EFFECT OF CROSS-WORD CONTEXT ON PLOSIVE IDENTIFICATION IN NOISE FOR NATIVE AND NON-NATIVE LISTENERS

M. Luisa Garcia Lecumberri and Martin Cooke

University of the Basque Country, University of Sheffield
garcia.lecumberri@ehu.es, m.cooke@dcs.shef.ac.uk

ABSTRACT

Studies of second language speech perception can highlight the role of prior knowledge in native language processing. The current study compared native and non-native identification of plosives in words spliced from natural utterances when presented in noise, with and without the context of the preceding word. Both listener groups performed at the same level in the absence of context at high noise, suggesting that the cues surviving energetic masking and splicing were similar for the two languages or that they had already been acquired by the non-native group. However, native listeners gained significantly more when contextual information in the preceding word was present, indicating that cross-word, extra-syllabic, cues are not so easily exploited by non-native listeners. An acoustic analysis revealed subtle durational differences in the preceding word rhyme, knowledge of which may contribute to the native advantage. Other possible explanations for the native benefit from cross-word context are discussed.

Keywords: speech perception, noise, non-native, context, plosive.

1. INTRODUCTION

Studies which compare the perceptions of native (N) and non-native (NN) listeners have the potential to distinguish between the use of signal-driven ("bottom-up") and knowledge-driven ("top down") processes in everyday speech perception [1]. Relative to Ns, NNs lack detailed knowledge of speech patterning in their second language (L2) and also suffer from interference from their own native language (L1). The use of knowledge-driven processes is critical when information is removed from the signal. Indeed, in a study comparing the speech perception performance in noise of monolinguals and bilinguals-since-infancy [7], it was found that the latter group never quite achieved the same levels of performance as the monolinguals. In the current study, listeners were exposed to stimuli in which information was removed in two ways: (i) by adding noise, which results in energetic masking, and (ii) via splicing, which leads to the removal of non-local cues.

Several recent studies compared N/NN performance in noise [3][4], but employed stimuli which forced listeners to use local (within-syllable) cues. The role of information external to a sound's immediate syllabic context is one of a range of effects which constitute the study of fine phonetic detail (FPD) in speech perception, in which it is argued that in natural speech, phonetic, morphological, syntactic and intonational structures have mutual influences which leave their mark on speech production and this information is available and used by listeners when interpreting the message [5].

In the current study, a monolingual English listener group and a native Spanish group studying English at university level identified word-initial plosives in words extracted with and without the immediately-preceding word from natural English utterances. Plosive consonants were chosen since both languages contain the same set, albeit with different realizations. Both the plosive *target word* and the preceding *context word* consisted of a single syllable. Target words varied only in the initial plosive consonant, while the context word nucleus was constant with a range of coda consonants. To introduce a null context for comparison with the spliced sentence-medial target words, an additional condition with the target word in sentence-initial position was tested.

Participants in the current study were relatively advanced learners of English as an L2. It might be expected that patterns of FPD specific to a given L1 are acquired late, if at all, by non-native listeners. Consequently, any differences in the ability of N and NN listeners to exploit cross-word

context in noise might reveal the presence of salient FPD.

2. EXPERIMENT

2.1. Speech and noise materials

Twelve single syllable context words with nucleus /ɪ/ and single consonant codas (Miss, Liz, Phil, Finn, Pip, Lib, Sid, Kit, Sig, Dick, Cliff, Viv) were combined with 6 target words consisting of one of 6 plosives + /eɪ/ (Pay, Bay, Tay, Day, Kay, Gay) to produce 72 word pairs such as "Miss Day" or "Dick Kay". Each word pair was embedded in a different carrier sentence (e.g. "You can see Miss Day tomorrow morning" or "She married Dick Kay three years ago"). Each carrier sentence contained a similar number of syllables and typically had 2-3 syllables prior to the context word. Target words were always followed by an obstruent for ease of segmentation. In addition, 6 further sentences were constructed with the target words in sentence-initial position (e.g. "Pay drives her career moves."). The resulting 78 sentences were read by 3 male native speakers of English in a sound-attenuating booth. The 3 speakers were from different regions of England but none had a strong accent. Speakers were asked to produce the utterances at their normal speaking rate.

Four segmentation points in each of the utterances were marked: (i) the onset of the context word; (ii) the onset of /ɪ/ in the context word, determined by analysis of the formant structure and energy changes in the waveform; (iii) the point in the stop closure of the target word 20 ms prior to its burst; and (iv) the offset of the /eɪ/ diphthong in the target word. For the sentence-initial target words, only (iii) and (iv) were marked.

Two sets of tokens were spliced from the utterances using these segmentation points. The *no-context* set (e.g. "pay") was created by segmenting target words from the interval between markers (iii) and (iv), while the *context* set (e.g. "Liz Pay") was spliced from the interval between markers (i) and (iv). Segmentation point (ii) was used in later analyses.

Speech-shaped noise, constructed by passing white noise through a filter whose magnitude spectrum matched the long-term magnitude of the utterances collected, was added to both sets of tokens at a fairly mild SNR (10 dB) and at a more adverse SNR (-3 dB). The noise started and ended 200 ms before/after the tokens and was ramped on and off to avoid sharp onset/offset effects. Different segments of noise were used for each target word, but to ensure that the only difference between the no-context and the context set was the presence of the context words, the noise signal added to identical target words was the same in the context and no-context sets.

In summary, 4 sets of stimuli were constructed by adding speech-shaped noise at 10 dB and -3 dB to a no-context set containing the spliced target words and to a context set containing the preceding context word and the target word.

2.2. Participants

Nineteen monolingual native English speakers drawn from students and staff of the University of Sheffield and 24 native Spanish speakers studying English Philology at the University of the Basque Country participated in the experiment. The two groups had similar age distributions.

2.3. Procedure

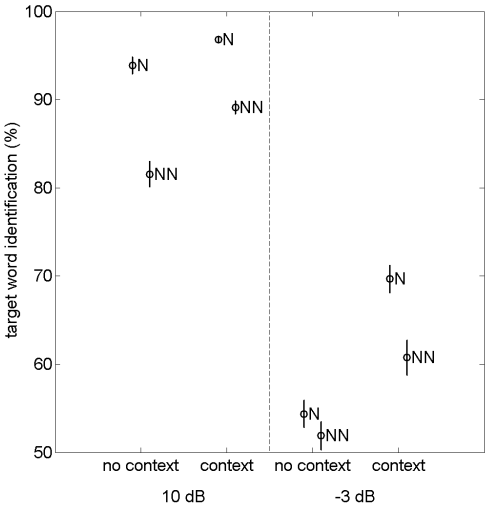
For each of the 4 conditions, stimuli were randomised in blocks of 216 (72 tokens x 3 speakers) for the context sets and 234 (78 tokens x 3 speakers) for the no-context sets. The order of the 4 conditions was randomised across listeners. English and Spanish participants were tested in near-identical conditions in Sheffield and Vitoria respectively. Participants were tested in groups of up to 15 in quiet rooms. Stimuli were delivered under computer control using Plantronics headphones and equivalent quality soundcards. The 4 conditions of the study were presented in a single session lasting approximately 40 minutes, which included a short practice in which participants identified 12 from each of the context and no-context sets in quiet.

3. RESULTS

3.1. Overall identification rates

Figure 1 shows native and non-native listener accuracy in identifying the target words with and without context, in both mild and intense noise.

Figure 1 Target word identification rates for the native (N) and non-native (NN) groups in mild and intense noise, for target words alone ("no context") and with the preceding word ("context"). Error bars indicate +/- 1 standard error.

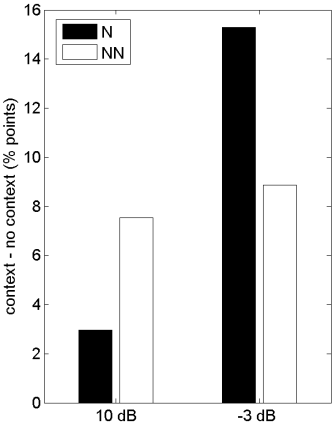


A 3-way ANOVA with factors of noise level, context and nativeness confirmed the overall effects visible in figure 1: native listeners outperformed non-native listeners, contexts were beneficial, and the high noise level was more detrimental than the lower level. Further analysis demonstrated that for both groups and both noise levels, the availability of the preceding context provided a significant benefit[1]. However, pairwise comparisons of the two groups in each of the 4 conditions indicated that although the English listeners significantly outperformed the Spanish group in 3 of the 4 conditions, there was no significant difference between the 2 groups in the condition requiring identification of target words without context in the intense noise background[2]. This finding is at odds with our previous study of consonant identification in VCVs in noise, in which the native advantage grew with increasingly adverse conditions [4]. In the current study, it appears that some of the information needed to support a native advantage has been removed by splicing.

3.2. Benefit of preceding context

Figure 2 depicts the benefit of context. While there is no significant difference in the effect of context for NNs at the two noise levels[3], Ns benefit substantially more[4] in the high noise condition, probably because their performance is near ceiling in the low noise condition. Due to the possibility of ceiling effects, subsequent analyses focused on the high noise conditions. Of most interest is the difference between native and non-native benefit for the high noise case[5], which suggests that native listeners can exploit cues in the prior context that non-native listeners do not take advantage of.

Figure 2 Beneficial effect of context (difference in context and no context scores).



3.3. Plosive identification rates

Figure 3 shows individual plosive identification rates in the -3 dB noise condition. Some clear trends are visible. First, voiceless plosives are better identified than voiced ones. This is particularly the case for /t/ which shows near-ceiling identifications even in the most adverse conditions. Second, identifications do not pattern according to place of articulation: the voiceless alveolar is best identified whereas the voiced alveolar is one of the most difficult. Finally,

[1] Context vs. no-context: F(1,41)=6.1, p < 0.05 for (N, 10 dB); F(1,41)=61.3, p < 0.001 for (N, -3 dB); F(1,41)=49.6, p < 0.001 for (NN, 10 dB); F(1,41)=25.9, p < 0.001 for (NN, -3 dB).

[2] N vs. NN: F(1,41)=44.7, p < 0.001 for (no-context, 10 dB); F(1,41)=1.24, p = 0.27 for (no-context, -3 dB); F(1,41)=82.3, p < 0.001 for (context, 10 dB); F(1,41)=11.8, p < 0.001 for (context, -3 dB).

[3] F(1,41)=0.7, p=0.41

[4] F(1,41)=48.4, p < 0.001

[5] F(1,41)=6.1, p < 0.05

identification patterns are similar for both listener groups, suggesting that the present group of non-native learners of English have already acquired some of the phonetic features that characterize English plosives and differ from Spanish ones (in particular, the aspiration which accompanies voiceless plosives).

On the other hand, the similar patterning of plosive identification between the two listener groups points to the influence of a "universal" factor of energetic masking. For the speech-shaped noise masker used here, whose spectrum falls off at higher frequencies, it is possible that voiceless plosives are better identified because masking is less effective for the high frequency energy of plosive burst and aspiration, and would explain why /t/ performance is high, since its burst has its energy concentrated at higher frequencies than that of /k/, which in turn is higher than /p/ [6]. In support of this suggestion, figure 4 provides some examples of the evidence on which target word identification is likely to be based at the -3 dB SNR level, according to a model of energetic masking [2].

Figure 3 Target word identification rates for the individual plosives in the -3 dB condition.

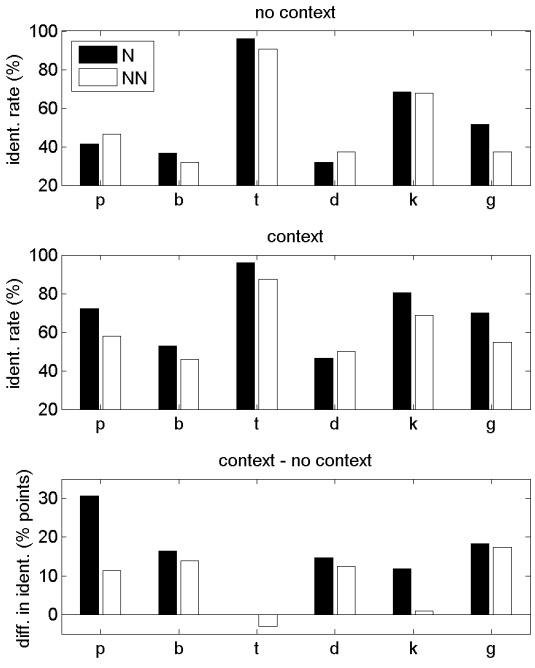
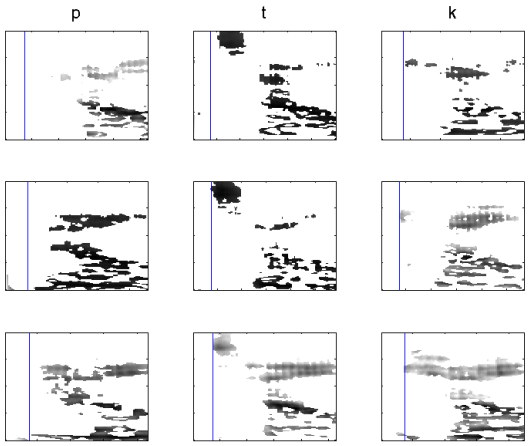


Figure 4 Spectro-temporal "glimpses" of target words "pay", "tay" and "kay" which are deemed to survive masking by speech-shaped noise in the adverse noise condition. Columns 1 to 3 depict examples of the words in the context "fin", although the part of the auditory spectrogram corresponding to the context word is not shown. Each row corresponds to a different talker. Vertical lines denote the release of the stop. Frequency range: 50-8000 Hz on an ERB-rate scale.

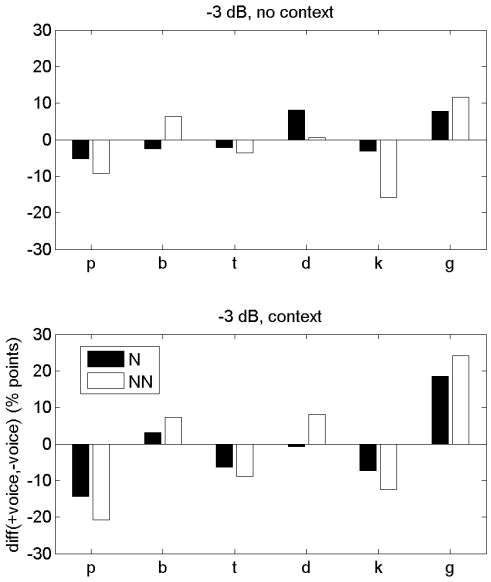


3.4. Effect of previous consonant

Plosive identification was affected by the preceding consonant, including the no context condition. For example, figure 5 displays the effect of coda consonant voicing in the context word on plosive identification in the target word for the -3 dB condition. The quantity depicted is the difference in correct identification of plosives preceded by voiced as opposed to voiceless consonants. This analysis suggests that voiceless plosives tend to be better identified when preceded by voiceless contexts and voiced plosives when preceded by voiced contexts. This tendency was observed both in the high and the low noise conditions but is more apparent for the former. The size of the effect was largest when the context was presented to the listeners, but, perhaps surprisingly, the effect is still present in the no-context conditions. Both listener groups showed these response patterns, but they tended to be stronger for the NNs. Indeed, given the fact that, in Spanish, voiced plosives are fully voiced, it is understandable that Spanish listeners are less accurate at identifying voiced plosives when the context may induce devoicing (as would be the case with a preceding voiceless consonant). English listeners can identify devoiced lenis

plosives, but for them too, more voicing is an additional facilitating cue in difficult conditions.

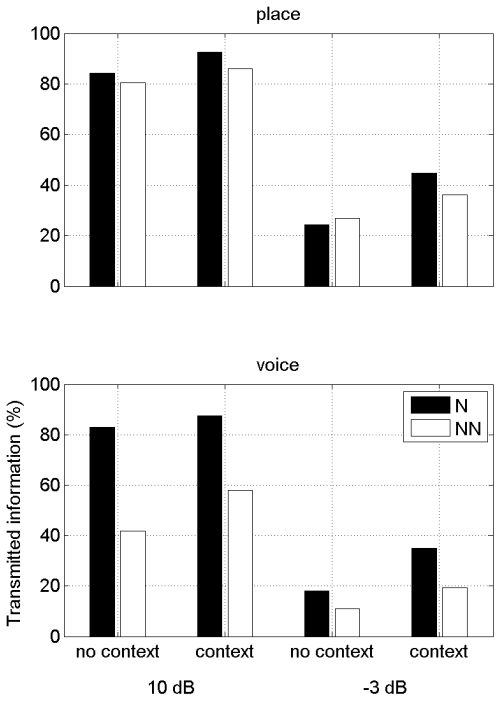
Figure 5 Differential effect of +voice and –voice in preceding consonant on plosive identification in the -3 dB noise condition.



It is of interest to note that the sentence-initial tokens were the least well recognised of all target words by both groups of listeners ($p < 0.001$). It is not clear why words spliced from the middle of sentences should be easier to identify than sentence-initial tokens in these stimuli. The current corpus is too small to permit definitive statements, but acoustic analyses revealed differences in duration and F0 contour of the target word in the two cases. Sentence initial duration of "pay" was significantly longer than sentence-medial "pay", for example. However, it is unlikely that longer duration itself is the cause of poorer identifications in noise, since it ought to provide more opportunities to glimpse the signal.

A transmitted information analysis [8], whose results are shown in Figure 6, demonstrates that for the place feature, the two listener groups behave similarly, as might be expected on the basis of their L1s, both of which exploit the three plosive places of articulation. It is mainly in the cueing of voiceless/fortis and voiced/lenis in which Spanish and English plosives differ, reflected in the proportion of transmitted information for the voice feature, in which the Ns outperform NNs.

Figure 6 Transmitted information for place and voicing features.



4. DISCUSSION

English and Spanish listeners differed in their ability to exploit cross-word contextual information in the identification of plosives in monosyllabic words spliced from English sentences and presented in a high level of speech-shaped noise. While both listener groups performed at a similar level when the target words were spliced out of their natural context, native listeners drew significantly more benefit from presentation of the preceding context word. This finding may indicate that splicing removed cues which native listeners can exploit in adverse conditions, cues which non-native listeners have yet to acquire.

Perhaps the clearest benefit of context comes from the availability of all those cues normally associated with sentence-medial plosive identification, namely voicing in a stop closure and transitions prior to the closure. Spanish normally has voicing throughout the closure of phonologically voiced stops, so it is possible that its absence, through either devoicing typical of many accents of English, or splicing, will cause problems for NN listeners. However, the fact that

both groups behave similarly for place supports the notion that similar formant transitions into the closure would be expected for Spanish and English.

Even finer phonetic detail may exist in the context word which can be exploited by native listeners. Acoustic analyses of devoicing, energy and duration of the context word were performed to determine whether these variables predicted the following plosive. While there were no significant effects of energy or devoicing, an effect of prefortis shortening emerged with small (~5-13 ms) durational reductions in the rhyme of the context word for the voiceless plosives of the target word. These were statistically significant[1] for the bilabials and alveolars but not for the velars. It is well-known that prefortis shortening occurs within the same syllable but this appears to be the first (albeit tentative) report of a cross-word effect.

The unnatural and potentially disruptive absence of contextual cues in the no context condition may also explain the levelling out of performance in two listener groups. Additionally, the fact that both groups scored similarly in the no context condition in the presence of high levels of noise suggests that either the local cues which resist energetic masking are similar for plosive identification in both languages, or that the non-native group had already acquired those cues.

There are other ways in which prior context might produce a native listener advantage. The presence of speech prior to the target may give native listeners a chance to "tune in" to speaker characteristics such as voice quality and accent which could aid in interpretation of the target word. The context word also indicates when to listen for the target word. Further investigations with a larger corpus are required to tease apart the various factors contributing to the patterns observed in this study.

Acknowledgements: Richard Ogden and Sarah Hawkins made many helpful comments on an earlier version of this article.

5. REFERENCES

[1] Cooke, M.P., Garcia Lecumberri, M.L., Barker, J. The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception. Submitted to J. Acoust. Soc. Am.

[2] Cooke, M.P. 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119, 1562-1573.

[3] Cutler, A., Weber, A., Smits, R., Cooper, N. 2004. Patterns of English phoneme confusions by native and non-native listeners. J. Acoust. Soc. Am. 116, 3668-3678.

[4] Garcia Lecumberri, M.L., Cooke, M.P. 2006. Effect of masker type on native and non-native consonant perception in noise. J. Acoust. Soc. Am. 119, 2445-2454.

[5] Hawkins, S. 2003. Roles and representations of systematic fine phonetic detail in speech understanding. J. Phonetics 31, 373-405.

[6] Kent, R.D., Read, C. 1992. *The Acoustic Analysis of Speech*. Singular Publishing Group.

[7] Mayo, L.H., Florentine, M., Buus, S. 1997. Age of second-language acquisition and perception of speech in noise. J. Speech, Lang. and Hearing Research 40, 686-693.

[8] Miller, G.A., Nicely, P. 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27, 338-352.

[1] Bilabials: $t(35)=-2.2$, $p < 0.05$; alveolars: $t(35)=-3.3$, $p < 0.001$; velars: $t(35)=-1.2$, $p = 0.13$.