Speech + noise = confusion: misperceptions arising from the interaction of speech and babble masker components.

Attila Máté Tóth[1], *Martin Cooke*[2,1], *Jon Barker*[3]

[1]Language and Speech Lab, University of the Basque Country, Vitoria-Gasteiz, Spain
[2]Ikerbasque, Bilbao, Spain
[3] Department of Computer Science, University of Sheffield, UK
a.m.toth@laslab.org, m.cooke@ikerbasque.org, j.barker@dcs.shef.ac.uk

1. Introduction

It is well established that the effect of competing talkers on speech intelligibility cannot be explained in terms of energetic masking alone [1, 2]. The additional loss of intelligibility is often attributed to informational masking, but the individual processes that jointly result in this effect are less well understood. One aspect of informational masking can be formulated as the allocation problem. How do listeners segregate speech fragments from the target talker that survive energetic masking from a multitude of salient competing speech fragments? In this work we aim to identify the role fragment misallocation plays in generating confusions in babble noise by identifying listeners' most likely spectro-temporal segregation given their reported percept.

2. Methods

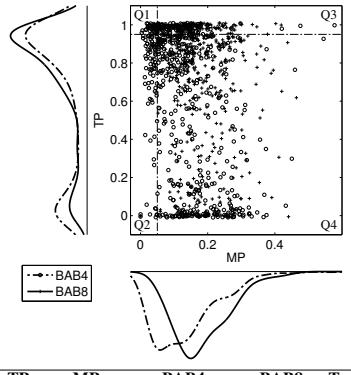
This study is based on a subset of the robust misperceptions corpus [3] involving babble maskers. In total 1057 misperceptions were studied in 4- and 8-talker babble. The most likely segregation was identified using the fragment decoding technique [4]. Originally developed for robust speech recognition, fragment decoding produces the best time-frequency segregation as a byproduct of the recognition process. To keep computation tractable the decoder requires the spectro-temporal plane to be partitioned into a number of regions such that each region originates from a single source. This fragment partitioning is done *a priori* using oracle knowledge to identify regions with positive local SNR for each source (including the target) in the mixture. The decoder returns the set of fragments that best explain a listener's reported percept through forced-alignment.

By considering the origin of each fragment in the best segregation we can characterise the type of misallocation. Allocation errors can involve incorporating masker fragments (Type I) or excluding available target fragments (Type II) from the speech hypothesis. We define Target Proportion (TP) as the combined area (in spectro-temporal pixels) of target fragments included in the best segregation divided by the total area of target fragments surviving energetic masking. Masker proportion (MP) is defined in a similar fashion for the fragments stemming from background voices. $MP > 0$ indicates a Type I error while $TP < 1$ indicates the presence of a Type II error.

3. Results and conclusion

Figure 1 shows all misperceptions in the corpus in terms of target and masker proportions together with the marginal densities for both maskers. The figure is partitioned into four quadrants

Figure 1: *Masker (MP) and Target Proportions (TP) for each misperception, along with their marginal densities.*



	TP	MP	BAB4	BAB8	Type
Q1	$> .95$	$< .05$	42 (7%)	4 (0.9%)	None
Q2	$\leq .95$	$< .05$	81 (13%)	3 (0.7%)	Type II
Q3	$> .95$	$\geq .05$	116 (19%)	100 (22%)	Type I
Q4	$\leq .95$	$\geq .05$	371 (61%)	340 (76%)	Type I & II

based on the type of allocation error involved. The table below gives the counts and proportions of misperceptions in each quadrant for both maskers. A key finding of the current analysis is that misallocation seems to play a role in most misperceptions. Cases involving no allocation errors (Q1), which are thus entirely attributable to other factors (e.g energetic masking) are relatively few. The large proportion of cases in Q3 and Q4, indeed 80% and 98% percent for BAB4 and BAB8 respectively, suggest that most confusions involve incorporating masker fragments to some degree. This involvement does not necessarily need to be large, as suggested by the marginal distribution of MP. At the same time the fact that $TP < 1$ for many misperceptions suggests that listeners are unable to make use of all available target fragments in most of the cases. At $TP \approx 0$ a cluster of cases represent an interesting subset of misperceptions where no target fragments were used. These cases most likely correspond to listeners reporting a salient word from one of the background talkers in its entirety. Most of these cases stem from BAB4, probably because the smaller fragments produced by BAB8 are unlikely to convey word level information. We conclude that misallocation plays an important role in generating misperceptions and as such can be considered an important component of informational masking.

Acknowledgements. The research leading to these results was funded from the European Community 7th Framework Programme Marie Curie ITN INSPIRE.

4. References

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

[2] 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.

[3] M. A. Tóth, M. L. García Lecumberri, Y. Tang, and M. Cooke, "A corpus of noise-induced word misperceptions for spanish," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. EL184–EL189, 2015.

[4] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Comm.*, vol. 45, pp. 45–25, 2005.