Article:
Beeston, A.V., Brown, G.J. and Watkins, A.J. (2014) Perceptual compensation for the effects of reverberation on consonant identification: Evidence from studies with monaural stimuli. Journal of the Acoustical Society of America, 136 (6). 3072 - 3084. ISSN 0001-4966

https://doi.org/10.1121/1.4900596

Perceptual compensation for the effects of reverberation on consonant identification: Evidence from studies with monaural stimuli [a)]

Amy V. Beeston[b)] and Guy J. Brown

Department of Computer Science,

University of Sheffield,

Sheffield S1 4DP,

United Kingdom

Anthony J. Watkins

Department of Psychology,

University of Reading,

Reading RG6 6AL,

United Kingdom

(Dated: September 26, 2014)

Abstract

Mounting evidence suggests that listeners perceptually compensate for the adverse effects of reverberation in rooms when listening to speech monaurally. However, it is not clear whether the underlying perceptual mechanism would be at all effective in the high levels of stimulus uncertainty that are present in everyday listening. Three experiments investigated monaural compensation with a consonant identification task in which listeners heard different speech on each trial. Consonant confusions frequently arose when a greater degree of reverberation was added to a test-word than to its surrounding context, but compensation became apparent in conditions where the context reverberation was increased to match that of the test-word; here, the confusions were largely resolved. A second experiment shows that information from the test-word itself can also effect compensation. Finally, the time course of compensation was examined by applying reverberation to a portion of the preceding context; consonant identification improves as this portion increases in duration. These findings indicate a monaural compensation mechanism that is likely to be effective in everyday listening, allowing listeners to recalibrate as their reverberant environment changes.

PACS numbers: 43.55.Hy, 43.71.Es, 43.71.An, 43.66.Lj

I. INTRODUCTION

When speech is heard in a room, the direct sound is accompanied by many reflections from the room's surfaces. These time-delayed and attenuated reflections combine additively to form reverberation, which reduces the modulation depth of the speech and adversely affects its intelligibility (Bolt and MacDonald, 1949; Houtgast *et al.*, 1985). However, speech perception in rooms is remarkably robust under diverse reverberation conditions; the message heard remains the same regardless of whether the listener and speaker are nearby or far apart. A mounting body of evidence (see for example: Watkins, 2005a,b; Ueno *et al.*, 2005; Longworth-Reed *et al.*, 2009; Watkins *et al.*, 2011; Brandewie and Zahorik, 2010, 2012, 2013; Srinivasan and Zahorik, 2013, 2014) suggests that this robustness is underpinned by auditory mechanisms that compensate for the effects of reverberation on the speech signal. Apparently, the auditory system achieves *perceptual constancy* in reverberation in much the same way as the visual system exhibits constancy for the properties of surfaces such as their size, colour and brightness (Adelson, 2000).

Many studies have shown that the perception of a reverberant sound is influenced by the properties of its temporal context (e.g., Watkins, 2005a; Longworth-Reed *et al.*, 2009; Brandewie and Zahorik, 2013; Srinivasan and Zahorik, 2014). Taken together, these studies suggest the following conceptual model: in order to achieve perceptual constancy, listeners exploit information gleaned from the acoustic context preceding a test sound, and accumulate this information over a period of time. The current paper focuses on two aspects of this model that require clarification: the nature of the information that is used, and the time

---

[a)]A portion of this work was presented in "Perceptual compensation for the effects of reverberation on consonant identification: A comparison of human and machine performance," Proceedings of 13th Annual Conference of the International Speech Communication Association (Interspeech), Portland, OR, September, 2012.

[b)]Author to whom correspondence should be addressed. Electronic mail: a.beeston@sheffield.ac.uk

period over which it is gathered.

Employing binaural speech identification tasks, in which reverberant speech stimuli are presented in spatialized noise, Zahorik and colleagues have demonstrated a binaural compensation effect for natural speech stimuli drawn from the Coordinate Response Measure and Modified Rhyme Test datasets (Brandewie and Zahorik, 2010, 2012), and from the PRESTO subset of TIMIT sentences (Srinivasan and Zahorik, 2013). Related work has also recently begun to probe the mechanisms that may account for these effects. Zahorik and colleagues (Zahorik *et al.*, 2012; Zahorik and Anderson, 2013) have found that prior listening to a particular room improves listeners' ability to detect amplitude modulation in that room.

In addition, a monaural compensation mechanism has been demonstrated in a set of behavioural studies using a phoneme identification task (e.g., Watkins, 2005a,b; Watkins *et al.*, 2011). From these experiments, it would appear that monaural mechanisms are primarily informed by the temporal envelope of the signal (which does not necessarily need to be speech). A recent study of neural coding appears consistent with this proposition. While studying responses of an inferior colliculus neuron in an unanesthetised rabbit, Kuwada *et al.* (2012) reported that monaural mechanisms seemed to underpin neural coding of both envelope synchrony and modulation gain. They observed a higher modulation gain in reverberant conditions, relative to anechoic conditions, and hypothesised that this may constitute a compensatory mechanism to redress the detrimental effects of reverberation on modulation depth.

Watkins' monaural identification task is highly sensitive to the way that reverberation tends to 'morph' one sound into another (cf. 'confusion heterogeneity' and 'threshold variability' in Phatak *et al.*, 2008). In Watkins' experiments, listeners identified the test-words 'sir' and 'stir' embedded in a fixed phrase. The test-words were drawn from a continuum of 11 steps that was created by interpolating between the temporal envelopes of naturally spoken tokens of 'sir' and 'stir' (Watkins, 2005b). Different amounts of reverberation were applied to the context phrase and test-word by convolving them with room impulse responses recorded at a 'near' or 'far' distance, and the step in the continuum at which the percept

switched from 'sir' to 'stir' (the *category boundary*) was measured. When the context was reverberated at the 'near' distance with the test-word reverberated at the 'far' distance, listeners tended to make more 'sir' responses (as though the dip in the temporal envelope that cued the [t] consonant had been concealed by reverberant energy). However, if both the context phrase and test-word were reverberated at the 'far' distance, more of the continuum steps were perceived as 'stir' again (even though the factors that had seemed to obscure the [t] were still present). Watkins concluded that listeners routinely use information about the temporal envelope of surrounding speech to compensate for the effects of reverberation on a particular word.

However, monaural compensation is not always apparent. In a recent study that used speech material from the Coordinate Response Measure dataset, only two of fourteen participants were reported to derive an appreciable benefit from monaural room exposure (Brandewie and Zahorik, 2010). We note, however, that the listeners' task in this study required identification of reverberant speech *in noise* (room reverberation was binaurally simulated, presenting speech directly in front of the listener and a masking noise to the side), and thus may potentially conflate the speech identification task with aspects of localisation and spatial unmasking. A second possibility is that monaural effects only emerge when small numbers of tokens are used in an experiment (phoneme-continuum identification has typically used minimal numbers of speech tokens) and might be less prominent in experiments where speech differs from trial to trial and the variation among sounds is thus more similar to everyday listening. Thirdly, performance differences in this sort of task might not result from the 'morphing' effects of reverberation seen with phoneme-continuum identification, where identification errors are consistently a single response-alternative, but from a rather more even distribution of errors across the response alternatives (Phatak *et al.*, 2008).

The time course of the monaural compensation effect has yet to be studied. However, a number of studies have recently queried the timescales on which binaural compensation effects operate. Shinn-Cunningham (2000) reported that localisation accuracy improves with long-term learning of a particular room condition (c. 5 hours). In contrast to this long-term

effect, Zahorik *et al.* (2009) reported that listeners' ability to determine the azimuth of a test pulse was impeded on a short timescale (just seconds) by inconsistent reverberation on the preceding context. For speech-based binaural tasks, a consistently reverberated context has several times been reported to provide benefit at the minimum temporal resolution of the experimental analysis. These effects were noted to occur on timescales measured in minutes for the sets of sentences examined in Longworth-Reed *et al.* (2009); within the first six sentences for material in Srinivasan and Zahorik (2013); and in just a few seconds for the two-sentence carriers used in Brandewie and Zahorik (2010). However, in a recent study designed specifically to measure the time course of the binaural effect, Brandewie and Zahorik (2013) reported that an exposure time of 850 ms was sufficient to achieve considerable speech intelligibility enhancement.

The current paper asks three questions relating to the conceptual model described above. First, we ask about the ecological relevance of monaural compensation using experiments in which the test-word, context-words and talker heard by a listener may vary independently from one trial to the next. The listening task allows consonant confusions to be measured, so that conclusions are not necessarily confined to the 'morphing' effects of reverberation. If there is a perceptual compensation in these conditions then it should reduce these consonant confusions.

A second question relates to the respective roles of the context and test-word in perceptual compensation. The perceptual compensation considered in our conceptual model might be termed 'extrinsic' since it is effected by information from the preceding speech context, which is external to the test-word (Watkins, 2005b). Watkins and Raimond (2013) observed a compensation effect that appeared to be 'intrinsic', in that it arose through information from within the test-word itself (including reverberation tails). Their experiment found a robust effect of intrinsic information, but only examined this for cases when test-words were presented in isolation (i.e., without any 'extrinsic' context). However, a context is generally present in everyday listening, so Experiment 2 asks whether intrinsic information plays a role in perceptual compensation when extrinsic information is also present. Extrinsic information

is removed by silencing the speech precursor, and some intrinsic cues to compensation are reduced with the method described in Watkins and Raimond (2013) where the reverberant tail at the very end of the test-word is gated (shortened).

Finally, we seek to clarify the time course of the monaural compensation effect that is suggested by the idea that information builds up over time in our conceptual model (cf. binaural timescales in Brandewie and Zahorik, 2013). Accordingly, Experiment 3 asks how much of the context phrase must be reverberated in order to compensate for the effects of reverberation in the test-word. This is achieved by applying reverberation to the context abutting the test-word, using temporal windows that had different durations.

II. METHOD

A. Speech material

Speech material was drawn from the Articulation Index Corpus (AIC), which contains around 2000 real-word and nonsense test syllables, among them the words 'sir' and 'stir', each embedded in a short phrase and spoken by 20 different talkers (Wright, 2005). The phrases used for each trial were similar in form to those used by Watkins, consisting of a single test syllable (TEST) within a sequence of three context words (CW):

[CW1][CW2][TEST][CW3].

Context words were drawn at random from a set of different words comprising 8 CW1 pronouns, 51 CW2 verbs, and 43 CW3 codas, resulting in a quasi-predictable temporal location for the test-word within a semantically unpredictable phrase (cf. Srinivasan and Zahorik, 2011) e.g., "people note sir typically" or "I evoke stir precisely". Prompts were generated separately for each talker; thus a given TEST was present in 20 individual CW sequences, each of which was spoken by a different talker.

Since reverberation tends to introduce more errors involving place of articulation than manner or voicing (Gelfand and Silman, 1979; Drullman *et al.*, 1994), our experiments examined unvoiced plosive consonants differentiated by horizontal place of articulation: alveolar

[t], velar [k], and bilabial [p]. In natural speech, these consonants are characterised by a brief silence or period of low amplitude that occurs when the airway is restricted by the articulators. They are particularly susceptible to the effects of reverberation since the dip in their temporal envelopes which helps to cue their identity may easily become obscured. Nábělek *et al.* (1989) reported that these consonants are even more vulnerable to reverberation when presented after an [s] sound than when they are presented alone, so the initial [s] of Watkins' test-words was maintained in all experiments.

To allow a direct comparison with Watkins' results, Experiment 1 used only the [ɜ˞] vowel that appears in his 'sir-stir' test-words. Later experiments used a larger number of vowels in order to widen the test material drawn from the AIC and increase the data obtained from each participant.

B. Convolution with room impulse responses

The experiments that follow used monaural stimuli, obtained by convolving speech with the left-channel of binaural room impulse responses (IRs) recorded with an acoustic manikin by Watkins (2005b) in an L-shaped office (volume 183.6 m^3). IRs were recorded at two source-receiver distances, denoted 'near' (0.32 m) and 'far' (10 m), which resulted in different levels of reflected sound. The early (50 ms) to late energy ratio in the impulse response was 18 dB at the 'near' distance, reducing to 2 dB at the 'far' distance. The later portion of the energy decay curve was practically linear (as shown in Watkins, 2005b, Figure 1), with an energy decay rate of 60 dB per 281 ms at 'near', and 60 dB per 969 ms at 'far'.

Test-word and context portions of the AIC utterances were independently convolved with 'near' and 'far' IRs, and then recombined to give the same- and mixed-distance reverberation conditions depicted in Figure 1. Accordingly, when the stimuli were presented monaurally over headphones to listeners seated in a sound-isolating booth, the sounds at their ear were the same as those for speech arriving from sources nearby or further away in the room.

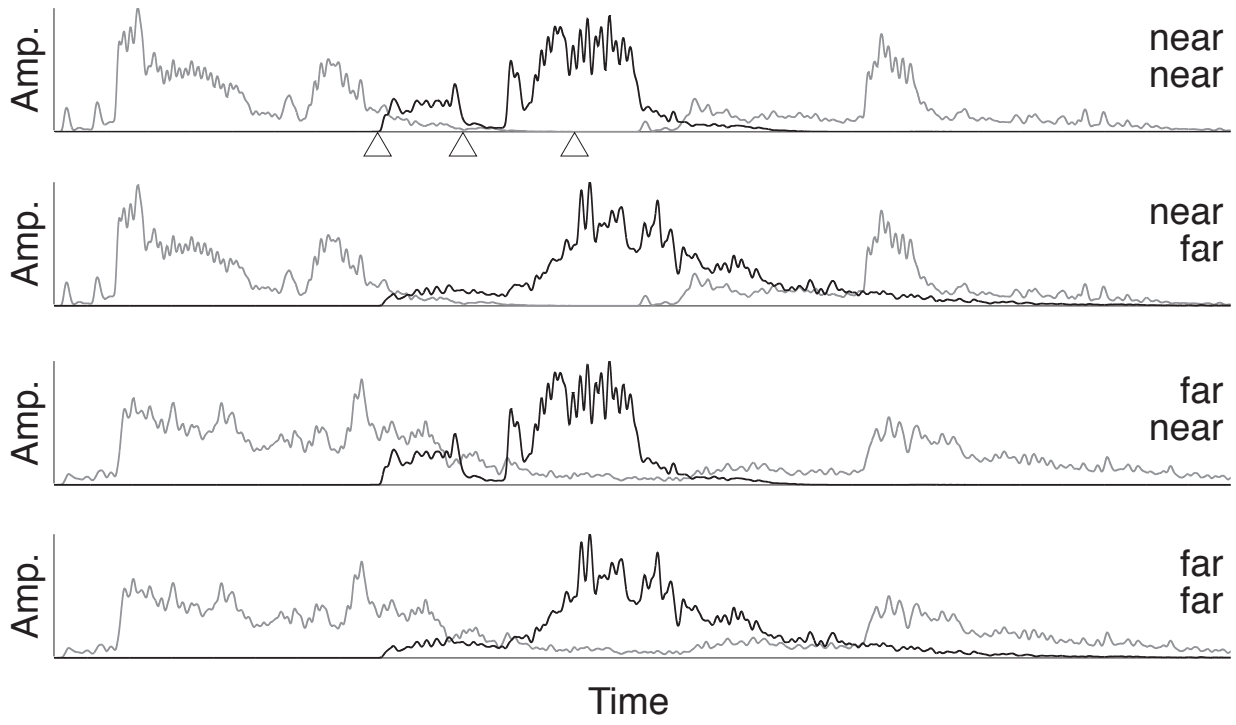

FIG. 1. Illustration of same- and mixed-distance reverberation conditions for one representative example of the 80 phrases used in Experiment 1. The traces are amplitudes (Amp.) of low-pass filtered (cutoff frequency 80 Hz) Hilbert envelopes derived from the temporally aligned context *(light line, upper label)* and test-word *(dark line, lower label)* before these two sounds were added, point-wise, to form the experimental stimuli. Before the addition, the context and test-word were independently reverberated at 'near' or 'far' room distances to give, from top to bottom: near-near, near-far, far-near and far-far *context-test* distance conditions. In the top panel, the test-word is annotated with pointers to show, from left to right, the start of frication, closure and voicing.

C. Measuring the constancy effect

Participant responses were recorded in consonant confusion matrices and analysed in terms of relative information transmitted (RIT) as described by Miller and Nicely (1955). This method regards participants as information channels that receive an input X and

respond with output Y, and measures their information transfer characteristic, given by

$$RIT = \frac{H(X;Y)}{H(X)} \tag{1}$$

where $H(X;Y)$ is the mutual information of X and Y, and $H(X)$ is the self-information (entropy) of X[1]. RIT summarises the consonant identification pattern of the confusion matrix with a single value that ranges from 1 with perfect transmission to 0 for random responses.

The RIT metric offers three benefits in characterising consonant identification over other measures such as percentage correct. Firstly, RIT is influenced by the pattern of *all* responses in the confusion matrix, whereas percentage correct only considers whether responses are on the main diagonal. Secondly, RIT factors the difficulty of the listener task into the metric so that it is not influenced by chance performance level. This allows confusion matrices of different sizes to be compared in a straightforward way (Smith, 1990). Thirdly, the RIT metric is a normalised measure of stimulus-response covariation that is free from listener response bias (Miller and Nicely, 1955).

Consonant identification performance is summarised in this paper with an error metric defined as $E = 1 - RIT$. A value of $E = 0$ indicates complete consistency in the participant's responses, whereas a value of $E = 1$ indicates a random response pattern.

III. EXPERIMENT 1: COMPENSATION FOR REVERBERATION IN CONSONANT IDENTIFICATION

Experiment 1 asks whether perceptual compensation for the effects of reverberation is apparent in a consonant identification task using speech produced by a range of different talkers and with varying speech contexts. To avoid ceiling effects in listener performance, the speech stimuli were low-pass filtered prior to their convolution with room impulse responses. Miller and Nicely (1955) have shown that cues to place of articulation are severely degraded in low-pass filtered speech, causing listeners to make more confusions. Additionally, Watkins *et al.* (2011) found that listeners gave more perceptual weight to high-frequency bands in

their 'sir-stir' experiments, partly because the temporal envelopes of the two test-words differ the most at high frequencies. Hence, by low-pass filtering speech stimuli at a range of cutoff frequencies, we aim to find a suitable operating point at which compensation for reverberation may be observed in our experiment. We expect that perceptual compensation will not be apparent in the lowest cutoff conditions, both because consonant identification is likely to be poor and because such filtering removes the temporal envelope information at the higher auditory frequencies, which tends to be more effective in compensation.

If perceptual constancy occurs in the consonant identification task, then it should become apparent in the following way. In conditions where a test-word is reverberated at the 'far' distance and a context is reverberated at the 'near' distance, listeners will make more confusions than in conditions where both parts of a trial's phrase are reverberated at the 'near' distance. However, the number of confusions caused by 'far' reverberation of a test-word should be reduced (i.e., compensation will be effected) in conditions where the context is also reverberated at the 'far' distance.

A. Stimuli

Eighty AIC utterances were selected, including the four test-words ('sir', 'skur', 'spur' and 'stir') each spoken by 20 talkers (12 male, 8 female). The utterances were segmented using Praat software (Boersma and Weenink, 2010), and word-boundaries were used to locate the context and test-word portions of the trial's phrase. Five versions of each phrase were created by low-pass filtering with an 8^{th} order Butterworth filter at cutoff frequencies of 1, 1.5, 2, 3 and 4 kHz (cf. Figure 3 of Miller and Nicely, 1955, which motivated this choice of cutoff frequencies).

Matched and mismatched reverberation-distance conditions were then created for each filtered phrase, following the method of Watkins (2005a,b). The context and test-word portions were isolated (zero-padded to retain the correct temporal alignment, as illustrated in Figure 1), allowing them to be independently convolved with either the 'near' or 'far' impulse

response as required. The resulting waveforms were scaled appropriately and summed to give same- or mixed-distance phrases, again as indicated in Figure 1. The near-near *context-test* condition and far-far condition were calculated first, and their root-mean-square (RMS) levels were equalised. Amplitude scaling factors were then derived for the context and test portions and these were applied to the mixed-distance phrases, resulting in stimuli for the near-far and far-near conditions that had the same RMS as the same-distance stimuli.

Finally, each signal was convolved with an impulse response that inverted the frequency characteristic of the Sennheiser HD480 headphones through which the stimuli were presented, and the signals were scaled *en masse* to be saved as WAV files without clipping. The set of sound files for Experiment 1 thus comprised 1600 stimuli (20 talkers $\times$ 4 test-words $\times$ 5 filter cutoff frequencies $\times$ 2 context distances $\times$ 2 test distances).

B. Procedures

The experiments reported in this study were approved by the local ethics committee, and informed consent was obtained for all participants. Sixty listeners without obvious or reported hearing deficiencies participated in the experiment. The group was a mixture of students and staff who were fluent native or non-native speakers of English. A sixth of the participants were recruited informally from the University of Sheffield's Department of Computer Science, and were not paid. The remainder responded to a university-wide email requesting volunteers, and were compensated for their time. A further 8 people completed the listening test but were discounted from further analysis since they did not meet the inclusion criterion (above 90% correct responses for the 4 kHz filter cutoff condition when both context and test-word were reverberated at the 'near' distance).

Each participant heard every one of the 80 selected AIC phrases just once; thus the test-word, the sequence of context words and the talker varied unpredictably for each trial that a listener heard. Stimuli were partitioned evenly among listeners to ensure that artifacts such as the association of a test-word with its context sentence were avoided, i.e. the 20 versions

(4 distances × 5 filters) of a given phrase were heard by different people. Participants were presented with each of the four word-initial consonant conditions at every combination of reverberation distance and filter cutoff frequency (4 consonants × 4 distances × 5 filters = 80 trials). The appropriate stimulus set was gathered for the participant, and its order randomised immediately prior to presentation.

Listeners were seated individually in a sound-attenuating booth (IAC single walled), and sounds were presented monaurally to the left ear over Sennheiser HD480 headphones at a maximum RMS level of 48 dB SPL (measured with an averaging time of 1 second). Before the experiment began there was a familiarisation session to allow the participant to become comfortable with the computer interface and the task.

Stimuli were presented by an iMac computer running Matlab v. 7.5 (R2007b) software through an M-Audio Firewire Audiophile sound interface. Each trial consisted of a speech context with an embedded test-word. Listeners identified the test-word with a click of the computer's mouse, positioned while looking through the booth's window at 'sir', 'skur', 'spur' or 'stir' alternatives displayed on the computer's screen. This click also initiated the subsequent trial. Stimuli were presented in a randomised order in a single session lasting approximately 6 minutes.

C. Results

Table I shows summary confusion matrices obtained from the data of all participants for the 4 kHz lowpass filter condition. Consonant identification is very robust to the low levels of reverberation present in the near-near *context-test* condition. However, confusions are frequent when more reverberation is added to the test-word alone (the near-far condition). The three most numerous confusions are 'stir', 'spur' and 'skur' being mistaken for 'sir'. However, when the preceding context is also reverberated at the 'far' distance (the far-far condition), the majority of these confusions are resolved, indicating perceptual compensation.

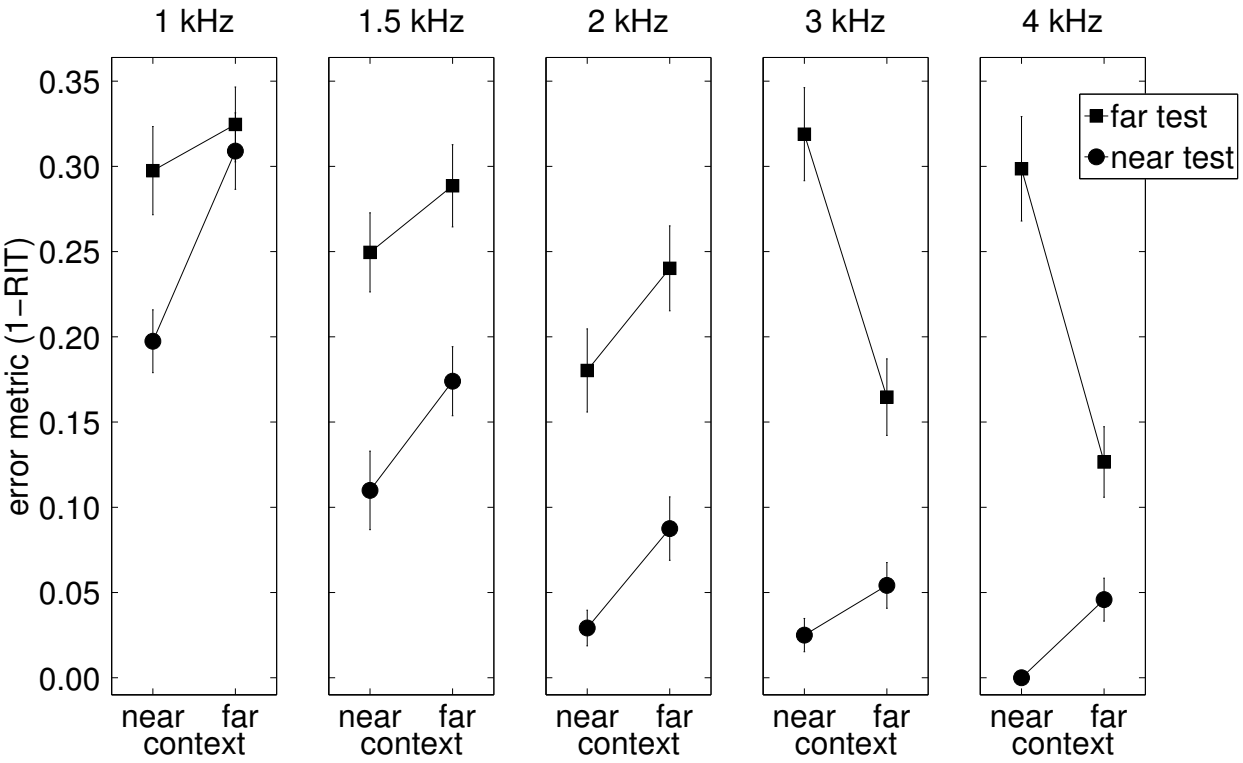

FIG. 2. Mean and standard error of the 60 participants' 1−RIT scores at the five low-pass filter conditions of Experiment 1. Compensation for reverberation is apparent in the downward-sloping upper line of the 3 and 4 kHz filter conditions. In these two conditions, an increased level of reverberation in the context (resulting from the increase in context distance) brings about an improvement in the identification of the far-distance test-words.

1. Perceptual compensation for reverberation

For numerical analysis, participants' responses were recorded individually in confusion matrices, and analysed in terms of their information transfer characteristics as described in section II.C. Figure 2 shows the mean and standard error of the 1-RIT scores at each reverberation distance and each filter condition. A three-way repeated measures analysis of variance (ANOVA) was performed on participants' arcsine-transformed RIT scores (Kirk, 1968) using IBM SPSS Statistics 20 software. All factors were within-subject; two factors had two levels each (context distance and test-word distance) and the third had five levels (filter cutoff frequency). Mauchley's test showed no cases of violation of sphericity.

A monaural perceptual compensation effect is apparent at the 3 and 4 kHz filter cutoff conditions shown in Figure 2. In these two conditions, a far-distance test-word is less often confused when it is preceded by a far-distance context than when it is preceded by a near-distance context. For filter cutoff frequencies of 2 kHz and lower, however, a far-reverberated context did not aid identification of the far-reverberated test-word. This pattern of results was indicated in the data by a three-way interaction among the factors for filter condition, test distance and context distance, where $F_{(4,236)} = 5.94$, and $p < 0.001$. Two main effects and three two-way interactions in the analysis, described below, largely arose from this higher-order interaction.

As anticipated, consonant identification performance is best in the near-near condition at each filter cutoff. This can be more clearly seen in Figure 3, in which the data from Figure 2 is redrawn as a conventional line plot. Increasing the distance of the test-word from 'near' to 'far' consistently increases the consonant identification error, giving a main effect of the test-word's distance with $F_{(1,59)} = 306.62$, and $p < 0.001$. In addition, consonant confusions became more prevalent as the cutoff frequency of the lowpass filter was reduced, giving a main effect of the filter cutoff frequency with $F_{(4,236)} = 53.99$, and $p < 0.001$. An interaction of these factors was also found, with $F_{(4,236)} = 9.16$, and $p < 0.001$, indicating that consonant confusions resulting from an increased level of test-word reverberation are more prominent when higher-frequency information is retained in the signal.

A two-way interaction between the factors for context distance and test-word distance, with $F_{(1,59)} = 28.32$, and $p < 0.001$, indicated that when the far-reverberated context did cause an improvement in consonant identification, this is confined to the far-reverberated test-words. As described above, however, the effect of context reverberation varies across the filter conditions, which showed as a significant interaction of context distance and filter cutoff frequency, with $F_{(4,236)} = 9.78$, $p < 0.001$. There were no other significant F ratios.

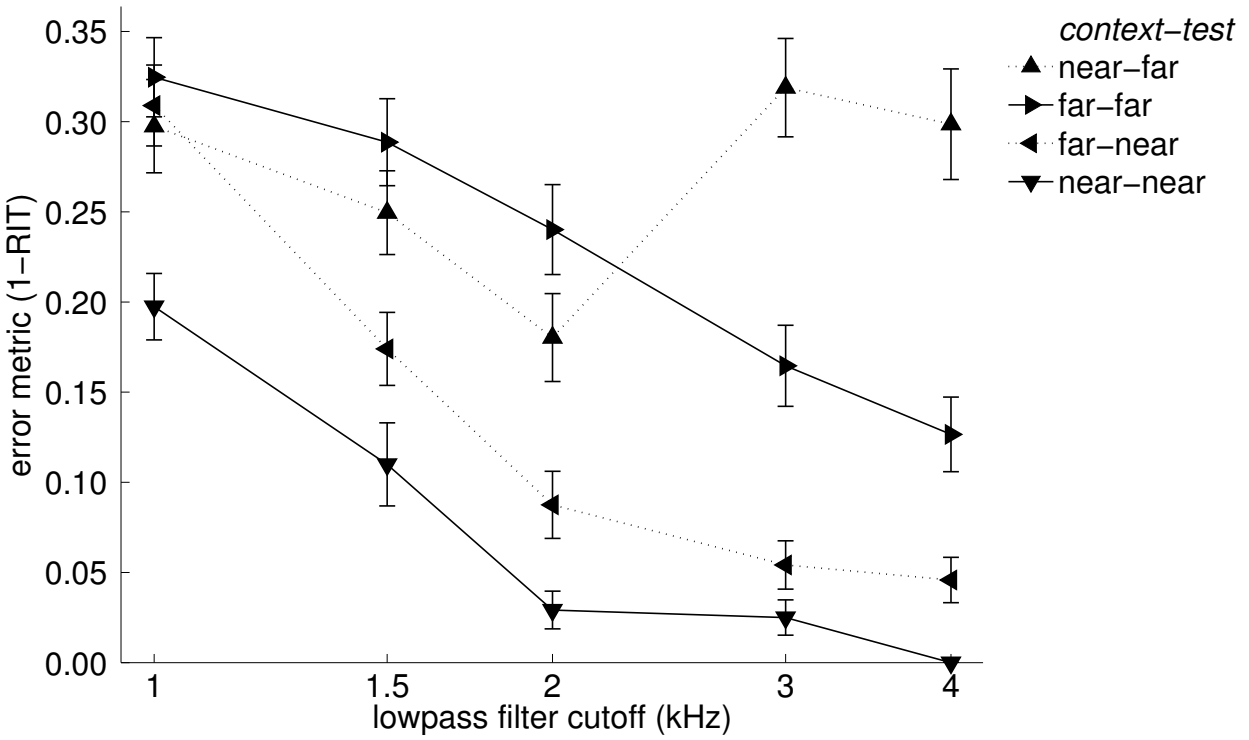

FIG. 3. Data of Figure 2 replotted to show the effect of lowpass filtering on each *context-test* condition. Consonant identification error decreased monotonically with increasing lowpass cutoff frequency, except when the context was 'near' reverberated and the test-word was 'far' reverberated.

2. Effect of low pass filtering on the near-far condition

It is apparent from Figure 3 that consonant identification error generally decreases as the lowpass cutoff frequency increases, as would be expected from the prior literature (e.g. Miller and Nicely, 1955). However, this trend is not observed in the near-far *context-test* condition; in this condition, consonant confusions increase when more high frequency information above 2 kHz is retained. A plausible explanation for this finding stems from within-channel processing relating to perceptual compensation (Watkins *et al.*, 2011). In the near-far condition, information in the context (e.g., about reverberation tails) suggests that the level of reverberation in the test word will be low, leading to errors in consonant identification because the test word is actually reverberated at the far distance. Such er-

rors will be most prevalent at cutoff frequencies above 2 kHz because the acoustic-phonetic cues that characterise stop consonants, a dip in the temporal envelope followed by a release burst, occur predominantly at high frequencies (e.g., Allen and Li (2009) note that [t], [k] and [p] are defined primarily by features in the range 4 kHz, 1.4–2 kHz and 0.7–1 kHz respectively). In other words, it is at high frequencies that there is the biggest discrepancy between the reverberation characteristics implied by the context, and the acoustic features that are actually encountered in the test word.

To seek further support for this explanation, listeners' responses in the near-far condition were analysed for each individual consonant. We reason that this would reveal whether there is a consistent pattern of behaviour for each consonant, or whether the form of the near-far curve in Figure 3 is only apparent because the data was pooled across all consonants.

Each participant's responses at the near-far reverberation condition were therefore analysed as follows. At each filter cutoff condition, the overall 4×4 confusion matrix was refigured into four 2×2 matrices quantifying, for each consonant stimulus-response pairing, the number of hits, misses, correct rejections and false alarms. As before, participants' error-rates were then quantified from these 2×2 matrices by calculating errors in terms of information transfer (1−RIT scores).

Figure 4 shows these results. The pattern is repeated across all test-words, showing a 'pivot point' in performance at 1.5 kHz for 'spur' and at 2 kHz for the remainder. This is consistent with the fact that the burst frequency for [p] is the lowest of the consonants considered here (Allen and Li, 2009). We conclude, therefore, that the increase in error rate in the near-far condition apparent in Figure 3 is not an artefact in the data caused by pooling across all consonants tested. Rather, it is most likely caused by conflicting high-frequency cues in the context and test-word of the phrase, which reduce the efficacy of within-channel compensation mechanisms.

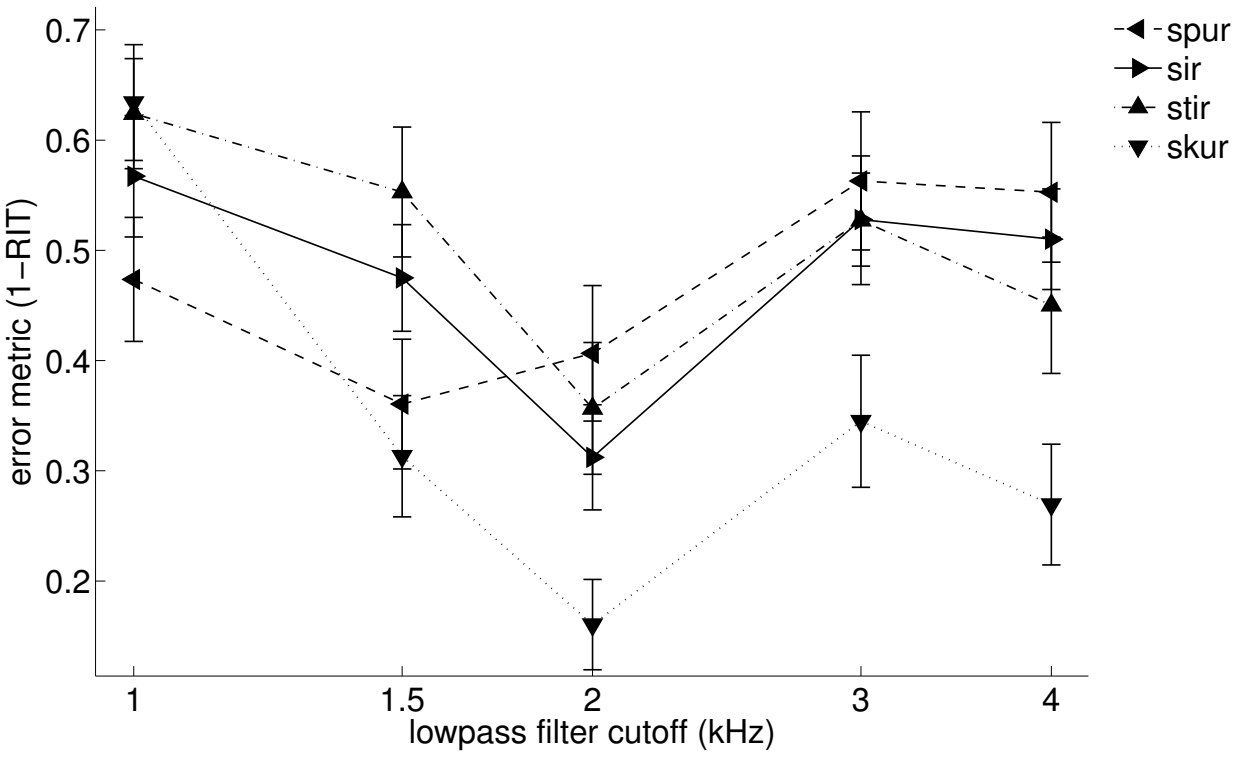

FIG. 4. Data of the near-far condition in Figure 3 replotted to show the cutoff filter effect on each test-word's responses. All test-words showed a similar pattern of performance, with a 'pivot point' at 1.5 kHz for 'spur' and at 2 kHz for the remainder.

D. Interim discussion

In the 'sir-stir' continuum experiments, Watkins (2005b) attributed his results to a monaural mechanism of perceptual compensation. Here, we have used speech material in which the acoustic-phonetic cues are much more variable as listeners heard different phrases on each trial. There is also much more temporal uncertainty in the stimuli of the current experiment, through uncertainty about the temporal location of the test-word, with context durations ranging from minimum 0.31 s to maximum 0.97 s. This uncertainty reduces listeners' sensitivity in other types of task, such as signal detection (Egan *et al.*, 1961) and gap detection (Green and Forrest, 1989). Despite this variability in our signals, we have observed perceptual compensation for the effects of reverberation that is qualitatively similar to that found by Watkins. It therefore seems that the compensation mechanism

will most likely be effective in everyday listening, where levels of stimulus uncertainty are generally high.

Experiment 1 found that perceptual compensation is not apparent when high-frequency components are removed from the speech signal. This is consistent with the finding of Watkins *et al.* (2011) that perceptual compensation is effected in a band-by-band manner, and that high-frequency bands carry more perceptual weight than low-frequency bands in conferring the 'sir' vs. 'stir' distinction. It seems likely that similar mechanisms of band-by-band processing underlie the effect of lowpass cutoff frequency on the 'sir-skur-spur-stir' distinction investigated here. However, the possibility remains that listeners are unable to compensate for the effects of reverberation at the lowest lowpass cutoff frequencies because the phonetic content of the speech signal was severely degraded (cf. Miller and Nicely, 1955).

IV. EXPERIMENT 2: AN INTRINSIC EFFECT

In the conceptual model discussed above, information for the compensation mechanism is gathered solely from the preceding context. However, Watkins and Raimond (2013) note that in addition to effects of such 'extrinsic' information, there are 'intrinsic' effects that arise through information from the test-word itself. In that study, reverberant test-words were presented in isolation (i.e., without a context). The following experiment asks whether 'intrinsic' information effects any compensation in conditions more similar to everyday listening.

The context phrase preceding the test-word was subjected to three different treatments: near-distance reverberation and far-distance reverberation (replicating conditions in Experiment 1), and a silencing treatment which removed the preceding context cues and gave conditions similar to those presented to listeners in Watkins and Raimond (2013) and Nielsen and Dau (2010). The test-word itself was first reverberated at the near or far room distance as before, and was subsequently gated in some conditions following the method of Watkins and Raimond. By shortening the reverberation tail that follows the test-word's final vowel,

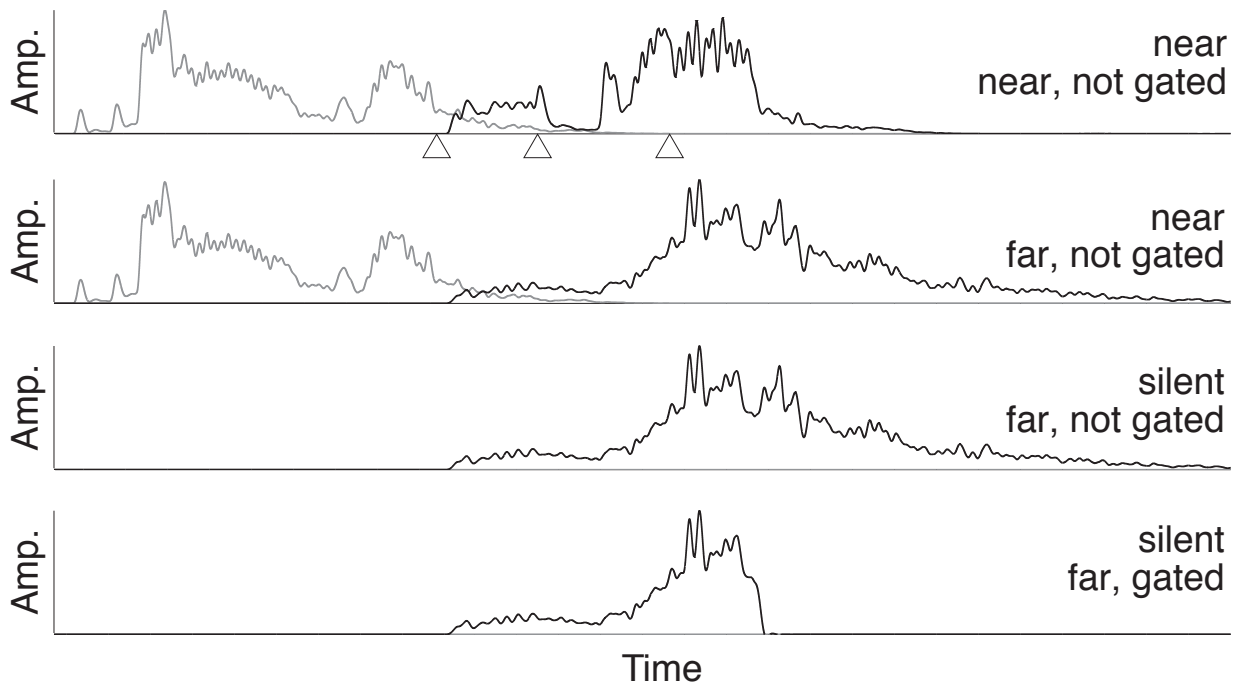

FIG. 5. Illustration of selected stimulus conditions in Experiment 2, here applied to the same phrase that was previously displayed in Figure 1. Context phrase *(light line, upper label)* and test-word *(dark line, lower label)* were independently reverberated at 'near' or 'far' distances as before *(upper half)*. In other conditions the context was silenced *(lower half)* and the test-word was gated *(bottom panel)*. Other details are the same as Figure 1.

the intrinsic information content was reduced. Selected stimuli are displayed in Figure 5 to depict the separate context silencing and test-word gating processes.

To avoid a likely ceiling effect in far-distance conditions, the effect of gating was only tested in the near- and silent-precursor conditions. However, far-precursor trials were included in a listener's set in order to maintain uncertainty about the level of reverberation. We expect listeners to make fewer consonant confusions when the tail is present and more confusions when the tail is removed by gating.

If an effect of gating *is* apparent, then it would suggest that the reverberation tail following the test-word's final vowel contributes to perceptual compensation even though it occurs some time *after* the consonant in the test-word, and that the conceptual model must

be updated accordingly to include intrinsic information sources.

It should be noted that this experimental design does not gauge the full size or importance of *all* intrinsic sources of reverberation information: the gating operation currently evaluates only the contribution of the vowel-end tail. Other indicators of reverberation in test-words are not evaluated, notably the tails that follow the test-word's consonant, which are closer in time to the crucial frication part of the sound.

A. Stimuli

The 4 kHz lowpass filter condition was selected for use in Experiment 2, since it gave a clear perceptual compensation effect in Experiment 1. Phrases were similar in form to those of the earlier experiment, however, to facilitate independent manipulation of the reverberation elicited by the test-word and the context, all phrases were truncated to remove the final context word. The stimuli with non-silent contexts used in Experiment 2 therefore had the form:

[CW1][CW2][TEST].

The set of test-word vowels was expanded further in this experiment, allowing more consonant confusion data to be gathered from each participant. Since the perception of [p] and [k] (but not [t]) depends on the following vowel (Liberman *et al.*, 1952), care was taken to ensure that the following vowel would have similar effects across the new set of test-words. Since coarticulatory variation is not prominent among front vowels, the vowels [eɪ, iː, ɛ, ɪ, æ, ɜ˞] were selected from the AIC, the last of which was the vowel used in Experiment 1. The vowel labelled [ɑ] was rejected because it was spoken inconsistently by the 20 talkers, with frequent mergers of the two back vowels [ɑ] and [ɒ] (Wright, 2005). The vowel [oʊ] was not included since it was the only remaining back vowel. Using the same initial consonant conditions as in Experiment 1, the experiment thus employed 480 AIC utterances (20 talkers $\times$ 4 consonants $\times$ 6 vowels).

Precautions were taken to position word boundaries so that the speech sounded naturally spoken after truncation and reverberation. The process of locating word boundaries was partially automated due to the large number of utterances involved. First, the AIC transcripts were expanded to phone sequences using the Carnegie Mellon University pronunciation dictionary (CMU, 2010). A hidden-Markov model-based automatic speech recognition system (HTK, 2010) was then used in conjunction with TIMIT-trained monophone acoustic models (Lee and Hon, 1989) to force-align each phone sequence with its corresponding speech signal and thereby identify the test and context regions. To overcome quantisation errors due to the 10 ms frame rate of the recogniser, the word boundaries were subsequently checked using Praat (Boersma and Weenink, 2010) and amended by hand where necessary.

Reverberation processing for the same- and mixed-distance phrases was undertaken as described in Section III.A, two examples of which are illustrated in the upper two panels of Figure 5. In the third panel, the preceding context words CW1 and CW2 were omitted in silent-context conditions, and silent intervals, SIL, of equal duration were introduced so that the phrases now comprised [SIL][SIL][TEST]. This further increased the uncertainty in the temporal location of the test-word since not only did the preceding context vary in duration for each phrase (ranging from 0.23 s to 1.24 s), but additionally any quasi-semantic cues from the preceding pronoun and verb were now removed.

Gated test stimulus conditions emulated those of Watkins and Raimond (2013), as illustrated in the final panel of Figure 5. A gating function was created using the right-hand-side of a Hann window of 10 ms duration, and was applied to 'near' and 'far' reverberated versions of the test-word, with the function time-aligned to begin its descent at the end of the test-word. Hence, the reverberant tail following the test-word was cropped off without shortening the word beyond its unreverberated duration.

Scaling factors were calculated across CW1, CW2 and TEST in order to ensure that the level of context and test portions maintained their balance in mixed-distance conditions. Finally, the twelve versions of each phrase were equalised in RMS level, the headphone correction was applied and the sound files were saved as previously described. The set

of sound stimuli for Experiment 2 thus comprised 5760 sound files (480 AIC phrases × 3 context conditions × 2 test-word distances × 2 gate conditions).

B. Procedures

Sixty participants were recruited from the student and staff population of Sheffield University, and were compensated for their time. A further 10 people took part but were discounted from subsequent analysis. In one case this was due to a reported hearing impairment. In the remaining 9 cases this was through failure to meet the inclusion criterion at the control condition (achieving above 90% correct responses in (not-gated) near-context, near-test conditions).

Stimuli were partitioned among participants as previously described to avoid any association between test-word and context phrase that might otherwise aid identification of the test-word. Each participant heard 40 phrases in each of the 12 experimental conditions. Vowels were divided evenly across the listener group, with participants hearing every test consonant either once or twice at each reverberation distance. In cases where listeners heard the same test consonant twice, the two instances were spoken by different talkers. Participants were not required to identify the test-word's vowel. Rather, they identified the initial consonant cluster only by choosing among buttons labelled 's', 'sk', 'sp' or 'st'. Stimuli were presented to the participants in a randomised order in a single session. Participants were instructed to take short breaks whenever needed, and the experiment was typically completed in around 25 minutes. Other aspects of stimulus presentation were the same as described in section III.B.

C. Results

As before, participants' responses were recorded in confusion matrices and analysed in terms of their information transfer characteristics. The main findings of Experiment 1 were replicated, as shown by the mean and standard errors of participants' 1-RIT scores in

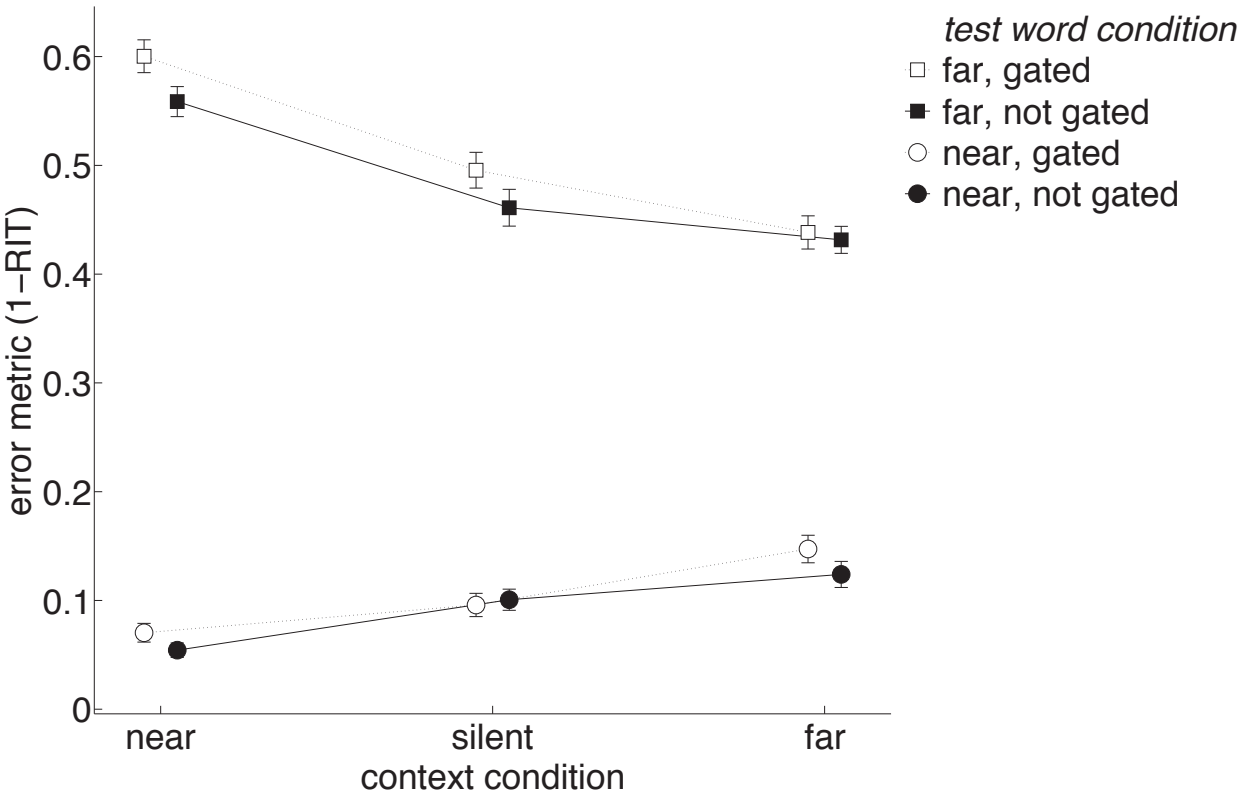

FIG. 6. Mean and standard error of the 60 participants' 1−RIT scores in Experiment 2. Conditions in which the reverberation tail following the test-word was removed by gating are shown with white markers. Conditions that preserve intrinsic information are presented with black markers. Extrinsic compensation seems to effect an overall reduction in consonant confusions between near-far and far-far *context-test* conditions, which replicates the corresponding main effect in Experiment 1. This reduction is seen for both gate conditions.

Figure 6. Firstly, an increase in test-word distance again brought about a large increase in the number of consonant confusions that participants make. Secondly, for both gated and not-gated stimuli, extrinsic compensation at the far-distance context condition resulted in a reduction in the number of consonant confusions recorded in comparison with the near-distance context condition. Since the final context word was omitted from the phrases used in this experiment, extrinsic information from the context portion *following* the test-word was clearly not necessary for perceptual compensation to occur.

A number of potential confounds preclude analysis along the lines of our earlier experiment (e.g., using a 3-way ANOVA with factors for test distance, context distance and gate condition). Looking left to right in Figure 6 we anticipate an increase in consonant confusions due to temporal uncertainty of the test-word as we move from near to silent contexts, since the silent context cannot cue the location of the test-word. Continuing towards the right from silent to far contexts, the reduced degree of temporal uncertainty at far might suggest a decrease in error; however, consonant confusions will likely increase due to overlap masking (Nábělek *et al.*, 1989) from the context in the far condition. Concurrently, extrinsic compensation effects would be expected to decrease the overall error rates of far test-words as we move from left to right.

Instead, for each context and gate condition we measure the difference between participants' scores for the two levels of test-word reverberation (so that constancy is greater when this difference is small). Difference scores were calculated for each participant as their RIT error at the far distance test-word minus their RIT error at the near distance test-word, the means and standard errors of which are shown in Figure 7 (*left*). A 2-way repeated measures ANOVA (all within-subject factors) was thus performed on participants' difference scores, using one factor for test-word gate condition (gated, not-gated) and a second factor for preceding context condition (near, far, silent). Mauchley's test showed that conditions of sphericity were met. A large, extrinsic effect showed in the data as a significant effect of context, with $F_{(2,118)} = 90.61$, and $p < 0.001$. Seen in Figure 7, the general reduction in near-far test-word difference when moving from left to right suggests that compensation increases in silent-context conditions, and increases further for far-distance contexts. There were no other significant F ratios.

It was argued above that gating effects were unlikely to be apparent in conditions with far-distance contexts due to a ceiling effect. These conditions therefore were excluded from the planned comparison in Figure 7 (*right*) which pooled data for the remaining silent- and near-context conditions. A paired-samples t-test revealed that there was some effect of test-word gating, with $t_{(119)} = 2.43$ and $p = 0.017$. Thus, results in the near- and silent-context

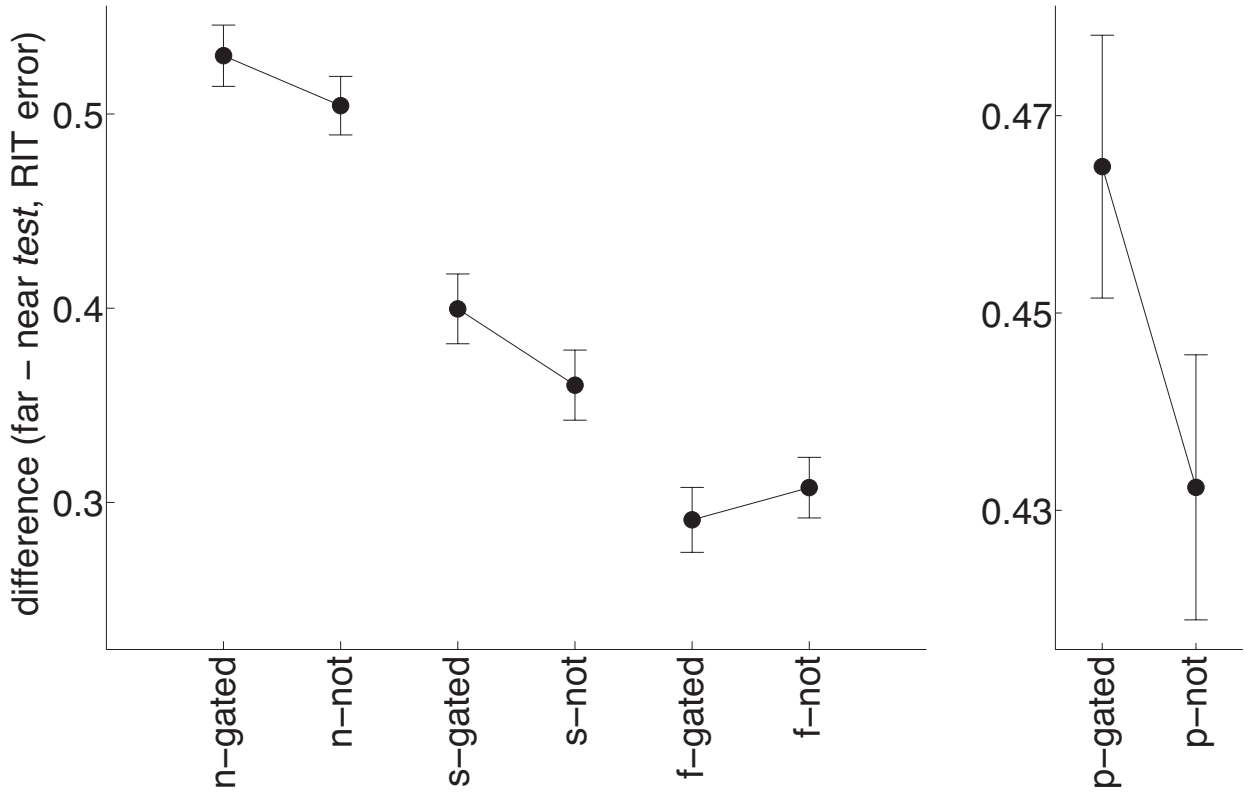

FIG. 7. *Left* – Mean and standard error of the 60 participants' difference scores (RIT error at far distance test-word minus RIT error at near distance test-word) for near (n), silent (s) and far (f) contexts at each gate condition in Experiment 2. *Right* – Pooled (p) results show some effect of gating on the test-word's vowel, indicating that intrinsic factors have probably helped to disambiguate far distance test-words in the near- and silent-context conditions. Note that the ordinate scales of the two panels differ.

conditions indicate that there is a role for intrinsic information which seems to help listeners identify reverberant test-words.

D. Interim discussion

The silent-context conditions in our experiment were intended to further investigate findings of Nielsen and Dau (2010) and Watkins and Raimond (2013) where test-words from the 'sir-stir' continuum were preceded by silence. The 'modulation masking' theory

suggested by Nielsen and Dau proposed that the dip cueing the [t] in a reverberant 'sir-stir' continuum test-word could be made less apparent (i.e., masked) by a preceding context, provided that that context contained a sufficient degree of modulation. The near context, with its relatively large amount of modulation, would induce substantial masking of the [t] in the far-reverberant test-word, and thus promote more 'sir' responses from the listeners (from which we may infer a greater degree of confusion in the AIC data). The far context on the other hand has a smaller degree of modulation which would promote less masking of the [t] dip, and would permit more 'stir' responses (or fewer consonant confusions). For silent contexts, where there is no modulation forward masking from the preceding context, Nielsen and Dau's proposal would predict a well-defined plosive dip, resulting in still fewer confusions in the AIC data. As in Watkins and Raimond (2013), however, a different pattern emerges in our listener results: far test-word consonant confusions were indeed less frequent for silent contexts than for near-distance contexts, but confusions were actually reduced still further by the presence of a far-distance context. This result may additionally cast earlier data in a new light, specifically where silence has been used as a 'control' condition to contrast against a reverberated speech carrier (cf. for example Ueno *et al.*, 2005; Nielsen and Dau, 2010; Brandewie and Zahorik, 2013), since it indicates that some compensation arises in silent-context conditions.

Our data supports the notion that perceptual compensation is influenced by both intrinsic and extrinsic factors. The overall extrinsic compensation effect seen in Experiment 1 is replicated here, seen in the reduction in error for far-distance test-words when they are preceded by a far- rather than near-distance context (cf. Figure 6). However, the reduction in far test-word consonant confusions when silent contexts are present (rather than a near-distance context) cannot be attributed to the extrinsic effects elucidated in Experiment 1 since the preceding context cues have been removed. Rather, we might attribute this reduction to an intrinsic influence. Further, by examining the intrinsic influence from tails at the end of the far-reverberated test-word's vowel we find that errors tend to be reduced in not-gated conditions with near and silent contexts. This suggests that the test-word's tail

influences the identity of the preceding consonant when intrinsic and extrinsic information are placed in conflict; in other words, if listeners are presented with an ambiguous stimulus, they use intrinsic information to help resolve the uncertainty. We might also suppose that tails from the test-word's initial consonant would be a further intrinsic influence. Experiment 2 therefore indicates that although compensation for reverberation is strongly informed by extrinsic information, intrinsic sources of information should not be discounted. Indeed, the conceptual model discussed earlier should be updated to include intrinsic information from the test-word.

In both Experiment 1 and Experiment 2, the context words preceding the test-word varied in duration from trial to trial; in other words, the amount of extrinsic information available to listeners was not constant. Nonetheless, both experiments found that inconsistent reverberation in the context and test-word degraded consonant identification performance, and that this degradation could be alleviated by making the context reverberation and test-word reverberation consistent. Experiment 3 now investigates the time course of this extrinsic compensation effect by carefully controlling the amount of extrinsic information available: starting with inconsistent reverberation between the context and test-word, we ask how much consistent reverberation is needed in order for compensation to be apparent.

V. EXPERIMENT 3: TIME COURSE

Experiment 3 investigates the time course of monaural perceptual compensation by varying the duration of the context speech that is reverberated at the far-distance. In terms of the conceptual model discussed above, it asks which portions of the preceding extrinsic context are influential in determining the compensation effect. In the previous experiments, the speech context preceding the test-word was wholly reverberated at either the far- or near-distance. Here, the context speech is divided into two regions; the first part is reverberated at the near distance, and the second part (just prior to the test-word) is reverberated at the far distance. By varying the boundary between these two regions among conditions,

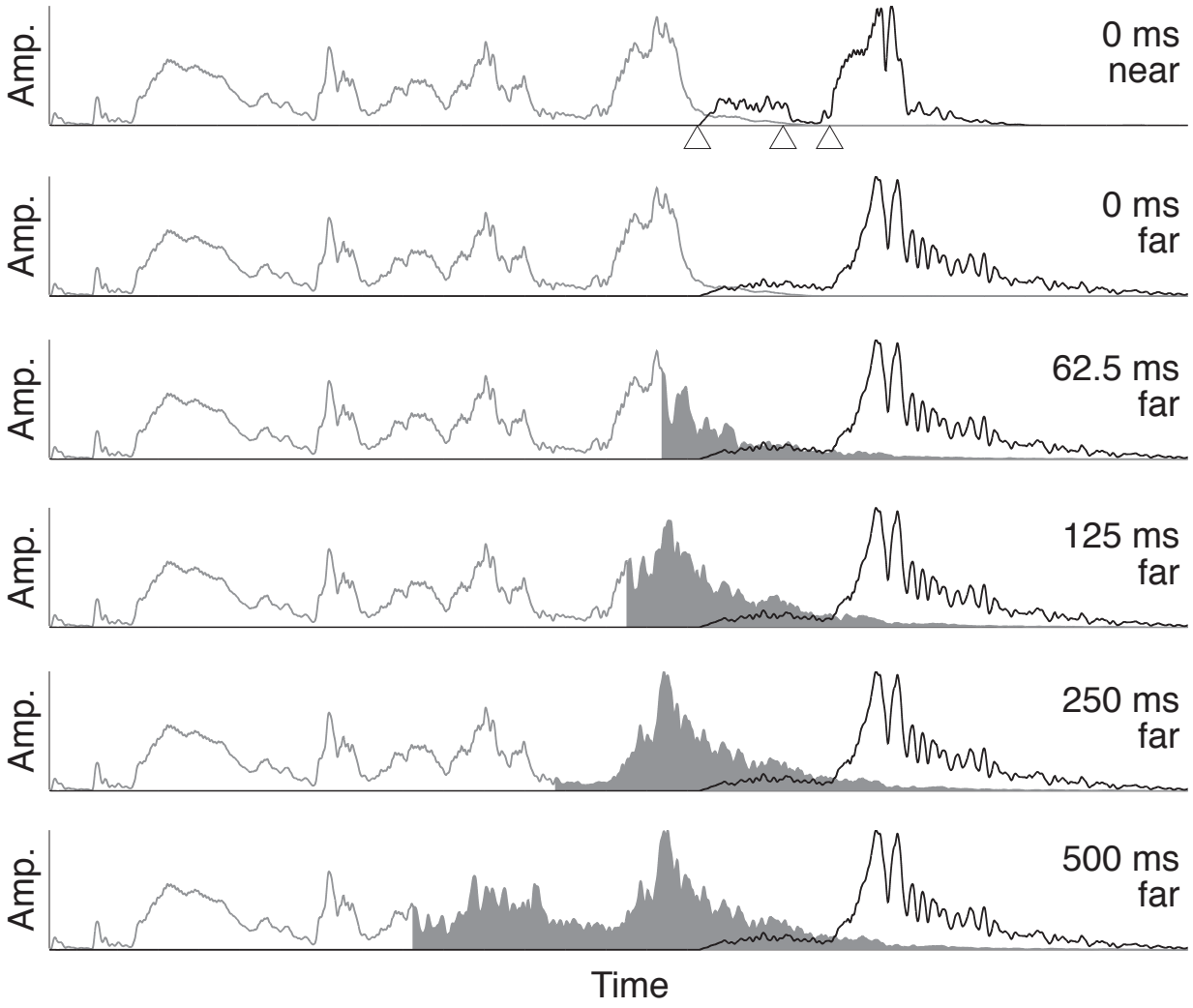

FIG. 8. Illustration of selected stimulus conditions for an Experiment 3 phrase. The test-word *(dark line, lower label)* is preceded by the context *(light line)* which is divided into an initial near-reverberated part and a subsequent far-reverberated part. The temporal position of the boundary is varied between these parts, so that less or more of the context immediately prior to the test-word is far-reverberated *(shaded area, upper label)*. The phrase differs from that used in Figure 1, but other details are the same.

we vary the amount of far-reverberated context while measuring compensation on a far-reverberated test-word. The following experiment thereby asks how compensation for the effects of reverberation may build-up.

A. Stimuli

Experiment 3 again used AIC speech material, low-pass filtered at 4 kHz as in earlier experiments. The listeners' task was simplified, with the response alternatives being reduced to a choice between 's' or 'st' for the start of the test-word, but stimulus variability was maintained by using speech from multiple talkers and by selecting five following vowels to complete the test-words: [eɪ, iː, ɛ, ɪ, æ]. Word boundaries between the test and context portions of each utterance were located as previously described in section IV.A. However, in Experiment 3 the AIC utterances were reordered and spliced so that all of the context words preceded the test-word:

[CW3][CW1][CW2][TEST].

This was done in order to maximise the duration of the context preceding the test-word, yielding phrases such as "often people determine stay". By limiting the number of corpus talkers to 10 of the available 20, the resulting 100 phrases (10 talkers × 2 consonants × 5 vowels) had preceding contexts of around one second duration or longer. Four phrases fell slightly short of this target, with context durations of 994, 979, 959 and 933 ms respectively.

The initial portion of the context phrase was always reverberated with the near-distance room impulse response. Thereafter, a portion of the context just prior to the test-word was reverberated at the far-distance. The duration of this far-distance window was nominally 0, 62.5, 125, 250 or 500 ms, as depicted in the shaded regions of Figure 8. In practice the window length was modified for each phrase so that it coincided with a zero-crossing in the audio signal. This ensured that reverberation of the context did not introduce an audible discontinuity in the signal. The duration of the far-context portion thus differed slightly from the nominal window length in almost all cases, but this variation was typically small (across the whole set of stimuli, the mean deviation from the nominal window length was 48.9 samples, corresponding to approximately 1 ms at the 48 kHz sample rate used).

The near- and far-distance portions of the context were recombined with the test-word using the RMS balancing techniques outlined in III.A, to create the same- and mixed-

distance phrases. Finally, the stimulus conditions for each phrase were equalised in overall RMS level, the headphone correction was applied, and the sound file was saved as previously described. The set of sound files for Experiment 3 thus comprised 1000 stimuli (100 phrases $\times$ 2 test distances $\times$ 5 context window durations).

B. Procedures

Forty participants were again recruited by university-wide email, and were compensated for their time. A further 5 people completed the experiment but were discounted from analysis. Two of these participants reported hearing losses, which contributed to considerable difficulties recognising test-words in all conditions. The remainder were excluded because they did not meet the inclusion criterion (above 90% correct responses for near-distance test-words at the 0 ms far-distance window condition).

Stimuli were partitioned among groups of 10 participants to ensure different versions of a given phrase (2 test distances $\times$ 5 context window durations) were heard by different people, avoiding any association between test-word and context phrase that would otherwise aid test-word recognition. Talkers were divided evenly across the listener groups, so that each participant heard 10 phrases in each of the 10 experimental conditions, each repeated 4 times. Every test vowel was used in each condition, once with a preceding [s] and once with [st]. As before, participants were not required to identify the test-word's vowel, but identified the initial portion by clicking on either an 's' or 'st' alternative on the computer's screen. Stimuli were presented in a randomised order in a single session lasting around 20 minutes, with breaks offered as desired. Other aspects of stimuli presentation were carried out as described in section III.B.

C. Results

Participants' responses were analysed in terms of the proportion of 's' responses at each test-word distance and context-window condition as shown in Figure 9. A two-way repeated

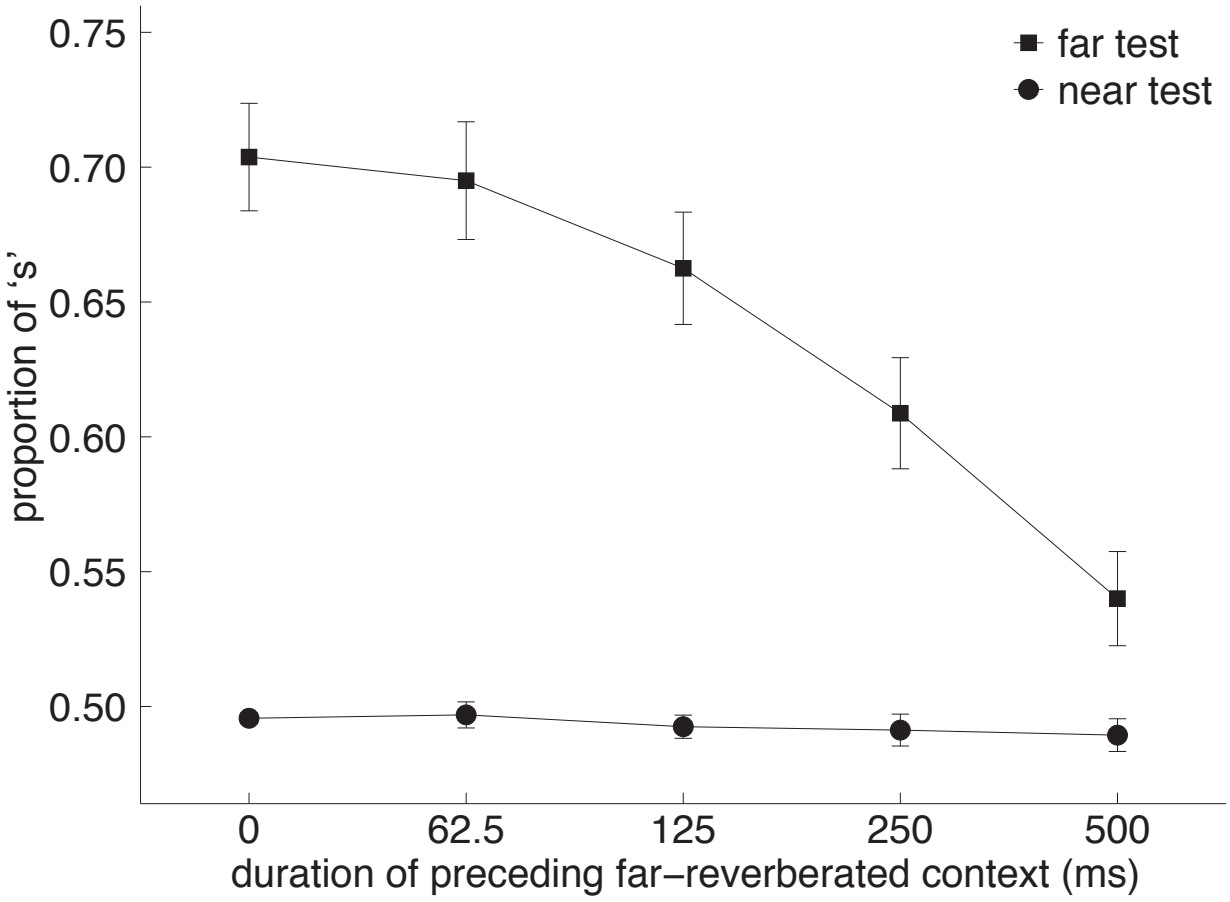

FIG. 9. Mean and standard error of the 40 participants' scores in Experiment 3. The lower line shows near-distance test-word scores, where the proportion of initial 's' reports is close to 0.5 and the data show no dependency on the duration of the far-reverberated context. The upper line shows the far-distance test-word scores. For the zero-length far-context window condition, where the test-word has more reverberation than the context, listeners often misclassified [st] as 's'. As the duration of the far-reverberated part of the context is increased, fewer misclassifications are made and the proportion of 's' reports decreases.

measures ANOVA was performed using one factor for test-word distance with two levels (near, far) and a second factor with five levels for the duration of the far-distance context preceding the test-word (0, 62.5, 125, 250, 500 ms). Once again, Mauchley's test showed no violation of sphericity. The two-way interaction between factors for test distance and far-reverberated context duration was significant ($F_{(4,156)} = 22.13$, $p < 0.001$) as were both

main effects; test-word distance ($F_{(1,39)} = 75.96$, and $p < 0.001$) and far-context duration ($F_{(4,156)} = 25.55$, and $p < 0.001$). A linear trend test (using a least-squares method) across the log-spaced window-duration conditions showed a linear decrease in the number of 's' responses with increasing duration of the far-reverberated context-window ($F_{(1,39)} = 82.90$, $p < 0.001$).

D. Interim discussion

Listener response data in Experiment 3 confirms our conceptual model's assertion that compensation for the effects of reverberation builds-up through time. Figure 9 shows a clear trend; in a task where listeners were required to determine whether the test-word started with [s] or [st], the number of 's' responses decreases as the duration of the far-reverberated portion of the context is increased. Complementing work on binaural compensation mechanisms by Brandewie and Zahorik (2013), it appears from our listener data that the monaural compensation mechanism acts on a similarly rapid timescale.

The experiment has an important limitation in that it is unable to determine whether a further improvement in consonant identification would occur if the duration of the far-reverberated context were to exceed 500 ms. It would have been possible to pad short utterances with additional speech material in order to produce longer contexts. However, this would have disrupted the consistent form of the utterances (which otherwise had exactly three context words preceding the test-word) and might have misdirected the listeners' attention towards the context and away from the test-word (cf. Ueno *et al.*, 2005, Experiment 1).

A further problem arises through selecting the 10 talkers with the fewest short utterances in the corpus. In doing so, it is likely that talkers with slower speaking rates were preferentially selected. In 'sir-stir' phoneme-identification, near-far conditions give a greater shift in the phoneme boundary at faster speech rates (see Watkins, 2005b, Figure 3), so a still larger perceptual compensation effect than that observed in Figure 9 might have been apparent

with faster-speaking talkers. However, inclusion of the slower talkers gives conditions that are more similar to every-day listening where talkers tend to slow down in reverberant rooms (Black, 1950) and listeners tend to prefer slower speech (Moore *et al.*, 2007).

VI. GENERAL DISCUSSION AND CONCLUSIONS

The experiments reported here complement previous studies by providing further evidence that *monaural* exposure to a reverberant environment is sufficient to bring about a significant improvement in consonant identification, here in read speech from 20 talkers. These effects, measured across 160 left ears, must be independent of any interaural processing attributable to binaural hearing, and cannot be directly attributed to 'echo-suppression' resulting from precedence effect buildup (cf. Zahorik *et al.*, 2009; Brandewie and Zahorik, 2010). Rather, our data are consistent with a compensation mechanism that has been attributed to temporal envelope constancy (Watkins *et al.*, 2011; Kuwada *et al.*, 2012; Srinivasan and Zahorik, 2014), which appears to enhance the amplitude modulation in reverberant signals (Zahorik *et al.*, 2012).

The current work, moreover, has revealed that these monaural mechanisms are relatively rapid, with the majority of consonant confusions being correctly resolved after only half a second of an appropriate preceding context. Previous experiments have reported compensation for reverberation despite various distortions to the fine-structure of the room's reflection pattern, for example by using impulse responses for the context and test-word reverberation that were recorded in different rooms (Watkins, 2005b), or by reversing the polarity of a randomly selected half of the samples in the impulse response (Watkins *et al.*, 2011). Such results contrast with the long-term binaural learning effect reported by Shinn-Cunningham (2000), where results seem to be due to more subtle learning of a particular room's detailed characteristics. Rather, the monaural constancy mechanism appears to establish a less subtle, but more rapid calibration to a new listening environment.

The starting point for this study was the following simple conceptual model of percep-

tual constancy: listeners use information obtained from the acoustic context preceding a test sound, and accumulate this information over a period of time. As noted above, the current study has clarified the timescale involved. However, our conceptual model also requires some refinement, since we have presented evidence that compensation is not only mediated by the preceding context, but also by information originating from within the test-word itself. Watkins and Raimond (2013) previously reported such an intrinsic compensation effect. However, their experiment used test-words presented in isolation. Given that natural conversation often consists of extended turns rather than isolated words, in this paper we assessed the contribution of intrinsic information when test-words were presented after a preceding speech context. Since we found a small effect of gating, even when a preceding context was present, it seems likely that intrinsic information contributes to monaural compensation in everyday listening situations.

Currently, our work has not yet demonstrated that the monaural constancy effect generalises to the identification of a full range of natural speech sounds, as has recently been shown for the binaural constancy effect (e.g., Brandewie and Zahorik, 2013; Srinivasan and Zahorik, 2013). Nonetheless, our findings are likely to be ecologically relevant because the consonants studied here appear so frequently in everyday speech. Mines *et al.* (1978) report that [t, s, k, p] account together for 15.28 % (respectively: 5.78, 4.61, 3.10, 1.79 %) of all phonemes encountered in casual conversational American English (including vowels). Moreover, since the consonants studied here are among those most vulnerable to the effects of reverberation (Gelfand and Silman, 1979; Nábělek *et al.*, 1989; Drullman *et al.*, 1994), our experiments address the very parts of the speech signal that are the most troublesome to hear in real reverberant listening situations.

While the speech material in our final experiment examined temporal effects of the preceding context reverberation, its phonetic content was not studied. Frequency regions around 4 kHz are likely to contribute most significantly to identification of the [t] when it is present in test-words (Allen and Li, 2009), and since compensation for reverberation appears to work in a band-by-band manner, the level of compensation achieved would be

similarly dependent upon these important frequency regions in the neighbouring context words (Watkins *et al.*, 2011). It is likely, therefore, that a context rich in sibilants and stops (e.g., "first people detect") would promote a higher degree of compensation for effects of reverberation on the test-word's stop consonant than would a phrase without (e.g., "now you remember"). Future work will be required to examine the implications of such phonetic variation on the time course of the constancy mechanism.

Acknowledgments

The authors thank Hynek Hermansky, Simon Makin, Ray Meddis, Kalle Palomäki and Andrew Raimond for discussion, and two anonymous reviewers and the Associate Editor for helpful comments on the manuscript. We additionally thank all listeners who took part in the experiments. The study was supported by EPSRC grant EP/G009805/1.

Endnotes

1. We define $H(X;Y) = \sum_{x,y} p_{xy}\log(p_{xy}/p_x p_y)$ and $H(X) = -\sum_x p_x \log(p_x)$, where p_x is the probability of occurrence of stimulus x, p_y is the probability of occurrence of response y, and p_{xy} is the probability of the joint occurrence of x and y. Probabilities were estimated from the finite sample of observations taken during the experiment, as described by Miller and Nicely (1955).

References

ISO 3382 (**1997**). "Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters," International Organization for Standardization, Geneva.

Adelson, E. H. (**2000**). "Lightness perception and lightness illusions," in *The New Cognitive Neurosciences*, 2nd ed., edited by M. Gazzaniga, (Cambridge, MA, MIT Press), 339–351.

Allen, J. B. and Li, F. (**2009**). "Speech perception and cochlear signal processing," IEEE Signal Proc. Mag. **26** (4), 73–77.

Black, J. W. (**1950**). "The effect of room characteristics upon vocal intensity and rate," J. Acoust. Soc. Am. **22**, 174–176.

Boersma, P. and Weenink, D. (**2010**) "Praat, version 5.0.40" [Online; accessed 1 Jul 2010] `http://www.praat.org/`

Bolt, R. H. and MacDonald, A. D. (**1949**). "Theory of speech masking by reverberation," J. Acoust. Soc. Am. **21** (6), 577–580.

Brandewie, E. J. and Zahorik, P. (**2010**). "Prior listening in rooms improves speech intelligibility," J. Acoust. Soc. Am. **128** (1), 291–299.

Brandewie, E. J. and Zahorik, P. (**2012**). "Adaptation to room acoustics using the Modified Rhyme Test," Proc. Meet. Acoust. **12**, 050007.

Brandewie, E. J. and Zahorik, P. (**2013**). "Time course of a perceptual enhancement effect for noise-masked speech in reverberant environments," J. Acoust. Soc. Am. **134** (2), EL265–EL270.

Carnegie Mellon University (**2010**) "CMU pronunciation dictionary, v. 0.7a" [Online; accessed 1 Jul 2010] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

Drullman, R., Festen, J. M. and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95** (2), 1053–1064.

Egan, J. P., Greenberg, G. Z. and Schulman, A. I. (**1961**). "Interval of time uncertainty in auditory detection," J. Acoust. Soc. Am. **33** (6), 771–778.

Gelfand, S. A. and Silman, S. (**1979**). "Effects of small room reverberation upon the recognition of some consonant features," J. Acoust. Soc. Am. **66** (1), 22–29.

Green, D. M. and Forrest, T. G. (**1989**). "Temporal gaps in noise and sinusoids," J. Acoust. Soc. Am. **86** (3), 961–970.

Houtgast, T. and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77** (3), 1069–1077.

Hidden Markov Model Toolkit (**2010**) "HTK, version 3.4.1" [Online; accessed 1 Jul 2010] `http://htk.eng.cam.ac.uk`

Kirk, R. E., (**1968**) *Experimental Design: Procedures for the behavioral sciences* (Brooks/Cole Publishing Co., Belmont, CA), p. 66.

Kuwada, S., Bishop, B. and Kim, D. O. (**2012**). "Approaches to the study of neural coding of sound source location and sound envelope in real environments," Front. Neural Circuits **6**, 42.

Lee, K.-F. and Hon, H.-W. (**1989**). "Speaker-independent phone recognition using hidden Markov models," IEEE T. Acoust. Speech **37** (11), 1641–1648.

Liberman, A. M. and Delattre, P. and Cooper, F. S. (**1952**). "The role of selected stimulus-variables in the perception of the unvoiced stop consonants," A. J. Psychol. **65** (4), 497–516.

Longworth-Reed, L. and Brandewie, E. J. and Zahorik, P. (**2009**). "Time-forward speech intelligibility in time-reversed rooms," J. Acoust. Soc. Am. **125** (1), EL13–EL19.

Miller, G. and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27** (2), 338–352.

Mines, M. A., Hanson, B. F. and Shoup, J. R. (**1978**). "Frequency of occurrence of phonemes in conversational English," Lang. Speech. **21** (3), 221–241.

Moore, R., Adams, E., Dagenais, P. A. and Caffee, C. (**2007**). "Effects of reverberation and filtering on speech rate judgment," Int. J. Audiol. **46** (3), 154–160.

Nábělek, A. K., Letowski, T. and Tucker, F. (**1989**). "Reverberant overlap- and self- masking in consonant identification," J. Acoust. Soc. Am. **86** (4), 1259–1265.

Nielsen, J. B. and Dau, T. (**2010**). "Revisiting perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **128** (5), 3088-3094.

Phatak, S. A. and Lovitt, A. and Allen, J. B. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124** (2), 1220–1233.

Shinn-Cunningham, B. G. (**2000**). "Learning reverberation: Considerations for spatial auditory displays," Proc. International Conference on Auditory Display, Atlanta, GA, 126–134.

Smith, A. M. (**1990**). "On the use of the relative information transmitted (RIT) measure

for the assessment of performance in the evaluation of automated speech recognition (ASR) devices," Proc. 3rd Australasian Speech Science and Technology Association Conference, Melbourne, Australia, 368–373.

Srinivasan, N. K. and Zahorik, P. (**2011**). "The effect of semantic context on speech intelligibility in reverberant rooms," Proc. Meet. Acoust. **12**, 060001.

Srinivasan, N. K. and Zahorik, P. (**2013**). "Prior listening exposure to a reverberant room improves open-set intelligibility of high-variability sentences," J. Acoust. Soc. Am. **133** (1), EL33–EL39.

Srinivasan, N. K. and Zahorik, P. (**2014**). "Enhancement of speech intelligibility in reverberant rooms: Role of amplitude envelope and temporal fine structure," J. Acoust. Soc. Am. **135** (6), EL239–EL245.

Ueno, K., Kopčo, N. and Shinn-Cunningham, B. G. (**2005**). "Calibration of speech perception to room reverberation," Proc. Forum Acusticum, Budapest, Hungary.

Watkins, A. J. (**2005a**). "Perceptual compensation for effects of echo and of reverberation on speech identification," Acta Acust. United Acust. **91**, 892–901.

Watkins, A. J. (**2005b**). "Perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **118** (1), 249–262.

Watkins, A. J. and Raimond, A. P. (**2013**). "Perceptual compensation when isolated test words are heard in room reverberation," in *Basic Aspects of Hearing: Physiology and Perception*, edited by Moore, B. C. J., Patterson, R. D., Winter, I. M., Carlyon, R. P. and Gockel, H. E. (Springer, New York), 193–201.

Watkins, A. J. and Raimond, A. P. and Makin, S. J. (**2011**). "Temporal-envelope constancy of speech in rooms and the perceptual weighting of frequency bands," J. Acoust. Soc. Am. **130** (5), 2777–2788.

Wright, J. (**2005**). "Articulation Index (LDC2005S22)," Linguistic Data Consortium, Philadelphia

Zahorik, P. and Brandewie, E. J. and Sivonen, V. P. (**2009**). "Spatial hearing in reverberant rooms and effects of prior listening exposure," Proc. International Workshop on the

Principles and Applications of Spatial Hearing, Zao, Miyagi, Japan.

Zahorik, P., Kim, D. O. and Kuwada, S., Anderson, P. W., Brandewie, E. J., Collecchia, R., and Srinivasan, N. K. (**2012**). "Amplitude modulation detection by human listeners in reverberant sound fields: Carrier bandwidth effects and binaural versus monaural comparison," Proc. Meet. Acoust. **15**, 050002.

Zahorik, P. and Anderson, P. W. (**2013**). "Amplitude modulation detection by human listeners in reverberant sound fields: Effects of prior listening exposure", Proc. Meet. Acoust. **19**, 050139.

TABLE I. Confusion matrices summarising 60 participants' responses at three of the 4 kHz low-pass filter cutoff conditions in Experiment 1. Reverberation conditions are labelled as *context-test* distance. Rows correspond to the stimuli presented; columns record the responses. In the near-near condition, no confusions were recorded. In the near-far condition, listeners frequently misreported 'skur', 'spur' and 'stir' as 'sir'. These confusions were largely resolved in the far-far condition.

	near-near					near-far					far-far				
		SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR		SIR	SKUR	SPUR	STIR
4 kHz	SIR	60	0	0	0	SIR	56	1	0	3	SIR	52	1	0	7
	SKUR	0	60	0	0	SKUR	9	46	3	2	SKUR	2	52	0	6
	SPUR	0	0	60	0	SPUR	27	3	27	3	SPUR	4	3	47	6
	STIR	0	0	0	60	STIR	23	2	1	34	STIR	2	0	0	58

List of Figures

FIG. 1 Illustration of same- and mixed-distance reverberation conditions for one representative example of the 80 phrases used in Experiment 1. The traces are amplitudes (Amp.) of low-pass filtered (cutoff frequency 80 Hz) Hilbert envelopes derived from the temporally aligned context *(light line, upper label)* and test-word *(dark line, lower label)* before these two sounds were added, point-wise, to form the experimental stimuli. Before the addition, the context and test-word were independently reverberated at 'near' or 'far' room distances to give, from top to bottom: near-near, near-far, far-near and far-far *context-test* distance conditions. In the top panel, the test-word is annotated with pointers to show, from left to right, the start of frication, closure and voicing. 9

FIG. 2 Mean and standard error of the 60 participants' 1−RIT scores at the five low-pass filter conditions of Experiment 1. Compensation for reverberation is apparent in the downward-sloping upper line of the 3 and 4 kHz filter conditions. In these two conditions, an increased level of reverberation in the context (resulting from the increase in context distance) brings about an improvement in the identification of the far-distance test-words. 14

FIG. 3 Data of Figure 2 replotted to show the effect of lowpass filtering on each *context-test* condition. Consonant identification error decreased monotonically with increasing lowpass cutoff frequency, except when the context was 'near' reverberated and the test-word was 'far' reverberated. 16

FIG. 4 Data of the near-far condition in Figure 3 replotted to show the cutoff filter effect on each test-word's responses. All test-words showed a similar pattern of performance, with a 'pivot point' at 1.5 kHz for 'spur' and at 2 kHz for the remainder. 18

FIG. 5 Illustration of selected stimulus conditions in Experiment 2, here applied to the same phrase that was previously displayed in Figure 1. Context phrase *(light line, upper label)* and test-word *(dark line, lower label)* were independently reverberated at 'near' or 'far' distances as before *(upper half)*. In other conditions the context was silenced *(lower half)* and the test-word was gated *(bottom panel)*. Other details are the same as Figure 1. 20

FIG. 6 Mean and standard error of the 60 participants' 1−RIT scores in Experiment 2. Conditions in which the reverberation tail following the test-word was removed by gating are shown with white markers. Conditions that preserve intrinsic information are presented with black markers. Extrinsic compensation seems to effect an overall reduction in consonant confusions between near-far and far-far *context-test* conditions, which replicates the corresponding main effect in Experiment 1. This reduction is seen for both gate conditions. 24

FIG. 7 *Left* – Mean and standard error of the 60 participants' difference scores (RIT error at far distance test-word minus RIT error at near distance test-word) for near (n), silent (s) and far (f) contexts at each gate condition in Experiment 2. *Right* – Pooled (p) results show some effect of gating on the test-word's vowel, indicating that intrinsic factors have probably helped to disambiguate far distance test-words in the near- and silent-context conditions. Note that the ordinate scales of the two panels differ. 26

FIG. 8 Illustration of selected stimulus conditions for an Experiment 3 phrase. The test-word *(dark line, lower label)* is preceded by the context *(light line)* which is divided into an initial near-reverberated part and a subsequent far-reverberated part. The temporal position of the boundary is varied between these parts, so that less or more of the context immediately prior to the test-word is far-reverberated *(shaded area, upper label)*. The phrase differs from that used in Figure 1, but other details are the same. 29

FIG. 9 Mean and standard error of the 40 participants' scores in Experiment 3. The lower line shows near-distance test-word scores, where the proportion of initial 's' reports is close to 0.5 and the data show no dependency on the duration of the far-reverberated context. The upper line shows the far-distance test-word scores. For the zero-length far-context window condition, where the test-word has more reverberation than the context, listeners often misclassified [st] as 's'. As the duration of the far-reverberated part of the context is increased, fewer misclassifications are made and the proportion of 's' reports decreases. 32