INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Misperceptions arising from speech-in-babble interactions

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

Abstract

The deterioration of speech intelligibility in the presence of other sound sources has been explained in terms of both energetic masking, which renders parts of the speech signal inaudible, and informational masking, in which audible components of the masker interfere with speech identification. The current study focuses on the role of a specific form of informational masking in which audible glimpses of both target and masker combine to produce an incorrect listener percept. We examine a corpus of word misperceptions in Spanish which occur when target words are combined with a babble masker. Glimpses originating in both the target and the masker are force-aligned to the reported misperceived word in order to identify the most likely acoustic evidential basis for the confusion. In this way, the degree of involvement of both target and masker can be quantified. In nearly all cases, the best explanation for the misperception involves recruiting evidence from the babble masker (type I error), and in more than 80% of the tokens some of the audible target evidence is ignored (type II error). These findings suggest misallocation of acoustic-phonetic material plays a significant role in the generation of speech-in-babble confusions.
Index Terms: informational masking, misperception

1. Introduction

Speech is almost always accompanied by extraneous sound sources, rendering its interpretation a non-trivial task. The challenge is greater still when the interfering signal is similar to the target, as in the case of competing speakers. The difficulties listeners face stemming from masker interference are two-fold. First, target speech cues can be occluded by more energetic portions of the masker, leaving the listener with incomplete target evidence, a phenomenon known as energetic masking. Second, listeners have to segregate those target components that survive energetic masking from a multitude of potentially similar acoustic fragments and integrate them into a coherent percept in order to interpret the intended message. This *allocation problem* can constitute a major component of non-energetic i.e., informational masking [1, 2, 3].

As a first step towards understanding the contribution of misallocation to informational masking, the aim of the current study is to measure the degree to which information from a speech-based masker is integrated into the reported percept, and the extent to which information from the target is omitted; both processes lead to the potential for a misperception. Fig. 1 shows an example confusion from the Spanish Confusions Corpus [4] which resulted when the target word "habrá" [there will be; /abra/] is reported as "acostumbrar" [to get used to; /akostumbrar/] by 9 listeners when presented in 4-talker babble at

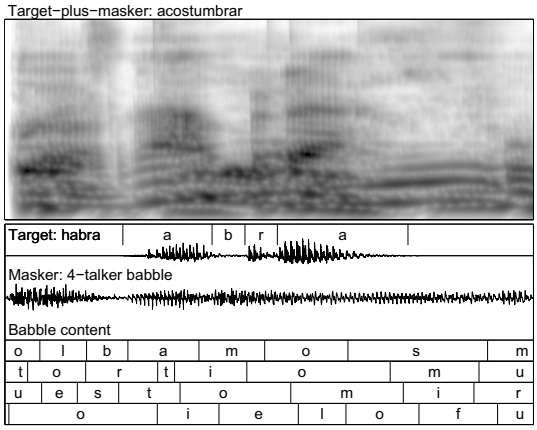

Figure 1: *An example robust misperception. Upper: Auditory spectrogram of a speech-in-babble mixture (see 2.2 for details). Lower: target and masker waveforms with phonemic content of target and each individual talker in the babble masker.*

a signal-to-noise ratio (SNR) of -0.8 dB. Phoneme transcriptions for the target word and the four individual babble tracks are also shown. While the sequence /bra/ is shared by the target and confusion, it is evident that additional processes are needed to explain the misperception. First, there is some evidence for the incorporation of babble segments corresponding to /o/, /st/, /m/, and later /r/ in temporal locations which are consistent with their inserted positions in the perceived word /akostumbrar/. In these cases it is possible that some part of the segment is energetically-dominant in the mixture during the relevant intervals. Second, elements of the initial /a/ of the target word may have been sufficiently masked to render its identity uncertain. Finally, for the initial segments /ak/, while there is no equivalent segment in the babble, segments with vowel and voiceless plosive characteristics occur at the right place. Here, listeners may be using lexical information to hypothesise "acostumbrar" in the absence of a lexical candidate congruent with the acoustic evidence. In sum, it is plausible that the confusion arises through an interplay of energetic masking, the incorporation (i.e. misallocation) of phonetic detail from the masker, and the failure to include certain details from the target itself.

To examine the extent of target-masker misallocations in the Spanish Confusions Corpus [4], we adopt here a 'microscopic' perspective [5, 6, 7, 8, 9, 10] in which each individual misperception is independently analysed for evidence of its composition. We use a tool from robust automatic speech recognition [11] to identify those time-frequency regions in the

http://dx.doi.org/10.21437/Interspeech.2016-24

speech-in-babble mixture that jointly provide the most likely explanation for the reported percept. Section 2 details the theoretical basis, and practical implementation, of the approach we use to force-align a set of fragmentary evidence from multiple speech signals to the misperceived word. The outcome of applying these techniques to over 600 misperceptions from the corpus is described in Section 3.

2. Identifying misallocation errors

2.1. Theory

The speech-in-noise recognition problem involves finding the most likely word sequence $\widehat{W}$ given noisy observations Y

$$\widehat{W} = \underset{W}{\operatorname{argmax}}\, P(W|Y) \tag{1}$$

The glimpse decoder [11] provides a formalism for recognising speech in noise, in which a useful by-product of recognition is the set of glimpses that forms the evidential basis for the most likely speech hypothesis. Here, a glimpse is understood as a connected spectro-temporal region where one source is energetically-dominant throughout the region. The current study uses the glimpse decoder to answer the question: given a misperceived word, which set of target *and* masker glimpses best accounts for that word? The solution can be expressed as a simultaneous search over words and segregations S to find the most likely word/segregation pair, given a set of glimpses G

$$\widehat{W}, \widehat{S} = \underset{W, S \in \mathcal{P}(G)}{\operatorname{argmax}}\, P(W, S|Y, G) \tag{2}$$

where S is understood as the set of partitions – i.e., members of the powerset of glimpses, $\mathcal{P}(G)$ – of the observations into those belonging to the target speech and those belonging to the masker. Via forced-alignment, the glimpse decoder can be used to find the most likely segregation hypothesis given the confused word. Implemented using hidden Markov models (HMMs), we look for the most likely HMM state sequence $\widehat{Q}$ and segregation given W, G and Y

$$\widehat{S}, \widehat{Q} = \underset{S,Q}{\operatorname{argmax}}\, P(Q, S|W, Y, G) \tag{3}$$

Here, we are only interested in the segregation component, $\widehat{S}$, since we hypothesise that this defines the set of glimpses from both target and masker that listeners are most likely to have used in making their response.

2.2. Input representation

Target speech and masker waveforms are summed at a given SNR to form the input to a model of the auditory periphery, leading to an auditory spectrogram (e.g., top panel Fig. 1), a spectro-temporal representation of auditory nerve excitation (the noisy observation Y in Eq. 3). This representation is computed by passing the mixture signal through a bank of 39 gammatone filters with centre frequencies in the range 50–8000 Hz equally-spaced on an ERB-rate scale, followed by extraction of the instantaneous Hilbert envelope at the output of each filter, which is subsequently temporally-smoothed and log-compressed prior to downsampling at 100 Hz.

2.3. Generation of glimpse set G

A glimpse is defined as an 8-connected spectro-temporal region originating from a single source, and which possesses a positive local SNR throughout the region. In the current study, the mixture consists of the target speech and four known background speech signals that constitute the babble masker. For a given source s_j in the mixture, we compute separate auditory spectrograms for s_j and the sum of the remaining $N-1$ signals. Auditory spectrograms are then compared to identify the regions where the jth source is dominant, i.e.

$$f(s_j) > f\Big(\sum_{i=1, i\neq j}^{N} s_i\Big) \tag{4}$$

where the function f maps a time-domain signal to the auditory representation defined in Section 2.2. The comparison is done for each time-frequency 'pixel'. A glimpse from source s_j is defined as a connected spectro-temporal region satisfying the inequality above. This process is repeated for each of the sources in the mixture. The set of fragments obtained for each source are combined to form G, the set of glimpses input to the decoder

$$G = G_T \cup G_M \tag{5}$$

where

$$G_M = \bigcup_{i=1}^{N-1} G_{M_i} \tag{6}$$

G_T denotes glimpses originating in the target source and G_{M_i} those stemming from the ith babble component. As in [5], small glimpses (here, those with area < 6 time-frequency pixels) are eliminated from G. The upper panel of Fig. 2 shows glimpses of target and masker components for the example token of Fig. 1. Regions in white correspond to spectro-temporal locations where no source is dominant.

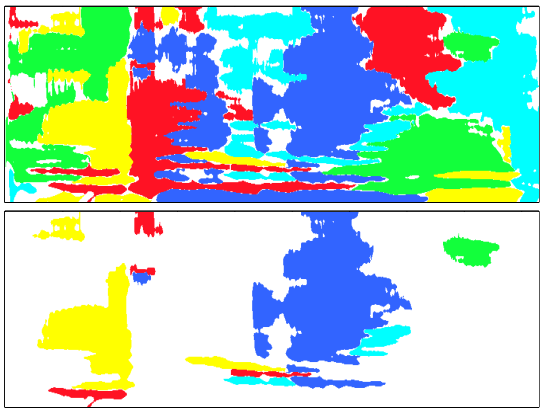

Figure 2: *Glimpse formation and decoding for the example of Fig. 1. Upper: glimpses from each source in the mixture (dark blue: target speech glimpses, G_T; remaining colors correspond to glimpses of the individual babble channels, collectively G_M). Lower: most likely segregation $\widehat{S}$ in which glimpses arise from both the target word and the masker.*

2.4. Selecting $\widehat{S}$ through forced alignment

The auditory spectrogram of the speech-plus-babble mixture Y, the glimpse set G, along with the confusion W reported by listeners, serve as input to the forced alignment process which produces the segregation hypothesis $\widehat{S}$, i.e. the subset of glimpses

in G which best explain confusion W. The lower panel of Figure 2 indicates the glimpses assigned by the decoder to the confusion W. In addition to the incomplete set of glimpses of the target, those stemming from one or more of the babble components are also included in the segregation hypothesis that best explains the misperception, illustrating how the misallocation of signal components might play a role in the formation of the confusion.

3. Evaluation

3.1. Misperceptions corpus

The dataset of speech misperceptions used in the current study is a subset of the Spanish Confusions Corpus [4], which consists of 1–3 syllable Spanish target words embedded in one of five types of masking noise at a range of SNRs. Here, we analyse 610 misperceptions which resulted from the 4-talker babble masker. In all cases at least 6 out of 15 listeners reported the same confusion (mean: 8.1). Table 1 lists some example confusions from this dataset.

target word	misperception	listeners	TP	MP
antes	alcohol	14	0.33	0.09
cerdo	entramos	12	0.84	0.13
comenzar	buscar	9	0.02	0.23
socios	sucios	8	0.98	0.07
sección	disección	7	0.54	0.07
casada	casarse	6	0.11	0.37
litros	comen	6	0.49	0.10

Table 1: *Some Spanish word confusions in 4-talker babble. The number of listeners out of 15 reporting the same confusion is shown in the column 'listeners'. Target and masker proportions (Sec. 3.3) determined by the glimpse decoder are also reported.*

3.2. Recogniser

Acoustic models for the glimpse decoder were speaker-independent 3-state triphone models trained on over 12 000 instances of Spanish words. 10-component Gaussian mixtures, with model- and state-level tying, were used to represent the feature distribution of each state. Triphone models were expanded into models of Spanish words for each of the 610 word confusions reported by listeners.

3.3. Target and masker proportion

Since the source (target or masker) of each glimpse is known, it is possible to quantify the involvement of both the target and masker in the most likely segregation $\widehat{S}$ for each misperception. We define the Target Proportion (TP) as

$$TP = \text{count}(\widehat{S} \cap G_T)/\text{count}(G_T) \quad (7)$$

where the 'count' function computes the total number of spectro-temporal pixels in the auditory spectrogram for a given set of glimpses. TP denotes the total area of target glimpses included in the most likely segregation hypothesis $\widehat{S}$ normalised by the total area of available target glimpses in the set G. The Masker Proportion (MP) is defined similarly:

$$MP = \text{count}(\widehat{S} \cap G_M)/\text{count}(G_M) \quad (8)$$

TP and MP provide a way to analyse the degree to which target and masker material contribute to each confusion in the corpus, and enable an analysis of the types of allocation errors made by listeners.

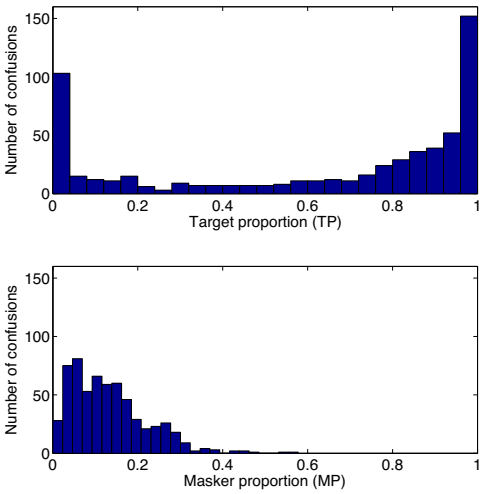

Figure 3: *Distribution of target and masker proportions across confusions.*

Figure 3 shows the distribution of target and masker proportions, MP and TP, for the 610 misperceptions analysed here. In around 150 cases, around a quarter of the total, all, or nearly all, of the available target glimpses are used in the misperception, according to the decoder; half of the confusions make use of at least 80% of the target glimpses. However, about 100 misperceptions use no material from the target at all. In some of these cases it is posible that listeners are hearing an entire word from the background babble. On average, misperceptions make use of 13% of masker glimpses, and in only 2% of cases is more than a third of masker material used. This is not surprising since the masker consists of 4 talkers in parallel, any one of which could in principle contribute sufficient phonetic information to create a confusion.

The proportion of time-frequency pixels taken up by the glimpses in the most likely segregation $\widehat{S}$ relative to the area of the target glimpses is shown in Fig. 4. On average the best hypothesis incorporates 23% more of the spectro-temporal plane than occupied by target glimpses, suggesting that the decoder frequently makes use of a substantial amount of information from the background babble.

Figure 5 shows the joint distribution of target and masker proportion, demonstrating that, at least from the decoder's point of view, a listener's interpretation makes use of both target and

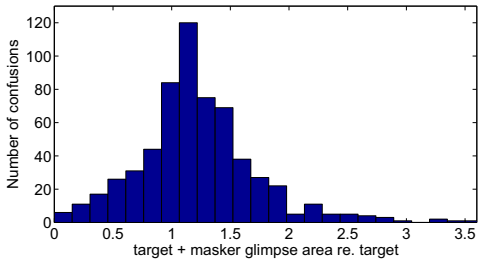

Figure 4: *Distribution of the area of glimpses in $\widehat{S}$ relative to the area of glimpses in G_T.*

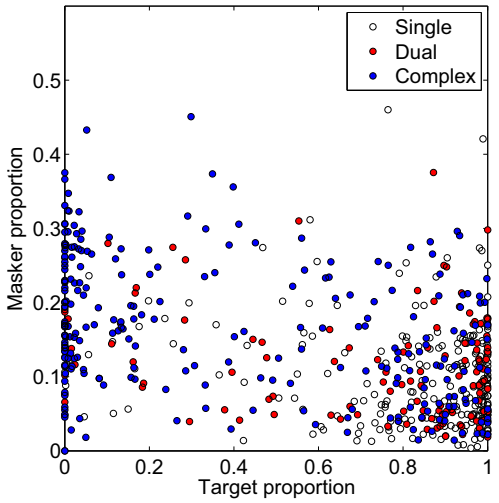

Figure 5: *Joint distribution of target and masker proportion, along with target-confusion phoneme distance class.*

masker glimpses in the majority of cases. That is, confusions arise due to both Type I errors i.e. the incorporation of masker material, as quantified by MP, and Type II errors i.e. failing to utilise target material, as quantified by $1 - TP$.

3.4. Relationship to phoneme distance

Correlations were computed between target and masker proportions and a measure of the phonemic distance between the target word and the reported confusion. First, phoneme sequences for each target and confusion were aligned using dynamic programming string alignment using penalties of 7 for insertion and deletion, 10 for substitution of vowel for vowel or consonant for consonant, and 20 for substitution of vowel for consonant (the latter penalty ensures that no such substitutions take place; instead, they are treated as an insertion and a deletion). Then, the phoneme distance was calculated as the number of phonemes inserted, deleted or substituted divided by the length of the alignment sequence, to obtain a value in the range 0 (no changes, which never occurs for confusions) to 1 (no phonemes in common). The phoneme distance computed in this way is negatively-correlated $[r = -0.64, p < 0.001]$ with target proportion, and positively-correlated $[r = 0.44, p < 0.001]$ with masker proportion.

Figure 5 also encodes three classes of phoneme distance, using a classification scheme similar to that of earlier 'slips of the ear' studies [12, 13]: *single* cases are those involving the deletion, insertion, or substitution of a single phoneme segment (e.g., socios $\mapsto$ sucios); *dual* cases correspond to changes involving a pair of segments (e.g., sección $\mapsto$ disección); all others are denoted *complex* (e.g., antes $\mapsto$ alcohol). While *single* cases tend to involve high values of target proportion, there remains a substantial number of cases where the target proportion is reduced. Conversely, *complex* cases typically correspond to low values of TP, but again there is a significant spread. In many such cases the misperception appears to be due to the masker material overriding the target signal entirely ($TP = 0$). Similarly, the amount of masker involved for all three classes is highly-variable across tokens. These findings suggest that while phoneme distance is correlated with target and masker proportion across the corpus, this kind of segmental metric alone is a poor predictor of the involvement of target and masker glimpses for any given misperceived token.

4. Discussion

When listening to speech in speech-like maskers, listeners have to confront not only the energetic masking effects of the background signal, which acts to reduce the availability of target speech cues, but also the informational masking effects which arise when the background has the potential to contribute misleading information, leading to a difficulty in allocating cues to the target source. This study is a first attempt to quantify the scale of the misallocation component of informational masking using a decoder which finds the most likely set of glimpses (regions escaping energetic masking) to contribute to the majority confusion reported by listeners.

Based on this model, one striking outcome (illustrated by the near absence of points on the $MP = 0$ axis of Fig. 5) is that in only a handful of cases is no information at all from the masker allocated to the overall misperception. Nearly all confusions are best-explained by incorporating both masker and target glimpses, or via masker glimpses alone ($TP = 0$).

Phoneme distance between the target and confused word explains some proportion of the misallocation effect, but the spread of individual cases is too wide for a segmental metric such as this to be a robust predictor. This is likely to be caused by the strictly temporal nature of the segmental metric. In contrast, the glimpse decoder takes into account the spectro-temporal decomposition of the signal. It is possible that more sophisticated forms of alignment which take into account the segmental constituents of the babble itself (as shown in Fig. 1) may lead to better predictions.

The results of the current study depend on the model of speech perception embodied in the glimpse decoder. This model has some limitations: for example, currently it does not take into account word frequency or familiarity effects which might affect the reported confusion in certain cases. Nevertheless, the proposed approach provides a microscopic modelling framework for evaluating target and masker interactions in speech perception which can be developed to incorporate additional factors which influence the reported percept. Future work will test the validity of the predictions by measuring listeners' responses to stimuli based on resynthesis from the proposed most likely set of glimpse components.

5. Conclusions

This study examined the role of signal component misallocation in misperceptions arising from babble maskers. Glimpse decoding, a robust speech recognition technique producing the most likely speech segregation as part of the recognition process, identified those spectro-temporal regions in the mixture that listeners were likely to be treating as evidence for the target speech. Most confusions were best explained by incorporating some of the masker material into the speech hypothesis, suggesting that misallocation plays a significant role in generating confusions in situations involving competing talkers.

6. Acknowledgements

We thank María Luisa García Lecumberri for earlier phonetic analysis of the confusions and Yan Tang for software support in the collection of the corpus.

7. 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] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, no. 6, pp. 4007–4018, 2006.

[4] 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.

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

[6] M. S. Regnier and J. B. Allen, "A method to identify noise-robust perceptual features: Application for consonant /t/," *J. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 2801–2814, 2008.

[7] 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.

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

[9] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *J. Acoust. Soc. Am.*, vol. 127, no. 4, pp. 2599–2610, 2010.

[10] L. Varnet, K. Knoblauch, F. Meunier, and M. Hoen, "Using auditory classification images for the identification of fine acoustic cues used in speech perception." *Frontiers in Human Neuroscience*, vol. 7, 2013.

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

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