# Non-native consonant acquisition in noise: effects of exposure/test similarity

Martin Cooke[1, a)] and María Luisa García Lecumberri[2]

[1]*Ikerbasque (Basque Science Foundation), Maria Diaz de Haro 3, 6, 48013 Bilbao, Spain*

[2]*Language and Speech Laboratory, Universidad del País Vasco, 01006 Vitoria, Spain*

(Dated: 21 June 2019)

The following article appeared in **J. Acoust. Soc. Am.** 146, 297–306 (2019) and may be found at https://doi.org/10.1121/1.5116575

When faced with speech in noise, do listeners rely on robust cues or can they make use of joint speech-plus-noise patterns based on prior experience? Recent studies have suggested that listeners are better able to identify words in noise if they experienced the same word-in-noise tokens in an earlier exposure phase. The current study examines the role of token similarity in exposure and test conditions. In three experiments, Spanish learners of English were exposed to intervocalic consonants during an extensive training phase, bracketed by pre- and post-tests. Distinct cohorts experienced tokens that were either matched or mismatched across test and training phases in one or both of two factors: signal-to-noise ratio (SNR) and talker. Cohorts with fully matching test-training exposure were no better at identifying consonants at the post-test phase than those trained in partially or fully mismatched conditions. Indeed, at more adverse test SNRs, training at more favourable SNRs was beneficial. These findings argue against the use of joint speech-plus-noise representations at the segmental level and instead suggest that listeners are able to extract useful acoustic-phonetic information across a range of exposure conditions.

[a] m.cooke@ikerbasque.org

## I. INTRODUCTION

Listeners are able to make sense of speech in a range of less than pristine conditions (e.g. Mattys *et al.*, 2012), but little is known about the detailed processes involved in decoding noisy acoustic input. One fundamental question concerns whether listeners exploit robust cues i.e. representations of the speech signal alone that remain after removing the effect of the masker, or whether they are able to make use of a joint representation of the speech-plus-noise signal acquired on the basis of prior experience with speech material in the presence of a masker.

One way to study the effect of noise on speech representations is to look at the consequences of different types of noise exposure on a group of listeners who are undergoing sound acquisition: non-native language learners. By examining such learners at the stage at which they are acquiring new sounds or modifying their existing native-language categories to accommodate non-native sounds (e.g. Best, 1995; Flege, 1995), it may be feasible to distinguish explanations of speech-in-noise processing that require joint representations of speech and masker at the phonological level from those that argue for the use of robust cues.

We recently demonstrated that exposure to noise during acquisition is beneficial for the identification of non-native consonants in matched noise conditions, and that such exposure presents no barrier to identifying them in quiet (Cooke and García Lecumberri, 2018). Spanish learners of English showed substantial post-test improvements in the identification of consonants in intervocalic contexts (VCVs) in the presence of speech-shaped noise (SSN) following eight training sessions in which they heard VCVs in the same masker, with feedback on incorrect responses. Gains far outstripped those of control groups exposed to vowels in consonantal contexts. Noise habituation was ruled out as an explanation, since a cohort trained on vowels in noise identified consonants in noise no better than a cohort trained on vowels in quiet conditions.

The finding that noise exposure also led to substantial gains on VCVs presented in *noise-free* conditions appears to support the interpretation that during training in noise, listeners were able to acquire cues that they could also deploy in the absence of noise. However, two other outcomes call into question an explanation couched solely in terms of speech cue acquisition as opposed to the learning of joint speech-noise patterns. First, the noise-training benefit did not transfer to a different, untrained, masker: the cohort exposed to noise produced equivalent gains to those of the quiet-trained group when tested in babble noise. If exposure to speech in SSN helped listeners acquire robust cues, or learn robust cue-weighting, it is not obvious why these were not more helpful in babble than any cues acquired by the group trained in quiet. Second, there was a small matched-condition benefit: the group trained in noise produced larger gains when tested in noise than the group trained in quiet, and vice versa. Both findings raise the possibility that some of the noise-training benefit came from the acquisition of joint speech-noise patterns at the phonological level.

The notion that mental representations of speech might contain more than just linguistic information emerged from the finding that exposure to words presented in the same voice led to increases in recognition accuracy in a subsequent test phase relative to words from a different voice (e.g., Goldinger, 1998; Nygaard and Pisoni, 1998; Pisoni and Levi, 2007).

Since words are frequently heard in the context of noise, later studies asked whether the lexicon might contain traces of masking noise in addition to indexical information.

Creel *et al.* (2012) presented listeners with novel words with or without white noise during an exposure phase, and subsequently measured identification performance in matched or unmatched conditions. Identification rates were highest, and responses fastest, for matched exposure and test conditions, indicating that experiencing tokens in noise benefits later presentation in noise. Whether this benefit arises from joint speech-noise representations or cue reweighting is less clear, although by analysing consonant and vowel confusions separately, Creel *et al.* (2012) found only weak evidence of increased weighting of vowel cues in noisy conditions.

While Creel *et al.* (2012) used novel words, the notion of joint representations of speech and noise has been extended to existing words and more complex maskers in a study by Pufahl and Samuel (2014), who found that a change in a co-occurring environmental sound from exposure phase to test phase led to impaired word identification. Recently, Strori *et al.* (2018) hinted that it is not simply the co-occurrence of words and maskers during the exposure phase that leads to subsequent recognition benefits, but rather the integrality of speech and masker. Identification improved when the amplitude envelope of a word was used to modulate the envelope of an accompanying masker, but not when the speech and masker envelopes were independent. Common envelope modulation acts to bind speech and masker into a perceptual object arising from a single acoustic source (Bregman, 1990), which Strori *et al.* (2018) argue is likely to promote the formation of a single unified memory encoding.

Although these studies suggest that noise can form part of an integrated memory representation of novel or existing words, Pufahl and Samuel (2014) and Cooper *et al.* (2015) note that the outcomes are also consistent with an explanation in which the noise itself does not form part of the memory representation: instead it is plausible that listeners make use of the residual incomplete speech-only pattern that results from masking. Cooper and Bradlow (2017) reasoned that these two possibilities can be disentangled by using stimuli in which the masker is spectrally-segregated from the speech, enabling the same speech stimulus to be present in same-noise and different-noise trials. Using this approach within a delayed recognition memory paradigm, Cooper and Bradlow (2017) demonstrated noise-specificity effects for monosyllabic words, suggesting that it is the joint encoding of speech and noise rather than the representation of incomplete speech patterns that is responsible for findings of noise-specificity.

Using the methodology of our earlier study (Cooke and García Lecumberri, 2018), the current investigation explored the issue of joint speech-noise representations by varying the degree of similarity between the material presented to non-native learners during training and test phases. Similarity was manipulated along both the speech and the masker dimensions. For the speech dimension, VCV tokens came from either the same group of talkers or from a different group of talkers during training and testing. For the masker dimension, the SNR was either the same or different in the training and test phases. This design allows us to explore the consequences of both indexical and acoustic similarity. The decision to use different SNRs rather than different maskers was taken to better control the degree of match between exposure and test conditions. The alternative of using different masker types might lead to confounds such as the presence of informational masking (e.g. in the case of babble maskers), or differences in properties such as temporal modulation rates between target and masker tokens (e.g. in the case of modulated noise maskers).

If listeners form joint speech-in-noise patterns based on materials presented during the exposure phase, we predict that subsequent identification performance will depend on the degree of match between the training and test experience. In the context of non-native learners, we hypothesise that the greater the degree of similarility between the tokens heard during training and testing, the larger the improvement from a pre-test baseline, with the greatest gains coming from the fully-matched regime, the smallest gains observed in the fully-mismatched regime, and intermediate gains when either the SNR or the talker set matches.

The main experiment (Expt. 1) explored the two factors (same/different SNR, same/different talkers) in a fully-crossed design. Separate listener cohorts underwent one of four training regimes with both factors matched, one factor matched, or both factors mismatched. Subsequent experiments examined effects of the degree of SNR match (Expt. 2) and talker match (Expt. 3) at less adverse SNRs.

## II. EXPERIMENT 1: MATCHED/MISMATCHED SNR AND TALKERS

Listeners identified consonants in VCVs in quiet and in noise, prior to and following training in noise. During the training phase VCVs came from either the same set of talkers as those used in the test phase or from a different set of talkers, and were mixed with noise either at the same SNR as that used in the test phase or at different SNRs. In the following, the factor SNR or TALKER is prefixed with a '+' or '-' to indicate matched or mismatched conditions e.g. +SNR/+TALKER indicates test and training regimes where both the SNR and talkers were the same, while -SNR/+TALKER indicates that only the talkers matched.

### A. Listeners

Some 96 participants took part in Experiment 1. All were students taking a degree course in English Philology at the University of the Basque Country, and all received course credit for participation. Listeners' results were excluded (numbers in parentheses) from subsequent analysis if any of the following conditions applied (i) their native language was not Spanish or Basque (2); (ii) they reported a hearing problem (1); (iii) they had undertaken intensive consonant training in the previous academic year (10); or (iv) they did not complete the post-test (9). Some 74 listeners (63 female; mean age 19.3, std. dev. 1.3) remained after application of these criteria.

### B. Speech and noise material

Speech material for training and test tokens came from the Consonant Challenge Corpus (Cooke and Scharenborg, 2008), an open collection of VCV sequences produced by female and male British English talkers. This corpus contains consonants from the 24-member set /p, b, t, d, k, g, tʃ, dʒ, f, v, θ, ð, s, z, ʃ, ʒ, h, m, n, ŋ, l, r, j, w/ in the context of combinations of the three corner vowels /æ, u, iː/, with either front or end stress e.g. /ˈaeθi/ vs. /aeˈθi/.

A speech-shaped noise (SSN) masker was used for all training regimes and also during the masked test phase. Noisy tokens were generated by adding VCVs to randomly-selected masker fragments of 1.2 s duration, where the speech onset was varied in the range 0 (synchronous with the masker) to 400 ms delay relative to the noise. Variation in VCV onset was employed for comparability with Cooke and García Lecumberri (2018), where the goal was to render the location of the VCV less predictable within the noise to encourage attentive listening. The masker was scaled to produce the required SNR in the region containing the speech signal i.e., discounting the leading and lagging noise-only sections of the waveform.

VCVs from two sets of talkers were used in the current study. One talker set, denoted 'matched', was used during all masker test phases and during the training regimes for the cohorts undergoing matched talker exposure. This set was composed of four female (talker ids: f1, f7, f12, f21) and four male (m3, m14, m16, m19) talkers. The other set, denoted the 'mismatched' set, was made up from VCVs from female talkers f6, f11, f20 and f23, and male talkers m2, m4, m5 and m17. The mismatched set was used in training regimes where the talkers differed from those used during the test phases. As in our earlier study (Cooke and García Lecumberri, 2018), multiple talkers were used to promote phonetic variability in order to encourage robust learning (e.g., Clopper and Pisoni, 2004; Logan *et al.*, 1991).

### C. Training regimes

Following the pre-test phase, listeners were assigned to one of four training regimes (Tab. I). In the +SNR/+TALKER regime, stimuli were drawn from the same set of talkers used for the test tokens, and the SNR was the same as that used in the test phase (-6 dB). In the +SNR/-TALKER regime the latter condition held but the training tokens came from the 'mismatched' set i.e. different talkers. For the -SNR/+TALKER regime, the talkers were the same as those used in the test set but the five blocks in each training session (see II D below) each had a different SNR, drawn from the set +2, 0, -2, -4, -6 dB. Finally, the -SNR/-TALKER regime consisted of both mismatched talkers and mismatched SNRs. The number of listeners assigned to each regime is indicated in Table I.

TABLE I. *Test and training regimes for Expt. 1. N denotes the number of listeners pursuing each regime.*

| Training regime | SNR (dB) | Talker set | N |
|---|---|---|---|
| +SNR/+TALKER | -6 | matched | 19 |
| +SNR/-TALKER | -6 | mismatched | 18 |
| -SNR/+TALKER | 2, 0, -2, -4, -6 | matched | 18 |
| -SNR/-TALKER | 2, 0, -2, -4, -6 | mismatched | 19 |
| Pre- and post-test | -6 | matched | 74 |

### D. Procedure

During the test phases (pre-test and post-test), listeners identified consonants using a 24-alternative forced-choice procedure. Following the presentation of each stimulus, listeners selected their response from an on-screen keyboard containing a grid of International Phonetic Alphabet symbols, one for each consonant. Participants were familiar with these symbols at the outset of the experiment. Each block contained 384 VCVs, made up of one exemplar of each of the 24 consonants from each of the eight talkers in the test set, with both initial and final stress (24 x 8 x 2 = 384). Each stimulus used a different speech token, and vowel contexts were chosen at random. Listeners underwent two test blocks on separate days, the first without noise (Quiet condition), and the second in the presence of the masker (SSN condition) at an SNR of -6 dB. Pre-tests had mean durations of 20.3 (st. dev. 3.8) and 21.2 (st. dev. 2.2) minutes for the Quiet and SSN conditions respectively. Following the pre-test, participants were assigned to one of the four experimental groups using an automated pseudo-random balancing procedure in such a way as to match group mean scores in both Quiet and SSN conditions to within 0.6%.

Listeners took part in eight training sessions, denoted t1–t8, at a rate of two per week, starting in the week after the pre-test. Eight sessions rather the 10 used in Cooke and García Lecumberri (2018) were deemed sufficient since in that study gains reached a plateau after around six sessions. In each training session participants heard five blocks of consonants, each containing 96 stimuli (four examples of each consonant), drawn from the eight talkers. In this way listeners were exposed to 160 examples of each of the 24 English consonants during the entire training process. The same screen layout was employed during training and testing. During the training phase, listeners received feedback on incorrect responses and had to listen exactly once again to the stimulus before moving on to the next token.

In the week following completion of the training phase listeners undertook a post-test that was identical in all respects to the pre-test. On average, the post-test required 16.7 (st. dev. 2.6) and 19.6 (st. dev. 2.7) minutes for the Quiet and SSN conditions respectively.

Stimuli were delivered via a custom Matlab program running on PCs in a quiet laboratory, through Plantronics Audio-90 headphones (Santa Cruz, CA). Listeners were able to set the volume to a comfortable level at the start of each test or training session.

### E. Postprocessing

Over 99% of the 398592 tokens heard during test and training phases had response times (measured from the offset of the VCV) in the range 0.5 s to 6 s. Some 17 (0.004%) and 3312 (0.83%) tokens were responded to more quickly or slowly respectively, and were excluded from analysis (statistical outcomes were identical across upper exclusion thresholds in the range 4-8 s). Test scores were expressed as the percentage of tokens correctly identified per listener and converted to rationalised arcsine units (RAUs; Studebaker, 1985) for display and statistical analysis.

### F. Results

Fig. 1 plots the percentage of consonants identified correctly in the presence of the SSN masker during the two test phases and in each training session. In the pre-test, participants identified English consonants correctly 52.9% (st. dev. 5.8) of the time. This figure is close to the 54.1% correct reported in Cooke and García Lecumberri (2018) for a similar listener cohort on identical stimuli. Scores at the point of the post-test were higher than in the pre-test for all training regimes, covering a range from 62.0% correct for the +SNR/-TALKER regime to 66.3% correct for the group who underwent -SNR/+TALKER training. Scores improved over the first six training sessions, with limited increases thereafter. Mean scores during training differed across training regimes: the two groups with mismatched SNR training produced higher identification rates than those with matched SNRs. Comparing pairs of regimes for which Talker is contrastive, it is clear that the mismatched speaker set

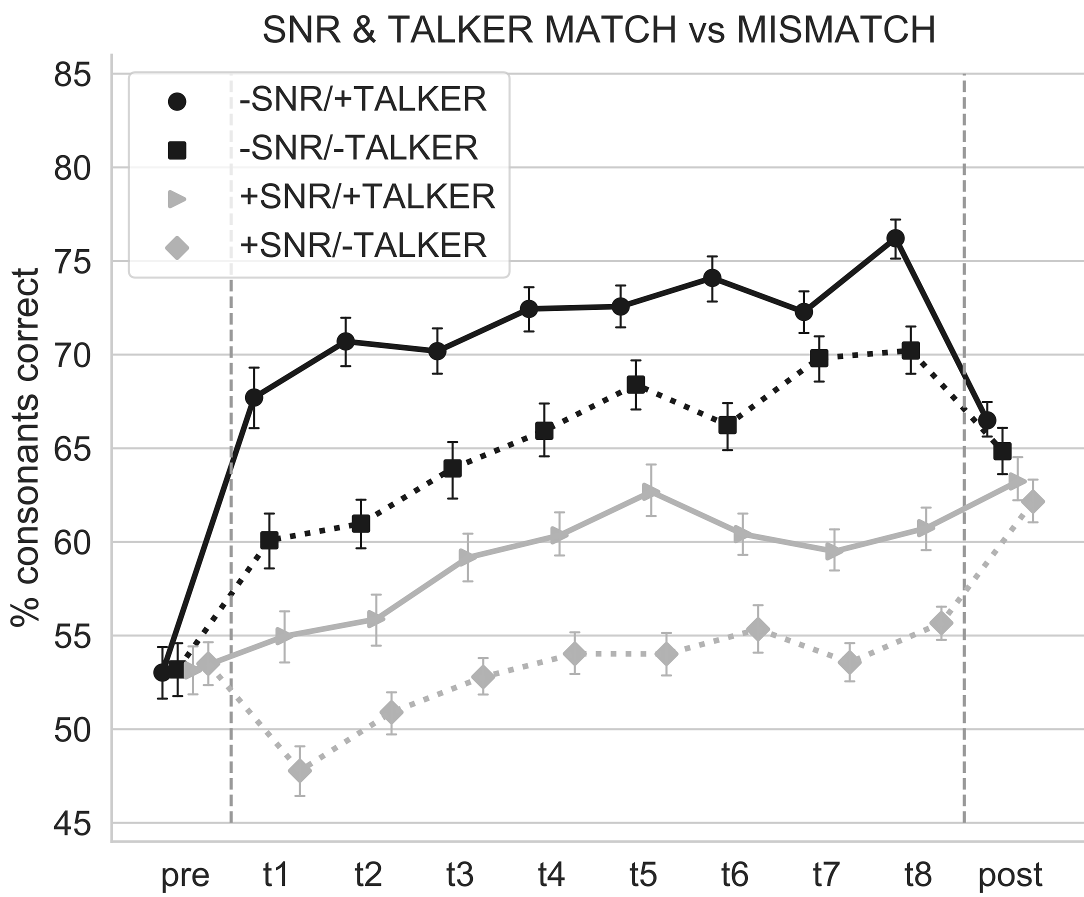


FIG. 1. *Consonant identification rates for the pre- and post-tests (SSN condition only), and in each training session (t1-t8), for the four training regimes of Expt. 1. The lighter line color indicates matched SNR training, while solid lines indicate matched talker training. Error bars here and elsewhere depict* ±1 *standard error.*

is intrinsically somewhat less intelligible than the matched set, with a deficit of around 6 percentage points for each of these contrastive pairings.

Gains, expressed as the difference in RAU-transformed scores between post- and pre-test (Fig. 2) indicate that all four groups benefitted from noise-based training, but to differing extents, with smallest gains for the two groups with matched SNR. A similar pattern of gains is seen for the Quiet and SSN test conditions.

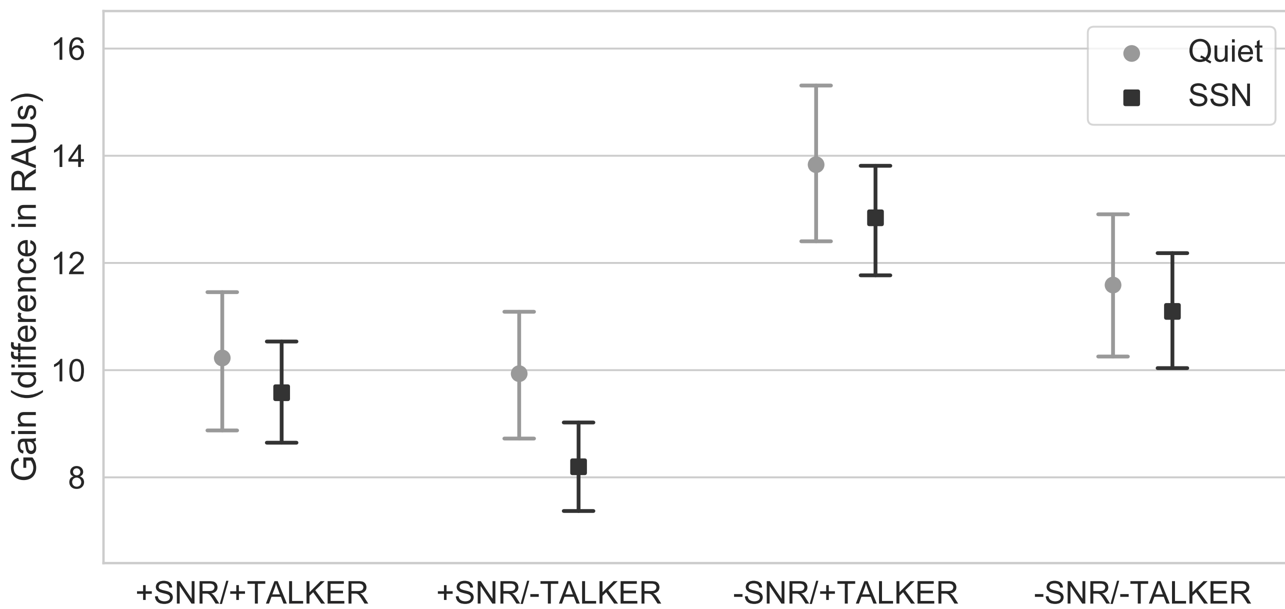


FIG. 2. *Changes from post-test to pre-test in consonant identification rates expressed as a difference in RAUs for the four experimental groups in Expt. 1.*

Potential condition effects for RAU gain scores were examined using a mixed-effects analysis of variance (ANOVA) with two between-subjects factors, Talker (matched/mismatched) and SNR (matched/mismatched) and one within-subjects factor, Masker (Quiet/SSN). Neither the 3-way nor any of the 2-way interactions were statistically-significant [min $p = 0.51$]. This analysis confirmed a significant effect of SNR: cohorts with matched SNR produced significantly *smaller* improvements than cohorts with a mismatch in SNR [$F(1,70) = 7.6, p < .01, \eta^2 = 0.076$]. Cohorts who heard matching talkers in training and test phases showed equivalent improvements as cohorts with mismatching train/test talkers [$p = 0.18$]. There was no effect of the presence or absence of masker on gains [$p = 0.11$].

### G. Interim discussion

We hypothesised that if listeners benefit from noise exposure by learning the joint pattern of speech and masker, gains would be ranked according to the similarity of the test and training conditions. Contrary to this prediction, the fully-matched cohort produced significantly smaller gains than a cohort with a mismatch in SNR between test and training. This outcome is not expected if listeners are learning joint speech-noise patterns, since SNR differences between exposure and test phases will produce a mismatch in the spectro-temporal pattern of the speech residual. All listener cohorts produced significant gains when tested in the Quiet condition, and moreover exhibited a similar ranking of gains across training regime in the Quiet and SSN conditions, suggesting that any acquisitional changes stemming from extensive exposure in noise also served in the absence of noise. Again, this would not be expected on the basis of joint representations of speech and noise.

However, Expt. 1 does not rule out the possibility that listeners acquire *multiple* representations during noise exposure. Specifically, listeners might create enriched representations of sounds based on the cues that survive masking, and also develop integrated representations with masking noise. Two lines of reasoning support this possibility.

First, while the SNR conditions have been expressed in terms of match versus mismatch, they could equally-well be described as ‘unfavourable’ versus ‘favourable’ in the sense that more of the speech target is audible in the mismatched condition. The value of reduced masking is clear in the identification rates during training (Fig. 1) where regimes with mismatched SNRs led to gains of over 13 percentage points over the corresponding matched SNR conditions. It is possible that the relative paucity of speech cues available during training at the more adverse SNR is not compensated for by a putative matched-noise benefit. In support of this notion, there is a striking relationship between the ranking of identification performance in training and in the post-test (although the same-talkers effect is not statistically-significant).

Second, SNRs in the mismatched condition actually overlapped 20% of the time with the more adverse SNR in the matched SNR conditions (Tab. I). It is possible that listeners undertaking mismatched SNR training were still able to construct joint speech-noise representations from the subset of matching stimuli.

Expt. 2 addresses these possibilities by testing whether a matched SNR benefit emerges at a more favourable SNR, and by measuring whether mismatched SNR benefits are also present when training SNRs are fully mismatched i.e. with no SNRs in common between training and test tokens.

## III. EXPERIMENT 2: MATCHED/MISMATCHED SNRS AT A FAVOURABLE SNR

This experiment required listeners to identify intervocalically-presented English consonants in quiet and SSN, but at a more favourable SNR (-3 dB) than the -6 dB used in Expt. 1. Participants were assigned to one of three training regimes which differed in the degree of SNR match during training and test. Except where noted below, methodological details for Expt. 2 were the same as in Expt. 1.

### A. Listeners

A new group of 105 listeners with the same characteristics as in Expt. 1 undertook the experiment. Some 85 listeners (74 female, mean age 19.1, std. dev. 1.0) remained after exclusion of participants' results using the criteria of Expt. 1 (5 non-native, 4 hearing impaired, 6 underwent previous consonant training, 5 did not finish).

### B. Stimuli

Test stimuli were identical to those used in Expt. 1 apart from an increase in SNR from -6 dB to -3 dB in the SSN condition. Training stimuli used the same talkers as those in the test set. Three training regimes were constructed. For the MATCHED regime, training tokens were mixed at the same SNR as the test condition (-3 dB). For the PARTIAL regime, tokens were mixed at the SNRs shown in Table II in equal number. This regime is similar to the -SNR/-TALKER condition of Expt. 1 in that 20% of the time listeners heard tokens at a SNR matching that of the test tokens. SNRs in the MISMATCHED training condition both avoided any match with the test tokens and were significantly more favourable overall than in the other two regimes (Table II). The ranges of SNRs were determined on the basis of pilot tests as values likely to produce significant increases in identification rates over Expt. I while remaining well below ceiling. The 3 dB gap between the lowest SNRs of the PARTIAL and MISMATCHED regimes was chosen to more clearly differentiate the two approaches, given that

TABLE II. *Test and training regimes for Expt. 2.*

| Training regime | SNR (dB) | Talkers | N |
|---|---|---|---|
| MATCHED | -3 | matched | 29 |
| PARTIAL | -3, -1.5, 0, 1.5, 3 | matched | 28 |
| MISMATCHED | 0, 0.75, 1.5, 2.25, 3 | matched | 28 |
| Test set | -3 | matched | 85 |

a smaller difference in SNR between test and training tokens might still be considered useable for integrated speech-noise representations. Listeners were assigned to training regimes using the same pre-test score balancing procedure applied in Expt. 1, resulting in per-regime participant numbers indicated in Table II.

### C. Results

Listeners correctly identified 68.1% of consonants in the SSN condition at the pre-test stage, substantially higher than the 52.9% correctness rate at the more adverse SNR of Expt. 1, confirming that a change of 3 dB leads to a significant performance gain. Scores at the post-test stage averaged 79.3% correct and were very similar for the three cohorts, differing by less than 0.8 percentage points, in spite of clear differences during the training phase (Fig. 3, upper panel). A mixed-effects ANOVA on RAU gains (shown in Fig. 3, lower panel) with a between-subjects factor of training regime and within-subjects factor of test condition (SSN or Quiet) indicated no effect of training regime [$p = 0.86$], test condition [$p = 0.45$] nor their interaction [$p = 0.56$].

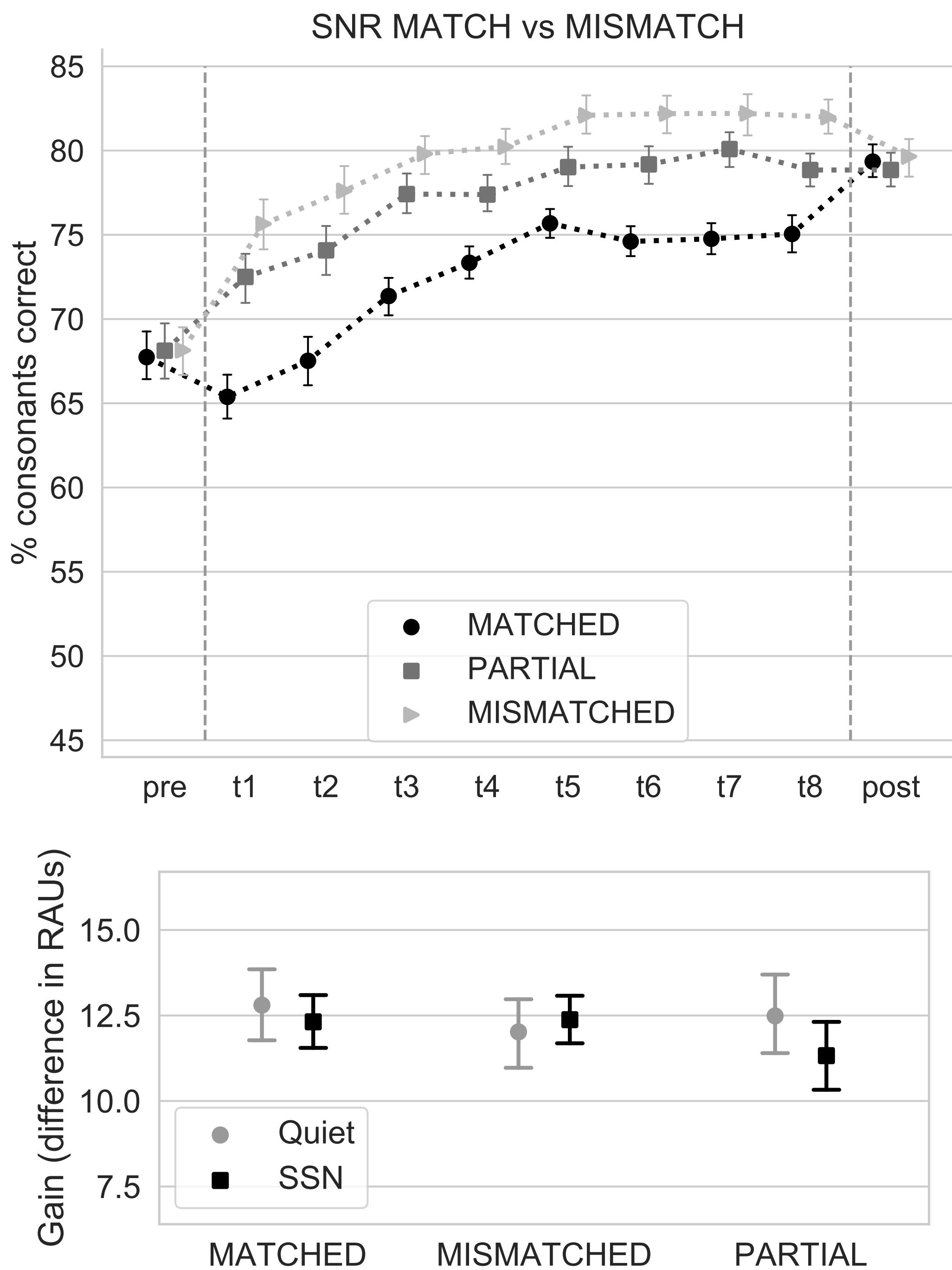


FIG. 3. *Upper: Consonant identification rates for the pre- and post-tests, and in each training session, for the training regimes of Expt. 2 (the vertical scale matches that used in Fig. 1). Lower: Gains from pre- to post-test for the SSN and Quiet conditions.*

### D. Interim discussion

Regardless of whether SNRs matched, partially matched, or mismatched during exposure and test phases, listeners produced similar pre-to-post gains in consonant identification rate. Since each of the three exposure regimes differed in adversity, this outcome suggests that performance gains do not depend strongly on the precise match between exposure and test conditions. Comparing the outcomes of Expts. 1 and 2, it appears that performance in the matched SNR condition of Expt. 1 was limited by cue paucity during the exposure phase, since increasing the test SNR by 3 dB in Expt. 2 led to equivalent gains across the three training regimes.

The fact that gains were identical in the MATCHED and MISMATCHED conditions shows that gains do not depend on having any SNRs in common during exposure and test phases. Further, the idea that listeners might make use of multiple representations is not supported by the finding of equivalent gains in the PARTIAL and MATCHED regimes. If listeners were both extracting robust cues from the more favourable SNRs and acquiring integrated representations from the matched SNRs, larger gains would be predicted in the PARTIAL than in the MATCHED regime. Overall, Expt. 2 is compatible with the hypothesis that listeners acquire robust cues or learn appropriate cue weighting rather than make use of joint speech-noise representations.

### E. Same-talker benefit at more favourable SNRs?

Unlike studies using words that found clear same-talker benefits (e.g., Goldinger, 1998; Nygaard and Pisoni, 1998; Pisoni and Levi, 2007), in Expt. 1 we found no unequivocal evidence of same-talker effects at the phonological level. However, noise is known to reduce indexical effects (Schacter and Church, 1992). Further, the finding of better predictions of relative speaker intelligibility at low SNRs in a study by Barker and Cooke (2007) might be interpreted as resulting from a reduced influence of indexical factors (and a greater reliance on pure energetic masking) at more adverse SNRs. Since Expt. 2 demonstrated that a too-adverse SNR during the training phase can limit the benefits of training, a further experiment was designed to determine whether a matched-talkers benefit would emerge at the more favourable SNR of Expt. 2.

## IV. EXPERIMENT 3: MATCHED VS MISMATCHED TALKERS AT A MORE FAVOURABLE SNR

### A. Listeners

A new group of 109 listeners with the same characteristics as in Expts. 1 and 2 undertook the experiment. Some 93 listeners (78 female, mean age 19.1, std. dev. 1.8) remained after exclusion of participants' results using the criteria of the earlier experiments (2 non-native, 7 had previous consonant training, 7 did not finish).

### B. Stimuli

Test stimuli were identical in all aspects to those used in Expt. 2. Training stimuli were either drawn from the same eight talkers as the test set (MATCHED condition), or came from different talkers (MISMATCHED condition). The talker subsets were the same as those used in the matched and mismatched talker conditions of Expt. 1. All masked stimuli, both test and training, were presented at an SNR of -3 dB (Tab. III).

TABLE III. *Test and training regimes for Expt. 3.*

| Training regime | SNR (dB) | Talkers | N |
|---|---|---|---|
| MATCHED | -3 | matched | 45 |
| MISMATCHED | -3 | mismatched | 48 |
| Test set | -3 | matched | 93 |

### C. Results

Mean identification rates in test and training phases (Fig. 4, upper) indicate that, as in Expt. 1, the cohort trained on matched talkers outperformed the group trained on mismatched talkers during the training phase. However, there was no effect of matched exposure and testing. A mixed-effects ANOVA on RAU gains (shown in Fig. 4, lower) with a between-subjects factor of training regime and within-subjects factor of test condition (SSN or Quiet) indicated no effect of training regime [$p = 0.89$], test condition [$p = 0.92$] nor their interaction [$p = 0.66$]. The clear absence of a matched-talkers effect following exposure at an SNR of -3 dB, a value shown in Expt. 2 to be sufficiently high to produce similar gains as those resulting from training at +3 dB, suggests that listeners were not able to preferentially exploit indexical information in this task.

## V. GENERAL DISCUSSION

The experiments reported here suggest that non-native listeners are able to extract information from a wide range of noise-based training regimes to support equivalent post-training gains in intervocalic consonant identification, as demonstrated in Fig. 5, which compiles outcomes from the 9 training regimes of the current study along with the two consonant regimes of Cooke and García Lecumberri (2018). Apart from the most adverse exposure conditions (-6 dB), RAU gains are strikingly similar and essentially independent of the amount of information available during exposure. Further, there is no evidence of any benefit of matched conditions during test and exposure, either in terms of SNR or talker sets employed. These findings argue against the formation of joint speech-plus-noise representations, and in favour of the use of robust speech cues (e.g., Lovitt and Allen, 2006; Wright, 2004).

There are a number of ways to reconcile the current findings with earlier studies which suggest the formation of joint speech-in-noise representations at the level of words (e.g. Cooper and Bradlow, 2017; Cooper *et al.*, 2015; Creel *et al.*, 2012; Pufahl and Samuel, 2014; Strori *et al.*, 2018). One possibility is that noise combines with speech at the lexical level but not at the phonological level that the current study targets. Alternatively, benefits may be heavily-dependent on using identical speech-in-noise tokens in exposure and test phases, a condition that did not apply in the current study. Finally listeners might behave

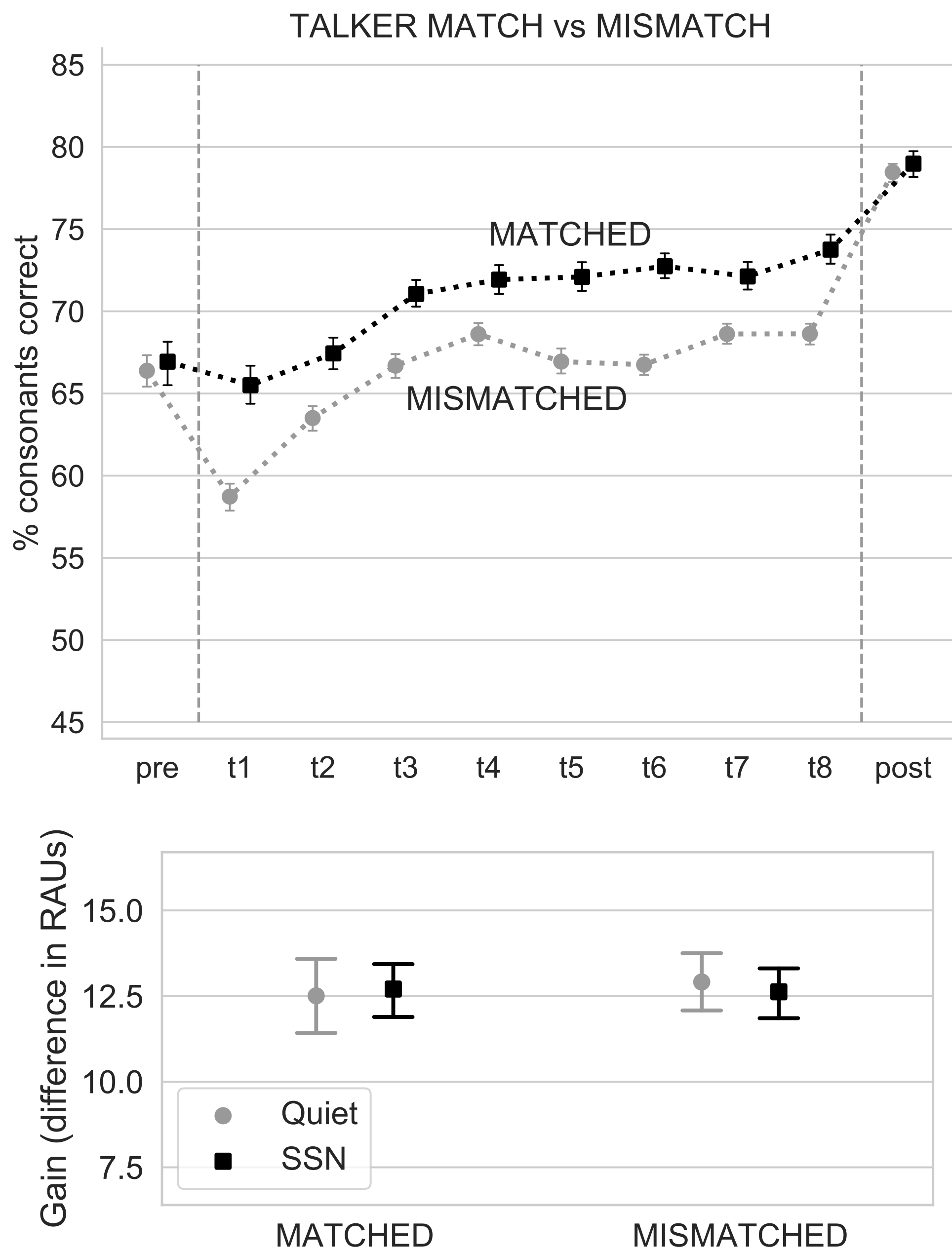


FIG. 4. *Consonant identification rates at the pre- and post-test stages and during each training session (upper), and RAU gains (lower), for the training regimes of Expt. 3.*

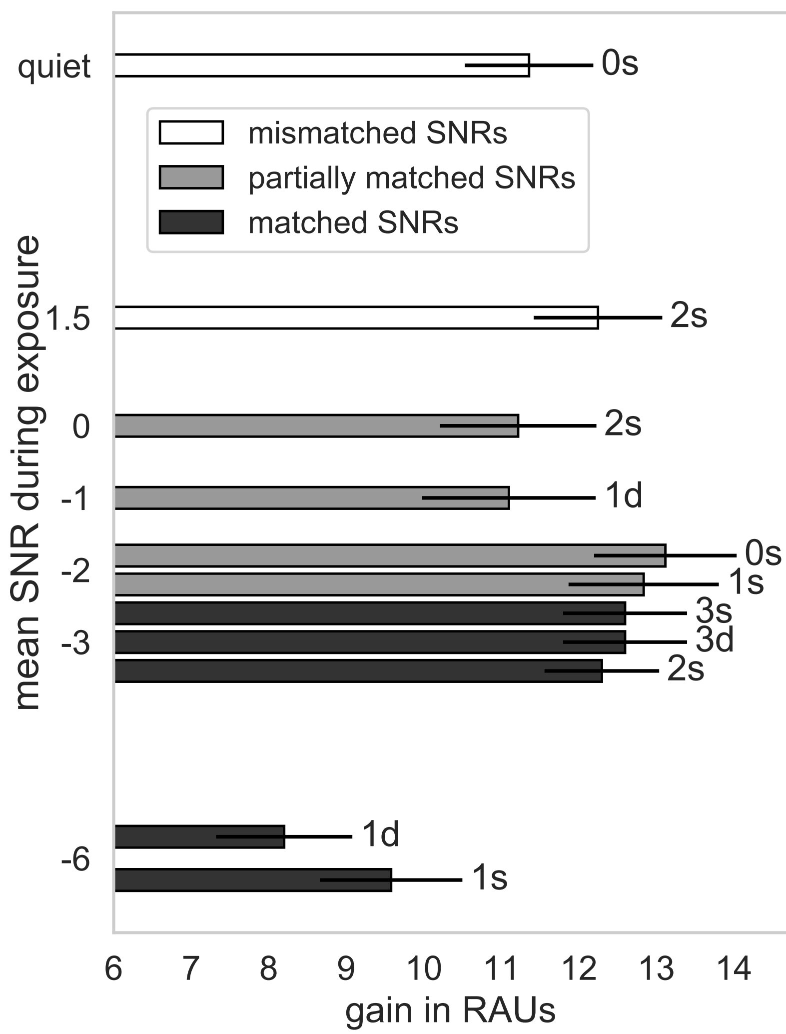


FIG. 5. *Summary of gains for each training regime as a function of mean SNR during training. Bars are labelled with a two character code indicating experiment number from the current study, or 0 to indicate experiments from* Cooke and García Lecumberri (2018)*, and and whether the same ('s') or different ('d') talker sets were used in training and testing. Shading denotes degree of SNR match. Error bars represent ± 1 standard error.*

differently during non-native category acquisition than when confronted by noisy native-language sounds. These ideas are examined below.

### A. Absence of joint speech-noise representations at the sub-lexical level?

While earlier studies have tested the effect of exposure to noise on existing words (Pufahl and Samuel, 2014; Strori *et al.*, 2018) or novel words (Creel *et al.*, 2012), it is an open question as to whether the construction of joint representations of speech and noise is contigent on speech tokens being existing (or potential) members of the lexicon, or whether noise can also influence the representation of speech at sub-lexical levels.

In some respects the VCV stimuli of the current study are similar to the CVCV words in the artificial lexicon used by Creel *et al.* (2012). The existence of a matched condition benefit for listeners exposed to novel CVCVs in Creel *et al.* (2012) might appear to argue for joint sub-lexical representations of speech-plus-noise, since at least on the first occurrence such sequences were presumably treated as lexically-meaningless. One key difference is that while their participants were encouraged to treat the CVCV tokens as new lexical items through association with pictures, our listeners were clearly focused on the segmental level in performing a forced-choice consonant identification task. Since novel CVCV was presented 24 times by Creel *et al.* (2012), it is possible that noise was integrated into the representation only after the sequence achieved lexical status.

Using a speeded classification task, chosen because it does not require processing at the lexical level, Cooper *et al.* (2015) found evidence for the early integration of noise and indexical information relating to speaker gender or identity. This outcome indicates that speech and noise are not segregated at an early stage of processing, lending support to the possibility that they remain in contact up to the phonological level. If integrated representations of speech and noise occur at the level of words but not for the segmental tokens of the current study, the question arises as to why noise might remain attached to the lexical representation while being purged from preceding 'lower' representational levels. One possible answer comes from the differential role played by phoneme-sized units compared to words. As pointed out by Cutler (2012), while the size of the phoneme inventory varies across languages, it is still 3-4 orders of magnitude smaller than the number of words in a typical listener's lexicon. Under this view, phoneme-sized units constitute an efficient compositional encoding mechanism. By analogy with the development of efficient feature layers in machine learning applications (e.g. Hinton and Salakhutdinov, 2006; Tian *et al.*, 2015), where the mechanism to encourage the formation of compositional representations is to choke off the capacity of the layer through a process known as bottleneck training, it is conceivable that only those types of acoustic-phonetic variability (e.g. allphones, reductions) that occur frequently are retained at the phonological level. In this way, it is possible that noise-related acoustic variations are treated as idiosyncratic, and while they pass through to the lexical level, do not form part of any sub-lexical representation. In this respect it is interesting to note that Jesse *et al.* (2007) found weaker same-talker effects (albeit in noise-free conditions) at the sub-lexical than the lexical level.

### B. Degree of stimulus similarity in exposure and test phases

Studies which have found advantages of prior exposure to noise at the lexical level (Creel *et al.*, 2012; Pufahl and Samuel, 2014; Strori *et al.*, 2018) have typically used identical stimuli during the exposure and test phases. In contrast, stimuli in the current study were similar (in the sense of having the same or similar SNRs or being spoken by the same set of talkers), but differed in terms of being independent exemplars drawn from potentially distinct vowel contexts. The use of similar but not identical tokens in the exposure and test phases of the current study was motivated by comparison with our earlier study (Cooke and García Lecumberri, 2018), whose findings permitted an interpretation in terms of joint encoding of speech and noise in spite of non-identical tokens. The absence of a benefit of prior exposure in similar conditions raises the unexplored issue of the extent to which gains from exposure are dependent on identity rather than similarity.

The use of multiple talkers in the current study was a design element to increase phonetic variability since it is well-known that such variability leads to more robust categories during non-native sound acquisition (e.g., Clopper and Pisoni, 2004; Logan *et al.*, 1991). While it is possible that the presence of multiple talkers weakened the degree of similarity between exposure and test conditions, nevertheless listeners heard 20 exemplars of each consonant from each of the 8 talkers during the training phase, a number substantially higher than the quantity typically used in word-based studies of noisy exemplars. Indeed, the number of repetitions has been found not to influence the size of the matched exposure-test benefit (Pufahl and Samuel, 2014, expt. 3), with significant effects from a single exemplar per word.

Although identical speech-plus-noise stimuli are of theoretical interest, they are not representative of a listener's real world experience of challenging speech communication conditions. For this reason, models such as Minerva 2 (Hintzman, 1988) that have been invoked by proponents of the more general episodic memory approach (e.g., Goldinger, 1998) that speech-plus-noise integrality is based on, do not require identical episodes during exposure and later recall, but instead function on the basis of similarity.

### C. Generalisability to native listeners

We chose non-native listeners in the current study for a number of reasons. First, they are in the process of phonological category enrichment for their L2, and the effectiveness of exposure has been clearly demonstrated here and elsewhere (e.g. Clopper and Pisoni, 2004; Cooke and García Lecumberri, 2018) in terms of substantial post-training improvements. We hypothesised that any differential impact of token sets during an extensive exposure phase would be readily measurable with this category of listener. Second, native listeners are close to ceiling performance in quiet conditions on a VCV identification task (Cooke and Scharenborg, 2008) and we were interested in measuring any transfer of exposure benefits to the noise-free condition. Finally, there is recent evidence that listeners are able to retune their non-native categories when presented with ambiguous non-native sounds, at least under lexical guidance (Drozdova *et al.*, 2016). However, it might be argued that non-native listeners process speech in noise, or speech from multiple talkers, in a different manner from native listeners, limiting the generalisation of the findings of the current study from the non-native listener population to native listeners.

Considering first the effect of SNR on non-native listeners, there is certainly evidence that native listeners suffer less in noise for words and sentences (e.g. Black and Hast, 1962; Cooke *et al.*, 2008; Jin and Liu, 2012; Meador *et al.*, 2000; Scharenborg *et al.*, 2018); for a recent review see Scharenborg and van Os (2019). However, other studies (e.g. Cutler *et al.*, 2004; García Lecumberri *et al.*, 2010; Rogers *et al.*, 2006) have demonstrated that native benefits are reduced or absent for the types of subword tokens used in the current study, suggesting that the impact of noise at the sub-lexical level is rather similar for native and non-native listeners.

There is also evidence that native and non-native listeners respond to sub-lexical tokens from multiple talkers in noise in a similar fashion. Bent *et al.* (2010) demonstrated that American English and Korean listeners showed a high level of consistency in ranking the intelligibility of 10 talkers producing vowels in bVd contexts at three SNRs. In a study of Mandarin tone identification in 4 levels of noise with tokens from 6 talkers, Lee *et al.* (2010) found that non-native listeners were no more adversely affected by either the presence of multiple talkers or by noise level than native listeners. These results are consistent with an earlier study by Bradlow and Pisoni (1999) using words presented without noise, in which it was found that non-native listeners responded similarly to native listeners in the face of indexical variability.

Taken together, these studies support the idea that at the sub-lexical level, multiple talkers and noise affect native and non-native listeners to a similar degree. This should not be surprising: while the impact of acoustic and indexical variability on L2 categories may differ in detail from their impact on native language categories, a non-native listener's everyday experience encompasses both noise and talker variation, and it seems likely that any processes or representations which handle variability in their L1 can also be deployed in an L2.

### D. Mismatched condition benefit in machine classification systems

Finally, we note that while listeners' performance might reasonably be considered to be optimal when the conditions under which sounds are acquired match everyday usage, recent studies in machine learning (e.g., Gonzalez and Abu-Mostafa, 2015) question the common assumption that classifier systems perform best in noise under matched exposure and test conditions. For example, Sivasankaran *et al.* (2017) have shown that training data with a mismatched selection of SNRs led to better performance than obtained when training using matched SNRs for a challenging speech separation and recognition task (Barker *et al.*, 2015).

## VI. CONCLUSIONS

Non-native listeners exposed to intervocalic consonants in noise did not exhibit greater gains from pre- to post-test when speakers or signal-to-noise ratios were matched between exposure and test phases than when one or both properties were mismatched. These findings highlight the flexibility of non-native sound acquisition in challenging listening conditions and suggest that listeners are capable of extracting robust cues to support consonant identification from a range of training regimes differing in adversity.

## ACKNOWLEDGMENTS

This study was carried with funding from the Basque Government *Consolidados* grant to the Language and Speech Laboratory at the University of the Basque Country.

Barker, J., and Cooke, M. (**2007**). “Modelling speaker intelligibility in noise,” Speech Communication **49**, 402–417.

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (**2015**). “The 3rd CHIME Speech Separation and Recognition Challenge: dataset, task and baselines,” in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 504–511.

Bent, T., Kewley-Port, D., and Ferguson, S. H. (**2010**). “Across-talker effects on non-native listeners’ vowel perception in noise,” Journal of the Acoustical Society of America **128**, 3142–3151.

Best, C. (**1995**). “A direct realist view of cross-language speech perception,” in *Speech Perception and Linguistic Experience*, edited by W. Strange (Timonium), pp. 171–204.

Black, J. W., and Hast, M. H. (**1962**). “Speech reception with altering signal,” Journal of Speech and Hearing Research **5**, 70–75.

Bradlow, A., and Pisoni, D. (**1999**). “Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors,” Journal of the Acoustical Society of America **106**, 2074–2085.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT Press).

Clopper, C., and Pisoni, D. (**2004**). "Effects of talker variability on perceptual learning of dialects," Language and Speech **47**, 207–239.

Cooke, M., and García Lecumberri, M. L. (**2018**). "Effects of exposure to noise during perceptual training of non-native language sounds," Journal of the Acoustical Society of America **143**, 2602–2610.

Cooke, M., Garcia Lecumberri, M. L., and Barker, J. (**2008**). "The foreign language cocktail party problem: energetic and informational masking effects in non-native speech perception," Journal of the Acoustical Society of America **123**, 414–427.

Cooke, M., and Scharenborg, O. (**2008**). "The Interspeech 2008 consonant challenge," in *Proc. Interspeech*, pp. 1765–1768.

Cooper, A., and Bradlow, A. R. (**2017**). "Talker and background noise specificity in spoken word recognition memory," Laboratory Phonology **8**, 1–15.

Cooper, A., Brouwer, S., and Bradlow, A. R. (**2015**). "Interdependent processing and encoding of speech and concurrent background noise," Attention, Perception & Psychophysics **77**, 1342–1357.

Creel, S. C., Aslin, R. N., and Tanenhaus, M. K. (**2012**). "Word learning under adverse listening conditions: context-specific recognition," Language and Cognitive Processes **27**, 1021–1038.

Cutler, A. (**2012**). *Native listening: Language experience and the recognition of spoken words* (MIT Press).

Cutler, A., Weber, A., Smits, R., and Cooper, N. (**2004**). "Patterns of English phoneme confusions by native and non-native listeners," Journal of the Acoustical Society of America

**116**, 3668–3678.

Drozdova, P., van Hout, R., and Scharenborg, O. (**2016**). "Lexically-guided perceptual learning in non-native listening," Bilingualism: Language and Cognition **19**, 914–920.

Flege, J. (**1995**). "Second language speech learning: Theory, findings and problems," in *Speech Perception and Linguistic Experience*, edited by W. Strange (Timonium), pp. 233–277.

García Lecumberri, M. L., Cooke, M., and Cutler, A. (**2010**). "Non-native speech perception in adverse conditions: A review," Speech Communication **52**, 864–886.

Goldinger, S. D. (**1998**). "Echoes of echoes? An episodic theory of lexical access," Psychological Review **105**, 251–279.

Gonzalez, C. R., and Abu-Mostafa, Y. S. (**2015**). "Mismatched training and test distributions can outperform matched ones," Neural Computation **27**, 365–387.

Hinton, G. E., and Salakhutdinov, R. R. (**2006**). "Reducing the dimensionality of data with neural networks," Science **313**, 504–507.

Hintzman, D. (**1988**). "Judgments of frequency and recognition memory in a multiple-trace memory model," Psych. Review **95**, 528–551.

Jesse, A., McQueen, J. M., and Page, M. (**2007**). "The locus of talker-specific effects in spoken-word recognition," in *Proc. International Congress of Phonetic Sciences*, pp. 1921–1924.

Jin, S. H., and Liu, C. (**2012**). "English sentence recognition in speech-shaped noise and multi-talker babble for English-, Chinese-, and Korean-native listeners," Journal of the Acoustical Society of America **132**, 391–397.

Lee, C.-Y., Tao, L., and Bond, Z. S. (**2010**). “Identification of multi-speaker Mandarin tones in noise by native and non-native listeners,” Speech Communication **52**, 900–910.

Logan, J. S., Lively, S. E., and Pisoni, D. B. (**1991**). “Training Japanese listeners to identify English /r/ and /l/: A first report,” Journal of the Acoustical Society of America **89**, 874–886.

Lovitt, A., and Allen, J. (**2006**). “50 years late: Repeating Miller-Nicely 1955,” in *Proc. Interspeech*, pp. 2154–2157.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). “Speech recognition in adverse conditions: A review,” Language and Cognitive Processes **27**, 953–978.

Meador, D., Flege, J. E., and Mackay, I. R. A. (**2000**). “Factors affecting the recognition of words in second language,” Bilingualism: Language and Cognition **3**, 55–67.

Nygaard, L. C., and Pisoni, D. B. (**1998**). “Talker-specific learning in speech perception,” Perception and Psychophysics **60**, 355–376.

Pisoni, D. B., and Levi, S. V. (**2007**). “Some observations on representations and representational specificity in speech perception and spoken word recognition,” in *The Oxford Handbook of Psycholinguistics*, edited by G. Gaskell (Oxford University Press), pp. 3–18.

Pufahl, A., and Samuel, A. G. (**2014**). “How lexical is the lexicon? Evidence for integrated auditory memory representations,” Cognitive Psychology **70**, 1–30.

Rogers, C., Lister, J., Febo, D., Besing, J., and Abrams, H. (**2006**). “Effects of bilingualism, noise and reverberation on speech perception by listeners with normal hearing,” Applied Psycholinguistics **27**, 465–485.

Schacter, D. L., and Church, B. A. (**1992**). “Auditory priming: implicit and explicit memory for words and voices,” Journal of Experimental Psychology: Learning, Memory and Cognition **18**, 915–930.

Scharenborg, O., Coumans, J. M. J., and van Hout, R. (**2018**). “The effect of background noise on the word activation process in non-native spoken-word recognition,” Journal of Experimental Psychology **44**, 233–249.

Scharenborg, O., and van Os, M. (**2019**). “Why listening in background noise is harder in a non-native language than in a native language: A review,” Speech Communication **108**, 53–64.

Sivasankaran, S., Vincent, E., and Illina, I. (**2017**). “Discriminative importance weighting of augmented training data for acoustic model training,” in *Proc. ICASSP*, pp. 4885–4889.

Strori, D., Zaar, J., Cooke, M., and Mattys, S. L. (**2018**). “Sound specificity effects in spoken word recognition: The effect of integrality between words and sounds,” Attention, Perception & Psychophysics **80**, 222–241.

Studebaker, G. (**1985**). “A rationalized arcsine transform,” Journal of Speech and Hearing Research **28**, 455–462.

Tian, T., Cai, M., He, L., and Liu, J. (**2015**). “Investigation of bottleneck features and multilingual deep neural networks for speaker verification,” in *Proc. Interspeech*, pp. 1151–1155.

Wright, R. (**2004**). “A review of perceptual cues and cue robustness,” in *Phonetically Based Phonology*, edited by B. Hayes, R. Kirchner, and D. Steriade (Cambridge University Press), pp. 34–57.