkers were trained to avoid list intonation. Recordings took place in an IAC single-walled acoustically-isolated booth using a Bruel & Kjaer (B & K) type 4190 $\frac{1}{2}$-in. microphone place 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. The resulting recordings were manually segmented and downsampled to 16kHz. A total of 12,489 items remained after removal of mispronounced or noise-contaminated items comprising of 3126, 3125, 3109 and 3129 words for talkers S1 to S4 respectively.

1.2 Maskers

In order to induce misperceptions, three different types of noise masker types were generated (Table 1): stationary speech shaped noise (SSN); four-speaker speech babble (BAB4); and three-speaker 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 was generated by estimating the envelope of a three-speaker babble signal and then using this to modulate a speech-shaped noise carrier. Speech-plus-noise stimuli were generated at SNRs within masker specific ranges shown in the table. These ranges were chosen based on the ranges reported in Toth et al. [2015] and after confirming their effectiveness in a series in pilot listening tests.

Table 1: Maskers used in the experiment. The column headed “Speech?” indicates those maskers containing speech signals.

	Masker	Speech?	Non-stationary?	SNR range (dB)
SSN	Speech shaped noise	✗	✗	−7 to −4
BMN3	Three-talker babble modulated noise	✗	✓	−8 to −3
BAB4	Four-talker natural babble	✓	✓	−3 to +1

1.3 Participants

A total of 214 listeners provided responses to the stimuli presented. Participants were students recruited at the University of Edinburgh (mean age 23.9 std. 6.5). They reported to be native English speakers with a range of different accents. Participants provided written consent to use their responses anonymously and were paid to perform the task.

1.4 Procedure

A total of 12 listening conditions were formed by considering all pairings of the four speakers with the three masker noise types. Stimuli were presented in blocks of 50 tokens in a single conditions. During a 1 hour 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 the upper bound SNR value of the condition as shown in Table 1. 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 was reduced throughout the block.

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 attempt correcting the spelling or skip to the next stimuli. Participants were allowed to complete up to four 1 hour sessions in total, but with no more than one session per day.

1.5 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 and using an online 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. 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. number of consistent confusions discovered per listener-token presentation. For details of the pruning rules see Toth et al. [2015]

1.6 Postprocessing

Listening experiments were conducted over two separate one week periods. At the end of this time, responses to all tokens that had been heard the required 15 times were collected. Of these, only responses to tokens that had been consistently confused were retained and recorded in the published dataset. Here, following the definition in Toth et al. [2015], a token is considered to be consistently confused if 6 or more listeners hear the token incorrectly but report the same word (or words in the same homophone set).

For most consistent confusions there are 15 separate lexical responses from listeners. However, it was noticed during analysis that a small number of the 214 listeners had not met the enrollment criteria (native English with self-reported normal hearing). Responses from these listeners were discarded leaving a number of tokens with a few less than 15 responses. Further, listeners occasionally chose to use the option to provide no response by skipping the token (see Section 1.4).

For each entry in the corpus, Arpabet and IPA transcriptions of the target word and the most consistent confusion have been provided. Arpabet transcriptions were taken from a merge of the BEEP dictionary and CMUdict lexicons, normalized to the same phone sets without stress markers. The pronunciation of the target word has been selected as the one that best matches what was actually 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.

Note, Arpabet transcription come directly from CMUdict and BEEP. The IPA transcriptions are based on a 1-to-1 translation of Arpabet symbols to equivalent IPA symbols.

2 Results

A total of 301,696 responses were collected from 41,437 unique stimuli presented. Only 9725 of the stimuli were not pruned and received responses by at least 15 listeners. Of these, 3135 passed the condition of minimal consistency where at least 6 of the listeners agreed in the response. On average each listener provided 763.8 (std. 234.6) responses per session. The number of times a participant skipped a response in each session remained rather low for most of the cases (avg. 35.9 std. 60.6).

Table 2: Counts of consistent misperceptions collected per speaker and noise condition.

		Masker				
Speaker	Gender	BAB4	BMN3	SSN	Totals	Percentage
S1	Male	197	248	174	619	19.34
S2	Male	227	294	215	736	23.00
S3	Female	334	353	294	981	30.66
S4	Female	313	302	249	864	27.00
Totals		1071	1197	932	3200	100.0
Percentage		33.47	37.41	29.12	100.0	

References

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

Table 3: Example corpus entry for the word “shrewd” in speech-shaped noise, misperceived by 6 out of 15 listeners as “intrude”

Field	Description	Example
ID	Integer used to identify the speech waveform corresponding to the entry	20123
Length	Speech signal length in samples	14720
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	129438
SNR	Signal-to-noise ratio 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; nonwords are identified by an asterisk; the first entry is the majority misperception	intrude|shrewd|rude|shrew|truth
N-Listeners	Number of listeners who heard the token	15
Counts	For each processed, in decreasing order	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	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 (number of occurrences per 10^6 word-forms) of target word according to word-frequency list SUBTLEX-UK	1.17
Confusion-Arpabet	As for Target-Arpabet	IH N T R UW D
Confusion-IPA	As for Target-IPA	ɪ n t ɹ u d
Confusion-frequency	As for Target-frequency	2.92
Phoneme-distance	Alignment distance computed using dynamic programming string alignment	24.0