Discovering consistent word confusions in noise

Martin Cooke[1,2]

[1]Ikerbasque (Basque Science Foundation)
[2]Language and Speech Laboratory, Faculty of Letters, Universidad del País Vasco, Spain
m.cooke@ikerbasque.org

Abstract

Listeners make mistakes when communicating under adverse conditions, with overall error rates reasonably well-predicted by existing speech intelligibility metrics. However, a detailed examination of confusions made by a majority of listeners is more likely to provide insights into processes of normal word recognition. The current study measured the rate at which robust misperceptions occurred for highly-confusable words embedded in noise. In a second experiment, confusions discovered in the first listening test were subjected to a range of manipulations designed to help identify their cause. These experiments reveal that while majority confusions are quite rare, they occur sufficiently often to make large-scale discovery worthwhile. Surprisingly few misperceptions were due solely to energetic masking by the noise, suggesting that speech and noise "react" in complex ways which are not well-described by traditional masking concepts.

Index Terms: speech perception, word confusions, noise

1. Introduction

Any account of speech perception must be able to explain the robustness of spoken communication in the face of reverberation, environmental noise, competing talkers and channel distortions [1, 2]. Over the years, increasingly accurate predictions of overall speech intelligibility for a wide class of distortions have been made [3, 4, 5]. However, these *macroscopic* models solely provide a numeric indication of how accurately speech in general will be perceived in a given condition, and are not designed to produce detailed insights into the processes of speech perception. Providing a good match to average listener recognition rates is one thing, but matching listener responses at the level of individual tokens is much more challenging.

In contrast, *microscopic* models [6, 7, 8] have been proposed to make specific predictions about how individual tokens of noisy speech will be perceived. These models are in their infancy and, to date, have had limited success, operating in restricted domains such as intervocalic consonant identification [9]. At present, much use is made of traditional information transmission measures [10], but these, like confusion matrices, are summary statistics which are less than precise in pinpointing model deficiencies. An important exception is [8], whose method for detailed analysis of individual tokens in noise is closest in spirit to what is proposed here.

The essential idea of the approach described in this paper is straightforward: to find examples of listener responses to speech in noise which demonstrate *consistent confusions* i.e. tokens for which a majority of listeners make the same error. Figure 1 illustrates consistent confusions for words embedded in babble noise. A large corpus of such confusions would be immensely valuable as a mechanism for both diagnosing and evaluating microscopic speech perception models at a fine-grained level of detail. In addition, a confusions corpus has the potential to address the issue of what aspects of speech make it robust or vulnerable to noise.

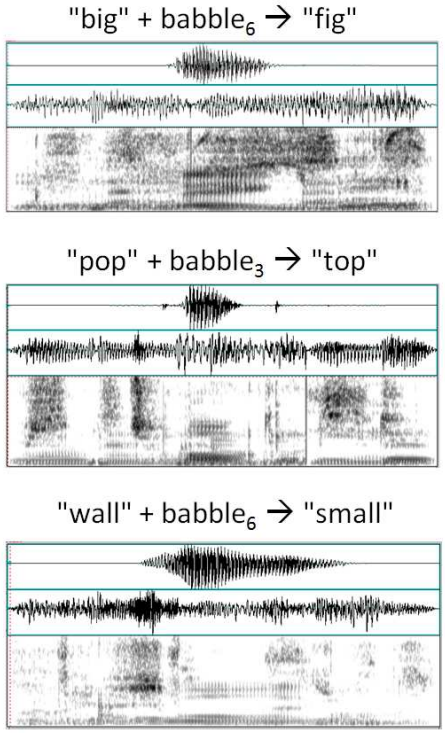

Figure 1: *Speech and babble noise "reactions" which result in confusions.*

However, it is far from clear how best to find examples of consistent confusions. The current study takes the form of a preliminary investigation to test the feasibility of large-scale confusion discovery. Two listening experiments were performed to (i) estimate the *discovery rate* for consistent confusions i.e. the proportion of noisy speech tokens which lead to a given rate of listener agreement on the incorrect answer; and (ii) to determine whether the cause of some confusions can be identified semi-automatically, allowing filtering of confusions into categories. Section 2 describes the word identification in noise task employed to provoke confusions in a group of listeners, while section 3 shows how speech and noise signals which result in interesting confusions can be manipulated to expose possible reasons for the confusion. Section 4 examines the feasibility of collecting a large-scale confusions corpus.

2. Discovering consistent confusions

Experiment 1 measured listener response agreements in order to estimate the *discovery rate* for consistent confusions in a variety of noise backgrounds and signal-to-noise ratios (SNRs).

2.1. Word material

A word identification task was chosen in preference to the use of logatomes or syllables, anticipating the eventual need to use a large population of phonetically-naive listeners for corpus collection. To maximise the likelihood of confusions, words were monosyllabic, relatively common, and had dense neighbourhoods of similar-sounding words. The following criteria were applied:

- **neighbourhood density** (defined as the set of words that differ by the insertion, deletion or substitution of a single phoneme) ≥ 20
- maximum of **written/spoken frequency** ≥ 10 per million

Word lists were obtained by filtering an existing monosyllabic lexical database which contained both neighbourhood density and frequency statistics [11, 12]. After the removal of homophones, 613 items met the criteria above. Table 1 shows a selection of words with the largest and smallest number of potential confusions. For example, insertion, deletion or substitution of a single phoneme in *bore* leads to 77 other words.

Table 1: *Example monosyllabic words, phonological neighbourhood density (ND) and frequency/million.*

	IPA	ND	freq (spoken)	freq (written)
bore	/bɔː/	77	19	41
pour	/pɔː/	73	27	83
awe	/ɔː/	73	0	12
...	...	...	...	...
source	/sɔːs/	20	111	129
learn	/lɜːn/	20	366	303
teach	/tiːtʃ/	20	276	132

Words were recorded at 50 kHz in an IAC single-walled sound booth by 2 male native British English speakers in citation form, then endpointed, filtered to remove energy at frequencies below 50 Hz, and downsampled to 16 kHz.

2.2. Experiment 1

Words were centrally-embedded in noise fragments drawn from the list in table 2 with 200 ms lead and lag time. Stimuli were presented in blocks of 100, with SNR varying incrementally from moderate to intense through the block, based on the extremes shown in table 2, values chosen on the basis of extensive pilot testing to avoid a low probability of confusions for the least adverse SNRs and a high probability of near-random responses at the more adverse end. While no claims can be made about the optimality of these figures, it is noteworthy that consistent confusions were likely to occur for rather a narrow range of SNRs for certain noise types. Tokens from the two speakers were used for each of the 12 noise backgrounds, for a total of 2400 stimuli.

Ten native British English listeners with no hearing problems identified words in noise under computer-control. Stimuli were delivered over Sennheiser HD 250 Linear II headphones in an IAC booth. Motivated by the goal of provoking consistent confusions, and unlike traditional speech perception studies, no randomisation of blocks or stimuli within blocks was performed i.e. all listeners heard the stimuli in the same order. This meant that context effects, such as priming from previously-presented words, were the same for all listeners. A single noise type and speaker were used within any given block. While randomisation of noise type, SNR and speaker might be expected to produce more confusions, the current task was more like the situation faced by listeners in the real world. Subjectively, increasing the noise level throughout the block had the effect of engaging the listener. Listeners proceeded through the 24 blocks at their own pace in two sessions of 40 minutes on separate days. Listeners responded by typing the word they heard. On average, listeners required 2.2 seconds per stimulus.

Table 2: *Noise types and SNR ranges for expt. 1.*

noise	SNR_{max}	SNR_{min}
competing talker	-5	-13
2-talker babble	-3	-9
3-talker babble	-2	-7
4-talker babble	-2	-7
6-talker babble	-1	-6
8-talker babble	0	-7
16-talker babble	0	-6
speech-shaped noise	0	-4
1-talker modulated SSN	-2	-5
4-talker modulated SSN	0	-4
reversed talker	-5	-11
factory noise	2	-3

2.3. Results

To assess listeners agreements, responses were first mapped to a single homophone in the cases where multiple equivalent orthographic forms existed (e.g. [by, bye, buy] $\mapsto$ by). No extensive manual checking for typos or spelling errors was performed, but a limited inspection suggested a very low incidence.

Table 3 summarises listener agreements for these stimuli. Listeners were unanimous in choosing the correct answer for 29.8% of the 2400 tokens, while a majority agreed on the correct word nearly 70% of the time. For consistent confusions, unanimity occurred for far fewer tokens (0.38%), but a majority agreed on the wrong answer nearly 7% of the time (161/2400). All noise types led to confusions, with the competing talker background resulting in one majority confusion for every 9 tokens.

Table 3: *Percentage of tokens with a given rate of listener agreement, for correct and confused responses. The lower row represented the discovery rate for consistent confusions.*

agreement	100%	$\geq 90\%$	$\geq 80\%$	$\geq 70\%$	$\geq 60\%$
correct	29.8%	44.2%	54.3%	62.9%	69.5%
confused	0.38%	0.96%	2.33%	4.25%	6.71%

Table 4 lists example confusions with agreement of 80% or more. Deletions and insertions as well as substitutions (or equivalently deletion followed by insertion) abound, particularly for consonants. /p/ is particularly detachable while word-initial /l/ seems to attract a preceding consonant.

Table 4: *Example majority confusions (format: sent-heard).*

80%	90%	100%
dull-doll, lip-flip	pad-had	big-fig
lock-block, pat-fat	howl-owl	ought-port
lash-flash, hall-fall	toll-told	dole-doll
boil-oil, cheap-cheek	peak-beak	force-horse
beer-fear, limb-live	ill-kill	boil-oil
wall-walk, eye-high	harm-hard	veil-fail
bow-bone, chew-tune	pop-top	wide-white
pat-path, more-normal	lure-law	pill-hill

3. Identifying the cause of confusions

Some listener confusions might have a trivial explanation such as mispronunciation of the original token. In other cases, energetic masking was likely to be the principal factor. To further explore the relative frequencies of these and other types of confusion, a second experiment asked listeners to identify the 161 majority confusions discovered in expt. 1 with the speech and noise signals undergoing various types of manipulation.

3.1. Experiment 2: Speech/noise manipulations

Five types of manipulations were performed. In *clean*, words were presented without noise in order to identify mispronunciations. Words were also *time-shifted* in 7 steps relative to the noise to determine the extent to which the confusion relied on a precise alignment of speech and noise elements. Similarly, words were *f0-shifted* using STRAIGHT resynthesis [13] in 5 steps to identify those confusions which might have an origin in incorrect grouping of harmonic components. Two further sets of conditions examined the possibility that confusions could be predicted solely on the basis of energetic masking. The *glimpsing* manipulation applied a model of energetic masking [7] to determine those spectro-temporal regions of the word most likely to survive the noise. These regions were then resynthesised. In the *SNR-shifted* conditions the original signal-to-noise ratio was altered, both up and down. In the former case, the goal was to reveal more of the target word. The purpose of a decreased SNR was motivated by the finding that listeners are able to exploit differences in level to separate target and background [14]. On the other hand, a decrease in SNR might also remove a majority confusion by decreasing audibility, in which case listener responses should show less consistency.

In all, applying these manipulations to the majority confusions discovered in the first experiment resulted in 17 new blocks (table 5) each containing 161 stimuli which were identified by 9 listeners. Note that two of the conditions (0 ms time-shift and a F0 shift factor of 1.0) were replications of the earlier experiment, designed to test the robustness of majority confusions. The quiet condition was presented last to prevent listeners hearing clear exemplars of the tokens at an earlier stage.

Table 5: *Experimental manipulations of majority confusions.*

manipulation	values
clean	
glimpsing	
time-shifted	-160,-80,-14,-20,0,20,160 ms
F0-shifted	factors: 0.8,1.0,1.1,1.2,1.5
SNR-shift	-5,+5,+10 dB

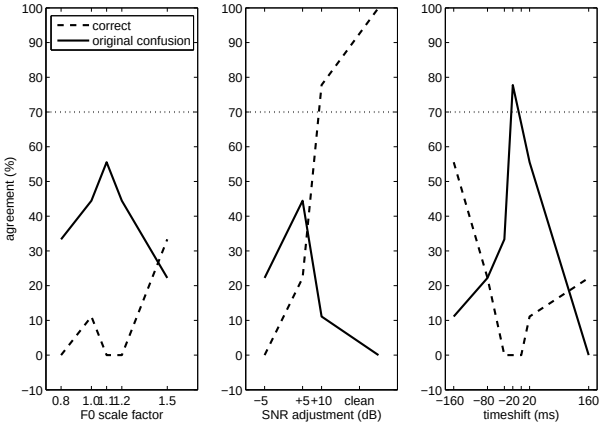

Figure 2: *Effect of manipulation for "wall" in babble noise. The percentage listener agreement for both the original confusion and for the correct response are indicated. The horizontal lines show the agreement level for the original confusion.*

3.2. Results

Responses were analysed by examining the number of listeners agreeing with the original confusion (from experiment 1) and the number identifying the correct word. While many cases resist a simple single-cause explanation, principal findings were:

(i) The test-retest rate for agreements on confusions was 82%. That is, while the agreement rate for some confusions was maintained (or increased), overall there was a fall in agreement. This may have been due to learning: the same set of words was used in each of the 17 conditions and while the quiet condition was presented last, it is likely that listeners learned about the words during the test, creating a closed-set task. Alternatively, since word order was different in experiment 2, this would suggest that some of the confusions in the first experiment came from semantic priming by recently-presented words.

(ii) 10% of items were misidentified due to ambiguous or incorrect pronunciation. In fact, 6% agreed with the original confusion but the other 4% revealed the original confusion only in noise, suggesting that the original ambiguous pronunciation was supplemented by a further reaction in noise.

(iii) Listeners identified the original confusion when resynthesised from glimpses on 24% of occasions. That is, pure energetic masking (EM) of parts of the target word was responsible for the observed confusion in those cases.

(iv) Overall, changes in fundamental frequency of the target word made little difference to agreements. However, in 18% of individual cases there was a clear F0 'tuning' effect, with small changes in F0 resulting in a change of the word perceived.

(v) Reports of the original confusion fell off slowly with target-background asynchrony, with about half of all confusions maintained even for the largest shifts of $\pm$ 160 ms. However, in specific cases even the smallest shifts of $\pm$ 20 ms could produce a significant change in majority response.

(vi) SNR shifts of $\pm$ 5 dB had little effect. Only the +10 dB improvement tended to cause a significant release from masking, though not to the levels seen in the glimpsing condition.

As an example, consider the case of [wall + babble$_6$ $\rightarrow$ small] from the Introduction. Here, resynthesis from glimpsing did not produce any correct responses (indicating that EM was not the sole cause of the confusion) and also did not result in any reports of the original confusion. Instead, listeners

tended to hear "mall" from the glimpses, suggesting that labiality and voicing were not masked. It appears that /s/ from the background was recruited. Fig. 2 illustrates the results of other manipulations to the speech and noise for the "wall" example. Here, we see an effect of changes in F0: as F0 of the target word is increased, there are more reports of the correct interpretation. Increasing SNR by +5 dB retains the original confusion. In the case of time-shifting, the striking finding here is the sharp 'tuning': at even quite small asynchronies the agreement with the original confusion is substantially weakened.

3.3. Discussion

Traditionally, the reduced intelligibility of speech in adverse conditions has been explained by the concept of masking, more recently refined into energetic (EM) and informational masking (IM) (e.g. [14]). EM concerns the loss of audibility of signal components due to the interaction of the target with an unwanted signal at the level of the auditory periphery while IM caters for everything else which reduces intelligibility (e.g. misallocation of signal components to target/background, cognitive load in attending to more than one source, interference from native language). One intriguing outcome of the current study is finding (iii) of section 3.2 that EM alone accounted for a minority of robust confusions. Other confusions would currently be categorised as resulting from IM.

Masking may not be the most appropriate way to interpret speech in noise. An alternative view is that speech and noise *react* in ways whose outcome depends on many factors, perhaps operating in sequence. Consider first that while noise is generally held to mask speech, the noise is itself masked in parts by energetic speech components. At an early stage, then, the time-frequency plane is fragmented into glimpsed and potentially audible components of both speech *and* noise. If speech components can be detected and then integrated into an ongoing utterance hypothesis, under the action of prior knowledge of words, then robust recognition is possible. However, there are many ways in which confusions can arise. The temporal, harmonic and energetic relationships among components such as patches of frication, formant transition or voice bars may well suggest groupings in which speech components are detached and allocated to the background noise, or background components attach themselves to the target. The weakening (e.g. by energetic masking) of a bond between parts of the original target speech paves the way for recruitment of background elements.

4. Towards a large-scale confusions corpus

The number of listeners, L, required to collect a corpus of N consistent confusions is given by (1)

$$L = \frac{k.N}{m.d_{agree}} \quad (1)$$

where k is the number of listeners required to screen each token, d_{agree} is the discovery rate for a desired strength of agreement between listeners and m is the number of tokens each listener is presented with. In the current study, $k = 10$, $m = 2400$ and d ranged from $d_{60} = 0.067$ to $d_{100} = 0.0038$, although after accounting for lack of test-retest consistency and poor pronunciations, $d_{60} \approx 0.05$ and $d_{100} \approx 0.0027$. A target corpus size of $N = 10^3$ would require more than 10^3 listeners each screening a large number of tokens. With adaptive token pruning techniques, k could be reduced somewhat, especially when high-levels of agreement are desired.

However, the scale of the task is well-suited to internet-based perception testing, with a reduced m and increased k to cater for uncontrolled variability. A large population test acts as an initial filter to generate confusion candidates to be assessed under formal listening conditions. A pilot test at `http://www.thebiglisten.org.uk` with $m = 50$ presentations and $k = 20 - 80$ listeners/token (after filtering for age, native-language, hearing impairment, ambient conditions and listening equipment) attracted responses from more than 2000 listeners and a full-scale test is planned.

Acknowledgements. Particular thanks go to Bruno De Cara and Usha Goswami for permission to use their lexical database and to Hideki Kawahara for the STRAIGHT resynthesis software. I have enjoyed productive discussions with Jon Barker, Odette Scharenborg, M.L. Garcia Lecumberri and Stuart Wrigley. This work was partially funded by the Sound to Sense (S2S) Marie Curie RTN.

5. References

[1] P. F. Assmann and Q. Summerfield, "The perception of speech under adverse acoustic conditions," in *Speech Processing in the Auditory System*, S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, Eds. Springer Handbook of Auditory Research, 2004.

[2] C. J. Darwin, "Listening to speech in the presence of other sounds," *Phil. Trans. Royal Society B*, vol. 363, pp. 1011–1021, 2008.

[3] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Am.*, vol. 19, pp. 90–119, 1947.

[4] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, pp. 318–326, 1979.

[5] K. Rhebergen, N. K. Versfeld, and W. A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *J. Acoust. Soc. Am.*, vol. 120, pp. 3988–3997, 2006.

[6] I. Holube and B. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Am.*, pp. 1703–1716, 100.

[7] M. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–1573, 2006.

[8] M. Regnier and J. Allen, "A method to identify noise-robust perceptual features: application for consonant /t/," *J. Acoust. Soc. Am.*, pp. 2801–2814, 123.

[9] M. Cooke and O. Scharenborg, "The Interspeech 2008 consonant challenge," in *Proc. Interspeech*, Brisbane, Australia, 2008.

[10] G. A. Miller and P. Nicely, "An analysis of perceptual confusions among some english consonants," *J. Acoust. Soc. Am.*, vol. 27, pp. 338–352, 1955.

[11] B. D. Cara and U. Goswami, "Similarity relations among spoken words: The special status of rimes in english," *Behavior Research Methods, Instruments, and Computers*, vol. 34, pp. 416–423, 2002.

[12] H. R. Baayen, R. Piepenbrock, and L. Gulikers, *The CELEX Lexical Database. Release 2 (CD-ROM)*. Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania, 1995.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, no. 27, pp. 187–207, 1999.

[14] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 109, pp. 1101–1109, 2001.