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

Undoing misperceptions: a microscopic analysis of consistent confusions through signal modifications

Attila Máté Tóth[1], *Martin Cooke*[2,1]

[1]Language and Speech Lab, University of the Basque Country, Vitoria-Gasteiz, Spain
[2]Ikerbasque, Bilbao, Spain

a.m.toth@laslab.org, m.cooke@ikerbasque.org

Abstract

Consistent confusions — word misperceptions reported in an open set task with a high agreement across listeners — can be especially valuable in understanding the detailed processes underlying speech perception. The current study investigates the origin of a set of consistent confusions collected in a variety of masking conditions, by applying signal-level modifications to the stimuli eliciting the confusion, and subsequently reevaluating listeners' percepts. Modifications were selected to provide release from either the energetic or the informational component of the maskers and involved manipulations of signal-to-noise ratio, fundamental frequency, and resynthesis of the noise-mixture in glimpsed regions of the target speech. Increasing signal-to-noise ratio and glimpse resynthesis showed the expected release from energetic and informational masking respectively. However, manipulations targeting informational masking release, including fundamental frequency modification, affected a surprisingly high number of confusions stemming from energetic maskers. The degree of fundamental frequency shift did not have a significant effect on the response patterns observed. Around 30% of confusions can be explained solely based on the information contained within the target glimpses surviving energetic masking, while for the rest of the cases additional factors, such as recruitment of information from the masker, appear to be involved.
Index Terms: speech perception,word confusions, noise

1. Introduction

When listening to speech, masking by extraneous sound sources can interfere with identification of the intended message. Past research on speech perception has focused on quantifying the intelligibility loss due to masking and other adverse conditions in terms of average recognition rate of the target utterance, with intelligibility quantified using univariate metrics such as word error rate (WER) or speech reception threshold (SRT). Several metrics have been proposed to predict 'macroscopic' intelligibility for a variety of conditions [1, 2]. While these predictions have become increasingly accurate, macroscopic models fail to provide a detailed explanation of how the interference from the masker alters a listener's percept on a token-by-token basis. More recently, 'microscopic' models have started to appear which aim to predict responses at the level of individual tokens [3, 4, 5]. Such models examine individual speech-noise interactions and the resulting percept in order to locate distinctive speech cues in the time-frequency representation of the mixture or describe the masking processes that lead to the reported percept [6, 7, 8]. Unravelling the mapping between noisy stimulus and the resulting percept can further our understanding of how speech is processed in everyday adverse conditions.

One of the main challenges faced by microscopic approaches is individual variability, as argued in [9], who examined the sources of variability in consonant perception. Potential sources of variability were examined both at the receiver, by investigating within- as well as across-listener variability; and at the source, looking at the effect of different acoustic realizations of phonetically identical speech tokens and noise tokens of the same masker type. They found that across-talker differences produced the highest source related variability, followed by differences in within-talker articulation and the masking waveform. Regarding receiver related variability, large inter-listener differences were found, while the intra-listener variability was quite small but proportional to the adversity of the condition. As the above study suggests, different listeners may respond differently to the same stimulus, especially in adverse conditions where uncertainty over the percept is high. Some studies that have investigated speech perception in nonsense syllables account for this listener-related variability by presenting the distribution of listener responses across experimental conditions [10, 9]. An alternative approach is to build a corpus of speech-in-noise tokens where listener agreement is high i.e., when faced with the same stimulus, a majority of listeners report the same confusion. The stimuli underlying such consistent confusions can then be further dissected and processed to examine the cause of the misperception. This method forms the basis for the current study.

In [11] we presented a large-scale elicitation of over 3000 consistent confusions in noise. In the current study we used a subset of this corpus to search for the basis for the majority percept reported by listeners. The goal was to determine the likely cause of each misperception by examining the agreement or otherwise with the original confusion of listeners' responses to modified stimuli. Following [12], three forms of manipulation were investigated. In one, the time-frequency regions where the target word is energetically-dominant were resynthesised in order to evaluate whether the confusion was more likely to be due to energetic masking, informational masking, or a mixture of the two. To confirm the possibility of energetic masking, a second form of manipulation modified the overall signal-to-noise ratio (SNR). Finally, the sensitivity of the confusion to auditory grouping based on common fundamental frequency (F0) was assessed by shifting the F0 of the target word. As an additional control measure, the robustness of confusions was assessed using a control condition with unmodified stimuli. The following section motivates the signal modifications explored in this study.

http://dx.doi.org/10.21437/Interspeech.2016-1030

2. Modifying speech-in-noise confusions

2.1. SNR increase

Increasing the speech-to-masker SNR leads to a straightforward reduction in energetic masking of the target word. The effect of a 3 dB increase in SNR was evaluated here (see second panel of Figure 1). We hypothesise that if listeners respond with fewer instances of the prior confusion and more of the correct target word following this manipulation, the original confusion is likely to have been caused by energetic masking.

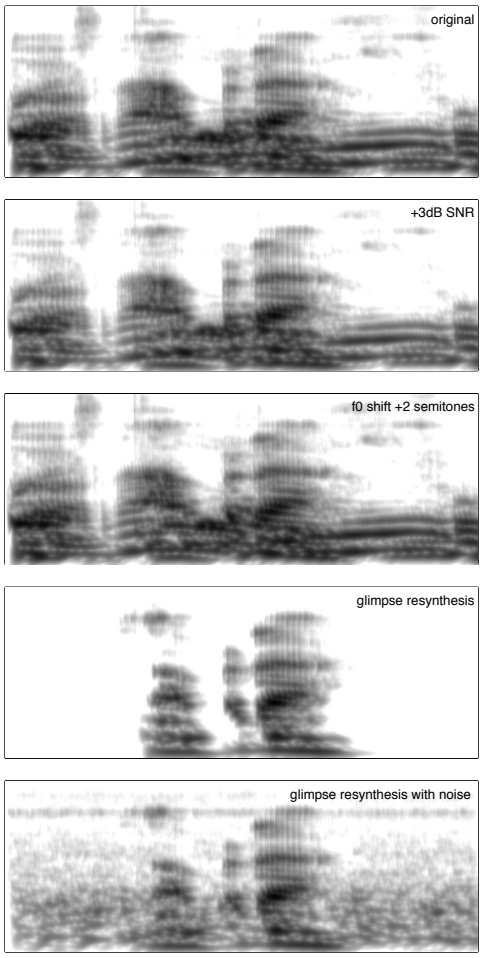

Figure 1: *Auditory spectrograms showing the original speech-in-noise token (target word "habrá", majority confusion "acostumbrar") and some of the experimental manipulations described in the text.*

2.2. Resynthesis from glimpses

One way to assess the extent to which listeners are utilising information from the masker in reporting a confused percept is to resynthesise just those parts of the target signal that are deemed to survive energetic masking. In this way, no parts of the masker are presented to the listener. We hypothesise that if listeners continue to report the original confusion following resynthesis, the misperception is likely to have been caused by energetic masking. Listeners reporting the correct target word instead implies that sufficient information exists in the target glimpses. We interpret this as a consequence of removing the informational masking effect of those parts of the stimulus not belonging to the target, a form of release from informational masking. A third possible outcome is that listeners report something other than the original confusion or the target word. Here, it seems likely that both energetic and informational masking have a significant role in causing the original confusion.

Resynthesis from glimpses is performed by first determining target glimpses — spectro-temporal regions in an auditory representation where the target word is more energetic than the masker [3] — then passing the speech-plus-masker signal through a zero-phase gammatone filterbank, selectively gating glimpsed regions in each frequency channel, and summing across channels. The zero-phase filterbank ensures that the resynthesised signal possesses the same phase structure as the original signal, and is implemented following [13] by filtering the signal, time-reversing the output, filtering the signal for a second time, and time-reversing the output again. Two experimental conditions were tested, one in which glimpses alone are presented (4th panel, Figure 1), the other where a speech-shaped noise is added at 12 dB SNR to the un-glimpsed spectro-temporal regions (lower panel, Figure 1).

2.3. F0 shift

Confusions might result from allocating parts of the masker to the speech hypothesis. One way in which this is thought to be catalysed in listeners is via similarity in F0 between target and masker. For instance, it is more difficult to identify simultaneously-presented vowels if they have the same F0 [14, 15]. By modifying the relationship between the F0 of the target and masker, we hypothesise that any confusions that revert back to the correct target word are dominated by informational masking. Four conditions were tested, corresponding to shifting the F0 of any voiced regions of the target word by -1, $+1$, $+2$ and $+3$ semitones. Larger shifts were avoided as they tended to change the perceived gender of the male target talkers. The target signal was manipulated as it has at most a single F0, while the maskers have no F0 or multiple concurrent F0s. STRAIGHT [16] was used to achieve F0 shifts. The 3rd panel of Figure 1 depicts the $+2$ semitone case.

3. Perception experiment

3.1. Stimuli

A subset of 800 tokens was selected from the Spanish Confusions corpus [11] for this experiment. Each token consists of a single word spoken by one of four talkers (two male, two female), centrally embedded in one of five masker types: speech shaped noise (SSN), speech modulated noise (BMN1), 3-talker babble modulated noise (BMN3) and 4- and 8-talker babble (BAB4 and BAB8). Further details of corpus elicitation, including listeners and SNR ranges, can be found in [11].

Tokens were selected randomly from the corpus after excluding those cases where the target and confusion differed in the insertion, deletion or substitution of a single phoneme, since many such cases were likely to be influenced by acoustic similarity, especially in an inflected language like Spanish (e.g., gender: "guapa/guapo"; number: "casa/casas"; person/tense: "veré/verá"). Tokens selected for the current experiment were balanced across the four talkers and five masker types.

Tokens were presented in the 8 conditions listed in Table 1 based on the manipulations described in Section 2.

manipulation	condition(s)
none	original (control)
SNR increase	SNR increased by 3 dB
F0 shift	-1, 1, 2, 3 semitones
Glimpse resynthesis	target glimpses alone
	target glimpses+low-level noise

Table 1: Experimental conditions

3.2. Listeners

72 monolingual Spanish or bilingual Spanish-Basque adults took part in our experiment after screening for hearing loss at 20 dB HL. Participants gave written consent and were paid for their participation.

3.3. Procedure

Of the 6400 unique stimuli (800 tokens x 8 manipulations), each listener screened 1600 stimuli in total in two 1 hr sessions separated by a break of at least an hour. The 3 dB increase, control and two resynthesis conditions were screened in the first session, and those involving F0 shifts in the second session. The experiment was conducted using custom MATLAB software in a sound-attenuated studio booth over Sennheiser HD 380 Pro headphones. Listeners were instructed to identify a single word after hearing each stimulus exactly once, and to type in their first impression. Stimuli were blocked by target talker and masker type, resulting in 20 blocks of 40 stimuli in each session. Prior to each block listeners heard four practice stimuli at a high SNR to familiarise themselves with the voice of the target talker and masker type, for the conditions where the masker was present. Block order was randomised first on speaker followed by masker so that blocks of a same target speaker are presented successively, in order to minimise the switching between target talkers. The order of stimulus presentation in each block was randomised. Each individual stimulus (i.e. token-condition combination) was heard by at least 15 listeners.

4. Results

4.1. Test-retest rate

The majority confusion in the unmodified condition matched the majority response in the original experiment in 636 cases (79.5%) of the sub-corpus used in this study. To ensure that the subsequent analyses are based on highly-robust confusions, we additionally insist upon a minimum listener agreement of 40% as in [11], which reduces the number of tokens to 505 (63%). The remaining analyses are based on this subset. We employ the following terminology to describe the relationship between the original confusion and the majority response elicited by the modified stimulus: listeners either MAINTAIN the original confusion, REVERT to the correct target word, or produce OTHER responses.

4.2. SNR increase

Following a 3 dB increase in the target relative to the masker, listeners MAINTAIN the original confusion in 339 cases (67.1%), REVERT to the correct target in 127 (25.2%) cases, and produce OTHER responses to the remaining 39 (7.7%) tokens. Figure 2 shows the breakdown of these responses across masker type. It is evident that the largest proportion of reversions to the correct target word occur for the SSN (36%) and BMN3 (33%) maskers. The response categories differed significantly across across masker type for the SNR increase condition [$\chi^2(8, N = 505) = 28.98, p < .001$].

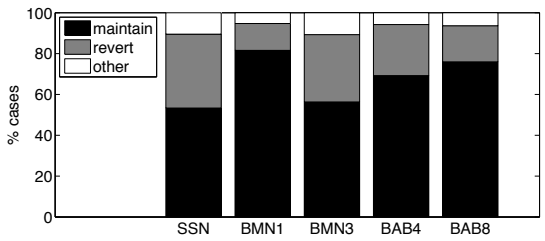

Figure 2: *Percentages of* MAINTAIN, REVERT *and* OTHER *responses per masker type for the SNR increase condition.*

4.3. Glimpse resynthesis

Figure 3 shows the distribution of MAINTAIN, REVERT, and OTHER responses following glimpse resynthesis with and without low-level noise, as a function of masker type. Here we see a striking difference between pure energetic maskers (SSN, BMN1, BMN3) and those which contain speech and hence also have an informational masking component (BAB4, BAB8). The former group have a larger proportion of cases where the original confusion is maintained, while babble-based maskers lead to many REVERT cases. In order to determine the significant associations between resynthesis condition, masker and response type, a hierarchical log-linear analysis [17] was conducted. A backward elimination procedure was used to select the best model. Model fit is assessed with the likelihood ratio chi-square test, which tests the difference between the observed counts and those predicted by the model, thus non-significant p values are associated with good models. The best model [$G^2(12) = 4.27, p = .98$] included significant interactions between masker and response type [partial $\chi^2(8) = 212.00, p < .001$], as well as response type and resynthesis condition [partial $\chi^2(2) = 10.61, p < .01$] and the corresponding main effects, out of which masker [partial $\chi^2(4) = 14.15, p < .01$] and response type [partial $\chi^2(2) = 44.03, p < .001$] were significant while resynthesis condition was not [partial $\chi^2(1) = 0, p = 1$]. The former significant interaction supports the differences of response categories across masker type mentioned above. The latter shows that the distribution of responses are significantly different for the two resynthesis conditions, with the noise in the gaps condition seeming to contribute more MAINTAIN responses.

4.4. F0 shifts

Figure 4 shows responses for the F0 shift cases. Overall, just over half of the 505 robust tokens (274; 54.3%) were unaffected by shifts in F0. In the remaining 231 cases at least one of the shifts had an effect. We used a log-linear analysis as for the resynthesis conditions to determine associations of masker, response type and the amount of F0 shift. The best model [$G^2(45) = 38.11, p = .76$] included a significant interaction between masker and response type [partial $\chi^2(8) = 90.65, p < .001$] and the corresponding main effects for masker [partial $\chi^2(4) = 28.29, p < .001$] and response type [partial $\chi^2(2) = 1691.45, p < .001$]. The factor F0 shift was not included in the best fitting model, indicating no significant main or interactive effects. These results show that the

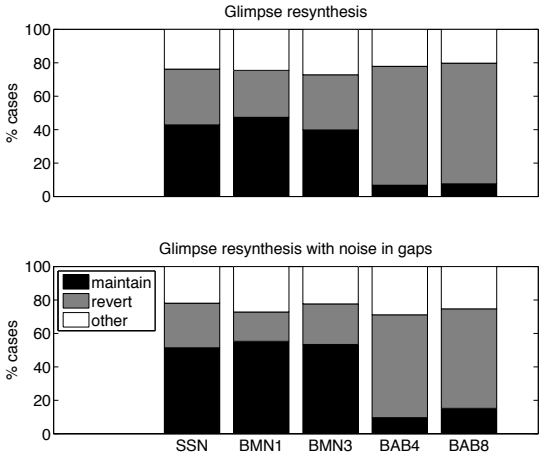

Figure 3: *Percentages of* MAINTAIN, REVERT *and* OTHER *responses per masker type for the glimpse resynthesis conditions.*

number of MAINTAIN, REVERT and OTHER responses did not differ significantly as a function of F0 shift. However, the differences in responses across masker type as shown in Figure 4 were found significant with the largest numbers of REVERT responses seen for the SSN and BMN3 maskers. Intriguingly, this is the same pattern as observed in the SNR increase case.

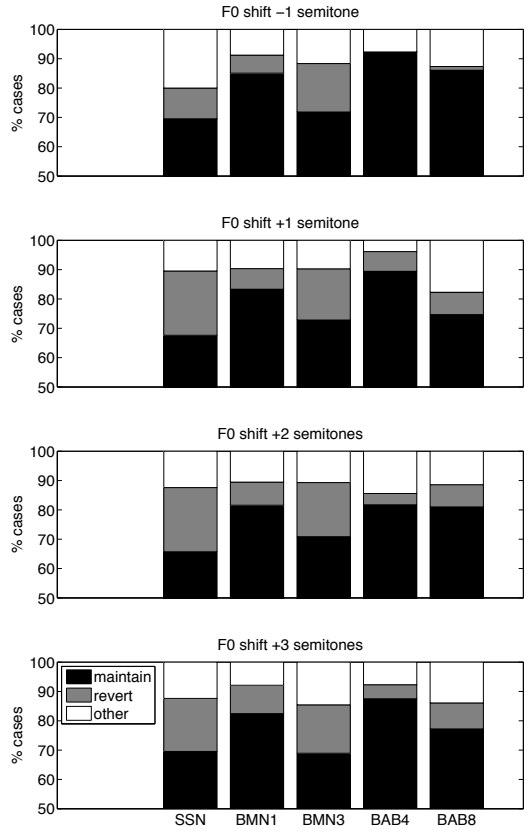

Figure 4: *Distribution of responses as a function of F0 shift. Note the change in axis range.*

5. Discussion

In Section 2 we hypothesised that confusions reverting to the target responses in the +3 dB SNR condition are likely to have been caused by energetic masking, as this manipulation results in a direct energetic masking release. This hypothesis is supported by the fact that the largest release from masking in this condition was observed for two of the three noise-based maskers (SSN and BMN3). On the other hand, BMN1 showed the smallest proportion of REVERT cases, perhaps since a 3 dB increase in SNR is not enough to bridge the gap between noise and speech energy in the masked regions caused by the large temporal modulations of this masker. The hypothesis that for the resynthesis conditions REVERT and MAINTAIN responses correspond to confusions caused by informational and energetic masking respectively was also supported by the distribution of response types across masker as shown in Section 4.3. From the modifications considered, F0 manipulations had the smallest effect on the confusions, which is in agreement with [12] who also found that F0 changes had little effect on the original percept. Contrary to expectations, we did not find evidence that F0 manipulations provide more release from informational masking compared to energetic masking.

Interestingly, both resynthesis and F0 modifications resulted in many REVERT cases for the noise-based maskers. Since in the former condition listeners have access to the same target glimpses as in the control condition — the only difference being the presence or the absence of the masker in the unglimpsed regions — these confusions cannot be attributed to simultaneous energetic masking. A potential explanation of some of these cases is forward masking, which is absent in the resynthesis condition. The large proportion of REVERT cases in the F0 conditions in SSN and BMN3 might also be attributed to a release from this same effect, as forward masking exhibits a sharper tuning curve than simultaneous masking [18], so this type of masking is more likely to be affected by shifts in F0. Perhaps the similarity between the F0 shift and 3 dB response patterns observed in Section 4.4 can also be attributed to confusions caused by forward masking as the latter manipulation is also expected to provide release from this effect.

6. Conclusion

In this study a set of signal manipulations were applied to speech-noise interactions that previously resulted in consistent confusions in order to uncover their cause. Modifications were mostly successful in achieving the expected type of masking release, providing a way to separate confusions based on whether they originated from informational or energetic masking. Our findings suggests that for the latter, both simultaneous and forward component might play a role in generating consistent confusions. Future work will focus on investigating speech-masker interactions on a spectro-temporal level using models of speech segregation to explain the resulting confusion.

7. Acknowledgements

The research leading to these results was funded from the European Community 7th Framework Programme Marie Curie ITN INSPIRE (Investigating Speech Processing in Realistic Environments).

8. References

[1] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Comm.*, vol. 52, pp. 678–692, 2010.

[2] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.

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

[4] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Comm.*, vol. 49, pp. 402–407, 2007.

[5] T. Jürgens 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.

[6] 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.

[7] 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.

[8] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of american english fricative consonants in noise," *J. Acoust. Soc. Am.*, vol. 132, no. 4, pp. 2663–2675, 2012.

[9] J. Zaar and T. Dau, "Sources of variability in consonant perception of normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1253–1267, 2015.

[10] J. B. Allen, "Consonant recognition and the articulation index," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2212–2223, 2005.

[11] 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.

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

[13] M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Stanford University, 1985.

[14] M. T. M. Scheffers, "Simulation of auditory analysis of pitch: An elaboration on the dws pitch meter," *J. Acoust. Soc. Am.*, vol. 74, no. 6, pp. 1716–1725, 1983.

[15] J. Bird and C. Darwin, "Effects of a difference in fundamental frequency in separating two sentences," *Psychophysical and physiological advances in hearing*, pp. 263–269, 1998.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 34, pp. 187 – 207, 1999.

[17] A. Agresti, "Loglinear models for contingency tables," in *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., 2006, pp. 204–243.

[18] B. C. J. Moore and B. R. Glasberg, "Comparisons of frequency selectivity in simultaneous and forward masking for subjects with unilateral cochlear impairments," *J. Acoust. Soc. Am.*, vol. 80, no. 1, pp. 93–107, 1986.