Temporal-envelope constancy of speech in rooms and the perceptual weighting of frequency bands

Anthony J. Watkins,[a] Andrew P. Raimond, and Simon J. Makin
Department of Psychology, The University of Reading, Reading RG6 6AL, United Kingdom

(Received 7 September 2010; revised 29 July 2011; accepted 23 August 2011)

Three experiments measured constancy in speech perception, using natural-speech messages or noise-band vocoder versions of them. The eight vocoder-bands had equally log-spaced center-frequencies and the shapes of corresponding "auditory" filters. Consequently, the bands had the temporal envelopes that arise in these auditory filters when the speech is played. The "sir" or "stir" test-words were distinguished by degrees of amplitude modulation, and played in the context; "next you'll get _ to click on." Listeners identified test-words appropriately, even in the vocoder conditions where the speech had a "noise-like" quality. Constancy was assessed by comparing the identification of test-words with low or high levels of room reflections across conditions where the context had either a low or a high level of reflections. Constancy was obtained with both the natural and the vocoded speech, indicating that the effect arises through temporal-envelope processing. Two further experiments assessed perceptual weighting of the different bands, both in the test word and in the context. The resulting weighting functions both increase monotonically with frequency, following the spectral characteristics of the test-word's [s]. It is suggested that these two weighting functions are similar because they both come about through the perceptual grouping of the test-word's bands. © *2011 Acoustical Society of America*. [DOI: 10.1121/1.3641399]

PACS number(s): 43.55.Hy, 43.71.Gv, 43.71.Es, 43.71.An [RYL] Pages: 2777–2788

I. INTRODUCTION

When a speech message is played at different distances from a listener in a room, the different amounts of reflected sound from the room's surfaces make the temporal envelopes of the signals very different. Nevertheless, these sounds are generally heard to have very similar phonetic content at diverse distances, suggesting that there is a "perceptual constancy" operation in hearing (Watkins and Makin, 2007b). This constancy would appear to arise from a mechanism that takes account of the amount of reflected sound in the surrounding context, and as a consequence, the level of the context's reflections is associated with compensation for effects of reflections in adjacent test-words (Watkins, 2005a,b).

Figure 1 shows waveforms of the "sir" and "stir" test-words that were first used in the experiments of Watkins (2005a). The amplitude modulation (AM) function is chosen so that a sound with the temporal envelope of a "stir" token is obtained when the function modulates the waveform of the "sir" token. Both of these sounds are easy for listeners to identify appropriately. A prominent effect of the modulation is the gap that precedes voicing in "stir," and gaps of different depths are generated using attenuated versions of the AM function. These test words are inserted in a suitable speech-context, such as "next you'll get _ to click on," and presented to listeners. When the level of room reflections in the test word is increased, the characteristic gap in [st] tends to be filled by the "tail" that is added to the end of the [s], and listeners identify more test words as "sir." However, when the reflections' level in the context is also increased, there appears to be a perceptual compensation that arises from the context, so that fewer test words are now identified as "sir."

Subsequent experiments (Watkins and Makin, 2007a,b) used steady-spectrum noise-contexts with sharply varying temporal envelopes, such as those that arise in auditory filters when speech is heard. With this type of context, compensation effects can also be substantial, even though these sounds are unintelligible noises (Watkins and Makin, 2007b). These and other results support the idea that phonetic perception is not responsible for the compensation effect and that rather, a more general perceptual-constancy type of mechanism is involved.

In one initial experiment, Watkins (2005a; experiment 1) found that when test words had the reflection pattern of a distant source, compensation effects increased with the context's level of reflections. This raises the question of how the level of reflections in the context is indicated in perceptual constancy. There are acoustic indices of speech intelligibility that are based on analyses of the temporal envelopes arising in different frequency-bands, giving measures that include the Speech Transmission Index (STI) and its relatives (Houtgast and Steeneken, 1973). This type of analysis assesses the degree of modulation in temporal envelopes at modulation rates that are significant in speech in order to obtain the relative attenuation of these modulation-frequency components. This attenuation increases with factors that adversely affect intelligibility, which include the amount of reverberation present in the sounds. However, it seems unlikely that constancy is cued by this particular aspect of the signal, as constancy is not obtained with reversed reverberation (Watkins, 2005), where the modulation attenuation is much the same as it is with forwards reverberation (Longworth-Reed *et al.*, 2009). An alternative suggestion is that constancy effects are

[a]Author to whom correspondence should be addressed. Electronic mail: syswatkn@reading.ac.uk

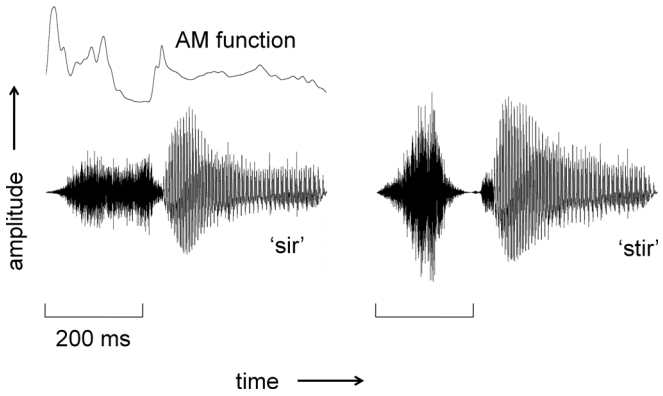

FIG. 1. The waveform on the left is the "sir" [s3] token used to make the other test words through amplitude modulation with the AM function shown above it. When this function is not attenuated the "stir" [st3] waveform on the right is generated. These two sounds are the end points of an 11-step continuum of test words whose intermediate steps are formed with suitably attenuated versions of the AM function.

associated with a different aspect of the temporal envelope, which is the prominence of the tails that reverberation adds at the ends of sounds (Stecker and Hafter, 2000) and at spectral transitions in auditory filters (Watkins, 2005a,b). This still implies that constancy operates through processing the temporal envelope, so according to this view, it should be independent of the fine-grained temporal structure of the signal.

There are, however, potential indicators of the amount of reflected sound in the "fine-structure" of speech signals. This structure is brought about by the signal's periodicity at voice-pitch rates (Drullman, 1995; Smith *et al.*, 2002), as well as by time delays between direct and reflected sounds when these sounds are played in rooms (Bilsen 1967/68; Bilsen and Ritsma, 1967/68). These two factors interact in perception, which can have deleterious effects, especially when fine-structure based cues to the perceptual segregation of sound sources are impaired by the presence of room reflections (Culling *et al.*, 1994, Culling *et al.*, 2003). Nevertheless, the degree of disruption of the signal's fine structure might be helpful to listeners as an indication of the amount of reflected sound that is present. Such a cue might be the strength of the signal's "harmonicity" (Roman and Wang, 2006), which reduces as the reflections' level increases.

To ask whether constancy is independent of the fine-grained temporal structure of the sound source, speech from a noise-band vocoder is used in the present experiments. This type of vocoder preserves the sound's narrow-band temporal-envelopes, but its fine structure is scrambled (e.g., Apoux and Bacon, 2004). Nevertheless, when the noise bands' center-frequencies span the speech range, the outputs obtained by adding the processed bands together are heard to be distinctly speech-like, and although these sounds have an unnatural "noise-like" quality, the original messages are quite intelligible (Shannon *et al.*, 1995). Such results indicate the wealth of phonetic information obtainable from these narrow-band temporal-envelopes. Moreover, these envelopes also seem to be the source of the phonetic information obtained by users of cochlear implants, as these devices work by conveying a selection of narrow-band temporal envelopes to their users (e.g., Dorman and Loizou, 1998; Friesen *et al.*, 2001). Experiment 1 asks whether constancy in temporal-envelope processing is also apparent with this type of speech.

Further conditions assess the contribution of effects from the fine structure of reflection patterns. Some of these effects remain when reflections are applied (by convolution) to speech from the noise vocoder, and they include the characteristic'comb-filter' shaping of the spectrum in sounds containing reflections (Bilsen 1967/68; Bilsen and Ritsma, 1967/68). In the present experiments, scrambling of the speech source's fine-structure is effected through a "signal-correlated-noise" (SCN) operation (Schroeder, 1968), wherein the polarity of a randomly selected half of the signal's samples is reversed. However, it is also possible to reposition this signal-processing stage so that it comes after the room-reflections have been introduced. When this is done, both the temporal fine-structure of the speech and the reflection-pattern are scrambled in the resulting sounds. Different conditions in experiment 1 compare the effects of these two types of scrambling to ask whether the fine structure of reflection patterns is important for compensation.

Information about processes underlying perception of the type of vocoder speech used in the present experiments can be obtained from measurements of the "perceptual weightings" that listeners attach to the individual bands (e.g., Apoux and Healy, 2010). Such weightings reflect the relative contribution of the bands to the perceptual effects in question. There are two effects in the present experiments; one is the shift in category boundary when the reflections' level in test words is increased, and the other is the compensatory reduction in this shift when the reflections' level in the context is also increased. If the processes underlying these effects are related, then the pattern of perceptual weightings obtained for the test word's bands should resemble the pattern for the context's bands.

For the test words, it seems likely that perceptual weightings of the bands will reflect the frequency characteristics of these sounds' initial frication, as listeners might be expected to place more weight in the region of the spectrum where there is more of a difference between the [s] and the [st] at the start of the words. The level of this frication can be seen in Figs. 2 and 3, which indicate that it is relatively low at low frequencies. However, its level becomes more substantial towards higher frequencies, where there is a corresponding increase in the difference between the [s] and [st] versions of the words. Experiment 2 asks whether the perceptual weightings of bands in test words have a correspondingly "high-pass" characteristic.

On the other hand, for bands in the context, it is possible that there are different factors influencing the frequency-pattern of perceptual weights, especially if compensation is effected through a gathering of information about the reflections' level across frequencies. Such a process might resemble calculation of "wide-band" versions of room-acoustic indices, such as the early-to-late index, C_{50}, or reverberation time, T_{60} (ISO 3382, 1997). For real-room reflection-patterns, there tends to be more reflected sound at lower frequencies. A reason for this variation is that typical room-surfaces generally have lower acoustic absorption at low frequencies where the reflections' level will tend to be larger than it is at higher

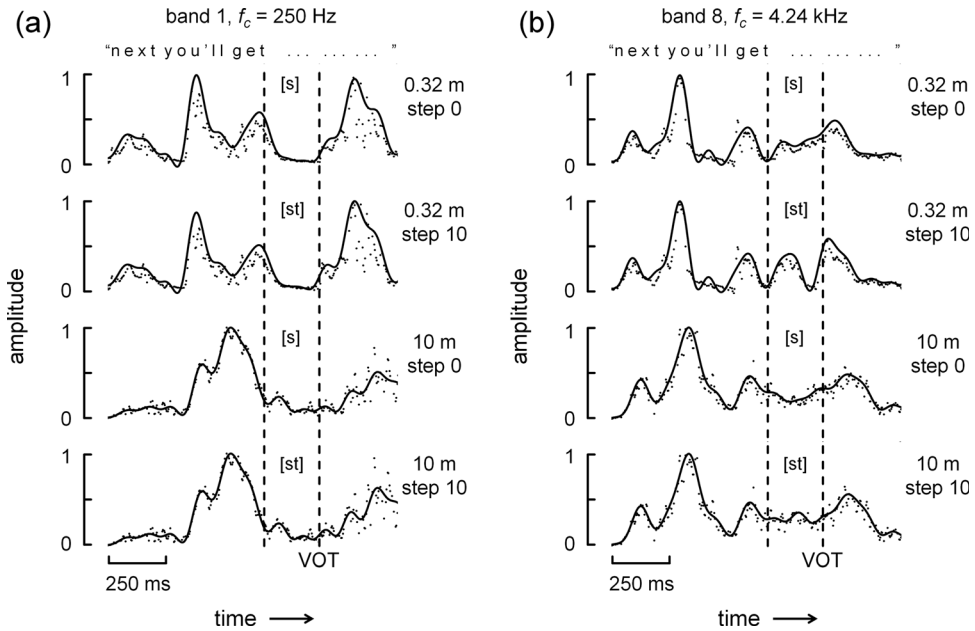

FIG. 2. The traces are temporal envelopes in narrow frequency-bands of test words from the continuum's "sir" [s3] and "stir" [st3] end-points along with the preceding part of the context, "next you'll get…." The speech was first convolved with the left channel of a binaural room impulse response (BRIR) obtained with a source to receiver distance of 0.32 m (upper two rows) or 10 m (lower two rows). These sounds were then played through an auditory filter with a center frequency, $f_c = 250$ Hz (a) or $f_c = 4.24$ kHz (b), followed by full-wave rectification and then low-pass filtering with corner frequencies of 50 Hz (points) or 10 Hz (continuous traces). The interval occupied by the test-word's [s] or [st] frication lies between the vertical dashed lines, which are the start of the test word (left-hand line) and voice-onset time (VOT) in the original dry recording.

frequencies. Figure 4 shows that these frequency characteristics are found with the real-room reflections used in the present experiments. As there is more compensation with contexts that have higher levels of reflections (Watkins, 2005a), it may be that compensation effects in this real room will be more substantial from the lower-frequency bands of the context, resulting in a "low-pass" pattern of perceptual weightings. To test this idea, experiment 3 uses this real-room's reflection patterns to assess perceptual weightings of the context's bands from their compensation effects, and it asks whether the pattern is similar to patterns for the test-word's bands measured in experiment 2. For this purpose, comparable measurement scales were used for the perceptual weightings in the two experiments.

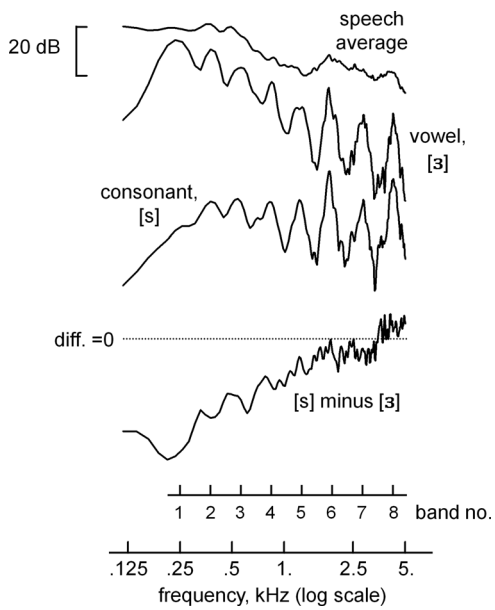

FIG. 3. The uppermost trace is the average spectrum of the context, "next you'll get" and "to click on." This was calculated by abutting the two parts and then averaging the FFTs of 170.7-ms Hann-windowed segments spaced at 21.3-ms intervals. The second-highest trace is the smoothed FFT of a 42.7-ms Hann-windowed segment centered on the test word's [ɜ] vowel, and the trace below this is a similar segment centered on the test word's [s]-frication (step 0 of the test-word continuum was used for this analysis). For clarity, the spacing between the fricative and vowel traces is made slightly larger than their actual difference. This difference is shown accurately in the lowest trace, where dB values from the vowel's FFT are subtracted from corresponding values in the fricative's FFT. The dotted horizontal line indicates a value of zero dB (diff. = 0) for this trace. The vowel's frequency-bands are relatively more powerful than those of the fricative at lower frequencies, but this difference diminishes towards the higher-frequency bands where it eventually reverses, so that the vowel's uppermost band is somewhat less powerful than the corresponding band in the fricative.

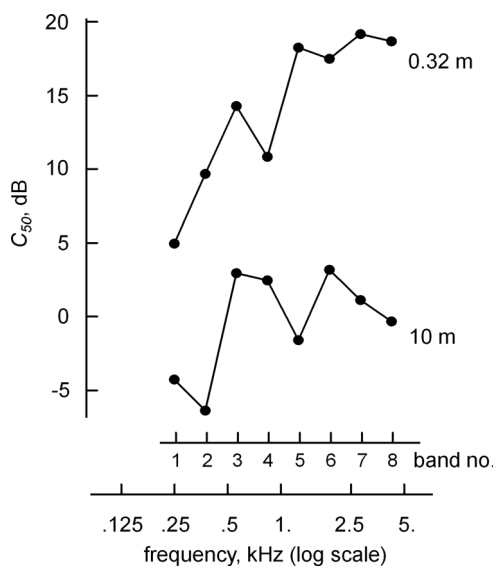

FIG. 4. The amount of reflected sound in different frequency-bands measured by the early-to-late index, C_{50}, which is the ratio of energy in the first 50 ms of the filtered impulse-response to the energy of the remainder (ISO 3382,1997). Values are shown for the center frequencies of the auditory-filter shaped bands used in the present experiments' vocoded speech. C_{50} tends to increase with frequency, consistent with a corresponding increase in the absorption of the room's surfaces. However, the low-pass characteristic of the head's shadow and its effect on contralateral reflected-sound also seems to contribute as there is a more pronounced increase in C_{50} with frequency at the nearer distance (0.32 m).

II. EXPERIMENT 1

A. Speech contexts and the test-word continuum

The methods described by Watkins (2005a) were used to obtain context phrases containing test-words from a continuum between "sir" and "stir," illustrated in Fig. 1. This method used the speech of one of the authors (AW) recorded with 16-bit resolution at a 48-kHz sampling rate using a Sennheiser MKH 40 P48 cardioid microphone in an IAC 1201 double-walled booth, giving "dry" speech. The context phrase was originally such a recording, of the phrase "next you'll get sir to click on," with the "sir" test-word excised using a waveform editor. A recording of a "stir" test-word was also obtained in this context phrase. The durations of the context's first and second parts were both 685 ms, and the original recordings of the test words were both 577 ms long.

To form a test-word continuum, the wide-band temporal envelopes of "sir" and "stir" were obtained by full-wave rectification followed by a low-pass filter with a 50-Hz corner frequency. The envelope of "stir" was then divided (point-wise) by the envelope of "sir" to give a modulation function, and clear "stir" sounds were obtained by amplitude modulating the waveform of "sir" with this function. The original "sir" along with the "stir" produced by the modulation are shown in Fig. 1. These two sounds were the 11-step continuum's end-points; nominally steps 0 and 10, respectively. The intermediate steps were produced from the recording of "sir" using appropriately attenuated versions of the modulation function.

Test words were re-embedded into the context parts of the original utterance. This re-embedding was performed by adding the context's waveform to the test word's waveform. Before the addition, silent sections were added to preserve temporal alignment, and to allow different reflection-patterns to be separately introduced into the test word and the context.

B. Category boundaries

When room reflections oppose the amplitude modulation that formed the continuum, cues to the presence of a [t] in test words are made less prominent. In Fig. 2(b), temporal envelopes in a high-frequency region show this effect, while the corresponding perceptual consequence will be that more of the continuum's steps will attract "sir" responses from listeners. To assess this perceptual effect of reflections, and to indicate differences among conditions, listeners' category boundaries were compared. The boundary is the step, or point between steps, where listeners switch from predominantly "sir" to predominantly "stir" responses.

Listeners were asked to identify four presentations of each of the continuum's steps played in the context, and category boundaries here were found from the total number of "sir" responses across all 11 steps. This total was divided by 4 before subtracting 0.5, to give a boundary step-number between -0.5 and 10.5.

C. Room reflections

The methods described by Watkins (2005a) were also used to introduce room reflections into the dry contexts and test words by convolution with the left-ear channel of a binaural room impulse-response (BRIR). This gives the effect of monaural real-room listening over headphones. The BRIRs were obtained in rooms using dummy-head transducers (a speaker in a Bruel and Kjaer 4128 head and torso simulator, and Bruel and Kjaer 4134 microphones in the ears of a KEMAR mannequin), so that they incorporate the directional characteristics of a human talker and a human listener. To obtain signals at the listener's eardrum that match the signal at KEMAR's ear, the frequency-response characteristics of the dummy-head talker and of the listener's headphones were removed using appropriate inverse filters.

The BRIRs were obtained in a disused office that was L-shaped with a volume of 183.6 m^3. The transducers faced each other, while the talker's position was varied to give distances from the listener of 0.32 or 10 m. This gave different levels of reflected sound, as indicated by the ratio of early (first 50 ms) to late energy in the impulse response (C_{50}, ISO 3382, 1997). The wideband (A-weighted) value of C_{50} in the room was found to be 8 dB at 0.32 m, descending to 2 dB at 10 m.

D. Eight-band vocoder

The individual vocoder-bands were narrowband noises, each with the temporal-envelope fluctuations that arise in an auditory filter when speech is played. The impulse response of a filter was a "gammatone" function with the parameter $\eta = 4$ and with the bandwidth appropriate for its center frequency, as given by the "Cambridge ERB" (Glasberg and Moore, 1990). There are two reasons for this departure from the broader flatter bands that are more usual in noise vocoders. One is to reduce interactions between adjacent bands, especially for conditions (in experiments 2 and 3) where the BRIR's distance is varied from band to band. The other reason is that these filters are convenient in giving an approximate indication of how the stimuli would be represented at the earliest stages of hearing (e.g., Fig. 2).

The eight center-frequencies were equally log-spaced across the speech range, starting at 250 Hz, and increasing by intervals of a musical fifth (7/12 octave) to 4.24 kHz. Bands were numbered from low to high center-frequency, using a band number, $n = 1,2,\ldots,8$. Figure 2 shows the temporal envelopes of the experiments' stimuli in two of these bands, while Fig. 3 shows the corrugation that this processing lends to these sounds' spectral characteristics.

The amount of reflected sound varies across these bands. This is partly due to variation of room-surfaces' absorption characteristics across frequency, but there is also a low-pass effect due to the head's shadowing of contra-laterally incident reflections when the source is nearby. Both these factors generally tend to give a decrease in the amount of reflected sound as frequency is increased, and the narrowband C_{50} values plotted in Fig. 4 show both these effects in the room measured here.

1. SCN-before

In these conditions, the fine-structure of the speech source is scrambled, but due to the positioning of the SCN

operation, effects from the fine-structure of the room-reflections' pattern are preserved. To obtain a vocoder noise-band in this condition, the speech was first played through an auditory filter, followed by the "signal correlated noise" (SCN) operation, where the polarity of a randomly selected half of the signal's samples is reversed. This operation gives a wideband signal, but it preserves the temporal envelope of the filter's output. The signal was then band limited by playing it through a version of the auditory filter that had its impulse response reversed, thereby correcting for delays introduced by the operation of the first filter.

Room-reflection patterns were applied to each band by convolution, and their rms levels were equated before they were all added together. Then, to give the resulting signal the long-term spectral characteristics of the original speech, the bands' levels were adjusted by playing their sum together through a "speech-shaping" filter whose frequency response is the long-term average spectrum of the original speech-context, as shown in Fig. 3.

2. SCN-after

In these conditions, the fine-structure of both the speech source and the room-reflections' pattern are scrambled, by a repositioning of the SCN operation. To obtain a vocoder noise-band in these conditions, the speech was played through an auditory filter as before, but now, a room-reflection pattern is introduced at this point, by convolution with the appropriate BRIR. These operations are then followed by the SCN operation so that it scrambles both of their fine-structure effects. The subsequent stages remain as before; i.e., band limiting, summing the bands, and speech shaping the result. As independent noise generators are used for each implementation of the SCN operation, the effect here is to vary the fine structure of the room-reflection pattern across frequency bands, as well as between the context and the test-word.

E. Procedure

The six listeners were undergraduate volunteers from Reading's Psychology research panel who reported no hearing problems. They each identified test-word continua in an unprocessed-speech condition as well as in the two SCN conditions, before and after. In each of these three speech-type conditions, the reflection pattern's distance was varied between 0.32 and 10 m in both the context and the test word to give all four combinations. For each listener there were 3 speech types $\times$ 2 test-word distances $\times$ 2 context distances $\times$ 11 continuum steps $\times$ 4 repeats $=$ 528 trials. Each listener received the trials in a different randomized order during single, unbroken listening-sessions that were administered individually.

Before these experimental trials, listeners were informally given a few randomly selected practice trials to familiarize them with the sounds and the set up, and to check that they could hear the eight-band sounds as speech. This amount of practice proved sufficient to leave listeners confident about identifying the vocoder speech, and the task did not seem demanding of their concentration. Factors likely to contribute to this task's apparent ease for listeners include the limited stimulus set, the regular occurrence of natural-speech trials, and the fact that there were only two response alternatives.

During the listening-sessions, sounds were presented at a peak level of 48 dB SPL through the left earpiece of a Sennheiser HD480 headset in the otherwise quiet conditions of the IAC booth. Trials were administered by an Athlon 3500 PC computer with MATLAB 7.1 software and with an M-Audio Firewire 410 sound card. On each of these trials, a context with an embedded test-word was presented. Listeners then identified the test word with a click of the computer's mouse, which they positioned while looking through the booth's window at the "sir" and "stir" alternatives displayed on the computer's screen. This click also initiated the subsequent trial.

F. Results

Means and standard errors of category boundaries are shown in Fig. 5.

Differences among these means were tested with an analysis of variance that had three within-subject factors; test distance, context distance (both with two levels), and speech type (with three levels). There is a substantial two-way interaction between the test- and context-distance factors, which seems to arise from perceptual compensation for effects of room reflections; i.e., the effect of increasing test distance is markedly reduced when the context's distance is increased ($F_{(1,5)} = 110.93$, $p < 0.0001$). This compensation effect appears substantial in each of the speech-type conditions.

The three-way interaction that tests whether compensation varied across the speech-type conditions was not found to be significant, although there are two-way interactions

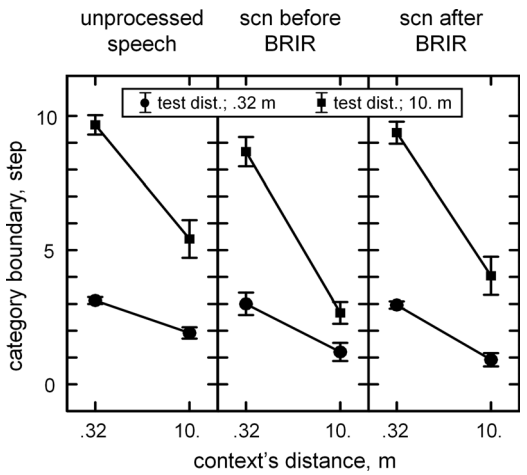

FIG. 5. The graphs show means and standard errors (bars) of category boundaries from experiment 1, where six listeners identified members of the test-word continuum in the context. The left-side panel shows results with unprocessed speech while the other two panels show results with noise-vocoded versions where the fine structure of this speech has been scrambled. The rightmost panel shows results in the "sen-after BRIR" conditions where the fine structure of the room's impulse response has also been scrambled. Neither of these manipulations appears to have reduced the compensation effect, which is substantial in all three panels, as indicated by the reductions in the category boundary in 10-m test words when the context's distance is increased to 10 m.

between speech type and context's distance ($F_{(2,10)} = 6.74$, Huynh–Feldt epsilon = 0.9562, $p < 0.0156$) and between speech type and test distance ($F_{(2,10)} = 5.47$, Huynh–Feldt epsilon > 1, $p < 0.0248$). Figure 5 indicates that for both of these interactions, it is the scn-before condition that seems different from the other two speech-types. Although this might suggest some small influence from the reflection-pattern's fine-structure, this influence does not seem to be apparent with the unprocessed speech, and so might instead be an artifact of vocoder-processing.

There are significant main effects for all three factors, which seem to come about through the pattern of the higher-order interactions described above. The effect of speech type is less marked ($F_{(2,10)} = 7.07$, Huynh–Feldt epsilon = 0.8256, $p < 0.0191$) than the effects of test distance ($F_{(2,10)} = 200.84$, $p < 0.0001$) and context's distance ($F_{(1,5)} = 200.84$, $p < 0.0001$).

Overall, the result-pattern is clear and consistent across the unprocessed and processed conditions. All three conditions show substantial increases in category boundaries when the distance of test-words' BRIRs is increased, while these effects of the test-word's distance are considerably reduced when the context's distance is also increased. The reliability of effects across conditions is clear from Fig. 5, consistent with listeners' reports that the vocoder's speech was about as easy to identify as the natural speech.

G. Discussion

Results with the unprocessed speech indicate a perceptual constancy effect that replicates earlier findings with these sounds. When the level of reflections in test words is increased by increasing their distance, more "sir" responses are made to them, and category boundaries increase. However, when the level of reflections in the context is increased as well, this seems to be taken into account in perception, giving rise to a constancy effect. As a result, the number of "sir" responses reduces, and category boundaries return to a level closer to that found in conditions with lower levels of reverberation.

A very similar result-pattern is found in scn-before conditions, indicating that these sounds can bring about the same sort of constancy effects. From these results it would seem that certain fine structure indicators of the reflections' level, such as the strength of "harmonicity" in the speech, are not necessary to cue these constancy effects. This finding is consistent with results of experiments that have used steady-spectrum noise carriers (Watkins and Makin, 2007a,b), and extends those results to sounds that have time-varying spectra.

The sen-after conditions also show substantial effects of increasing the level of reflections in test words, as well as substantial compensation effects in conditions where the level of the context's reflections are also increased. So constancy remains when the reflection-pattern's fine-structure is scrambled along with the signal's fine structure. This would appear to rule out the possibility that the reflections' level is indicated to listeners by aspects of the fine structure of reflection patterns. Constancy seems to be cued through a processing of the temporal envelope, in a way that is independent of the fine-grained temporal structure of the signal. These results are therefore consistent with a perceptual mechanism that assesses the level of reflections from the prominence of tails in the temporal envelope (Stecker and Hafter, 2000; Watkins, 2005a,b).

III. EXPERIMENT 2: PERCEPTUAL WEIGHTING OF BANDS IN THE TEST WORD

A. Rationale

Figure 2 shows temporal envelopes of the experimental sounds in a low- [Fig. 2(a)] and a high- [Fig. 2(b)] frequency band. When the source is nearby, at 0.32 m, the distinction between "sir" and "stir" that listeners are trying to make in their identification task is associated with differences between these envelopes in the temporal interval occupied by the frication of the [s] or [st] parts of the words. This "fricative interval" lies between the two vertical dashed lines; one at the end of the context and the other at the onset of voicing. In the high-frequency band in Fig. 2(b) there is a prominent gap in the fricative's interval of the temporal envelope of "stir," distinguishing it from the temporal envelope of "sir" in this interval. This distinction is much less apparent in the low-frequency band of Fig. 2(a), where there is relatively little power in the fricative's interval of either of the test-words. Hence, for the "sir"-"stir" distinction, bands at the higher frequencies are likely to be more heavily weighted, as this is the frequency region where there are bigger differences between the two words' temporal envelopes.

The variation of power in the fricative's interval across frequency is indicated in the spectral-envelope plots shown in Fig. 3. To assess the power in this interval relative to other parts of the signal, the differences between the consonant and vowel envelopes are also plotted. The resulting trace shows a generally monotonic increase with frequency. This experiment asks whether there is a corresponding increase in perceptual weightings of the test word's bands as their center frequency increases.

A problem with assessing the perceptual weights of individual bands is that variation of the reflections' distance in a single-band context does not have a substantial effect on these test words. (Watkins and Makin, 2007a,b). To get around this, the method used here compares effects between subsets of bands, and asks which subsets give the largest effects. Each subset contained half (4) of the test-word bands, and the BRIR's distance was varied (between 0.32 and 10 m) in these bands, while it was fixed (at 0.32 m) in the remainder. Each subset was paired with a subset that contained the other half of the test-word's bands, so that bands in one subset of a pair have a higher set of center-frequencies than the bands in the other subset. If perceptual weightings have a "high-pass" characteristic, then effects with a higher-frequency subset will be larger than with the corresponding lower-frequency subset.

Three subset-pair patterns were chosen so that results with them could be pooled to give a symmetrical approximation to the perceptual weighting function. Results of this pooling indicate the degree of association between variation

of the BRIR's distance in a particular band, and the attendant variation of category-boundary differences across the subsets. The assumption is that this association will be stronger for bands that are more heavily weighted by listeners. The subset-pair patterns act similarly to a basis-function set, in that this pooling will capture certain spectral shapes, including the smoothly varying high- or low-pass patterns of interest here. Sharply varying and highly asymmetrical spectral characteristics will be poorly represented however.

B. Method

Listeners identified test words in contexts whose BRIRs were held at 0.32 m throughout, while the test-word's "nominal" distance was varied between 0.32 and 10 m. The term "nominal" distance is used here because the BRIRs on half of the test-word bands were actually fixed at 0.32 m at both distances, while the distance of BRIRs on the remaining bands were at the nominal distance. In different conditions, partitioning of test-word bands between the two types followed one of the patterns shown in Fig. 6. In this figure, six conditions, (cond = 1,2,…,6), have different patterns of coefficients ($V_{n,\mathrm{cond}}$) that indicate whether the BRIR on a test-word band (n) should be held at 0.32 m ($V_{n,\mathrm{cond}} = -1$), or varied with the condition's nominal distance between 0.32 and 10 m ($V_{n,\mathrm{cond}} = +1$).

The experiment measured the perceptual weighting in each condition from the difference between category boundaries for test words at the two nominal distances. These "shifts," S_{cond}, were found for each of the test-word's bands, and an estimate was made of their association with variation in the reflection's level (i.e., with, $V_{n,\mathrm{cond}}$). This was done by multiplying the values of the two variables for a condition before summing across conditions, so that for bands in the test-word:

$$\text{perceptual weighting of band } n = \sum_{\mathrm{cond}=1}^{\mathrm{cond}=6} S_{\mathrm{cond}} V_{n,\mathrm{cond}}. \quad (1)$$

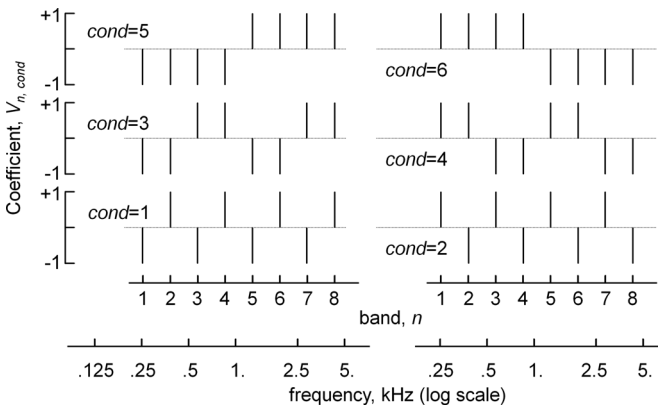

FIG. 6. Coefficients ($V_{n,\mathrm{cond}}$) used with the vocoder's eight bands (n = 1,2,…,8) to indicate whether the band's distance is to be held at 0.32 m ($V_{n,\mathrm{cond}} = -1$) or whether the band's distance is to be varied with the condition's nominal distance from 0.32 m to 10 m ($V_{n,\mathrm{cond}} = +1$). Patterns are chosen for conditions (cond = 1,2,…,6) where the perceptual effects (category-boundary shifts) can each be multiplied by the corresponding value of $V_{n,\mathrm{cond}}$ and then summed to indicate degrees of association between these variables. This pooling gives estimates of bands' perceptual weightings.

The resulting pattern of perceptual weightings across bands gives an indication of the extent to which it shows a symmetrical high-pass or a low-pass characteristic. Perceptual weightings therefore measure category-boundary shifts, and intervals between their values are weighted continuum-step differences.

Over an experimental session, listeners heard eight-band speech with 2 test-word distances × 6 cond × 11 continuum steps × 3 repeats = 396 trials. Responses on these trials were used to calculate category boundaries, with the total number of "sir" responses across all 11 steps being divided by 3 before subtracting 0.5 to give boundary step-numbers between −0.5 and 10.5. In addition, these trials were randomly interspersed with 'filler' trials where listeners heard the unprocessed versions used in experiment 1. The aim here was to maintain listeners' "speech-like" perceptions of the eight-band speech throughout the experiment. These fillers comprised 2 test-word distances × 11 continuum steps × 6 repeats = 132 trials.

Different randomized orderings of all 528 trials were used for each listener, so that on average, they heard unprocessed speech versions on every fourth trial. Other aspects of the method were the same as for experiment 1.

C. Results

Figure 7 shows means and standard errors of category boundaries from the six listeners. Differences among these means were tested with a two-way analysis of variance that had within-subject factors for test-word distance (two levels) and for condition (six levels). The two-way interaction between these factors is significant ($F_{(5,25)} = 44.4$, Huynh–Feldt epsilon = 0.8256, $p < 0.0001$) as are both main effects; test-word's distance ($F_{(1,5)} = 158.44$, $p < 0.0001$) and condition ($F_{(5,25)} = 45.64$, Huynh–Feldt epsilon = 0.3295, $p < 0.0001$). This result pattern shows substantial effects of increasing the level of reflections in the test-word's bands, but there is considerable variation among conditions that have reflection-level variations in different subsets of bands.

Figure 8 shows the perceptual weighting function obtained with Eq. (1) from the means and standard errors of listeners' category-boundary shifts. There is a clear monotonic increase with the center-frequency of the test-word's band.

D. Discussion

The high-pass pattern of bands' perceptual weights is also clear from Fig. 7, which indicates that for each of the three types of pair, the higher-frequency member gives a bigger category-boundary shift than the corresponding lower-frequency member. The largest shifts are shown in condition 5, where the BRIR's distance in the upper four bands is varied. Effects seem mostly to arise from the frequency region of these four bands, as the effects from the opposite pair-member, condition 6, are insubstantial. Effects in conditions 1 and 3 are both smaller than in condition 5, and there is some contribution from the opposite pair-members in both cases, indicating contributions to effects from bands outside these four-band subsets.

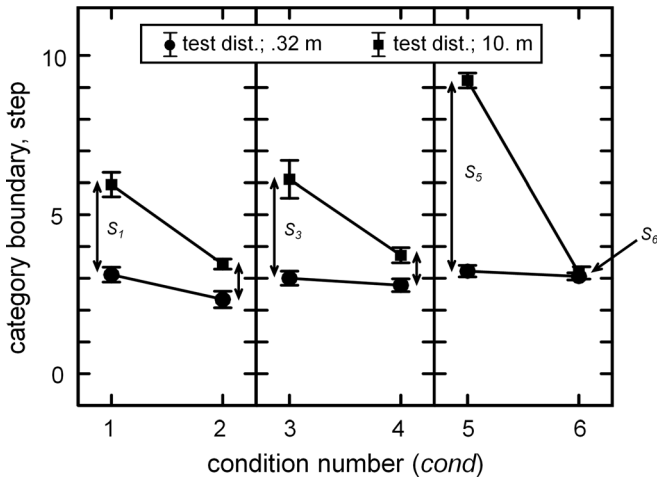

FIG. 7. The graphs show means and standard errors (bars) of category boundaries from experiment 2, where 6 listeners identified test words whose nominal distance varied between 0.32 and 10 m in contexts whose distance was held at 0.32 m throughout. In different conditions, some of the test-word's bands were also held at 0.32 m, as indicated by the coefficient-patterns associated with these condition numbers (cond) in Fig. 6. Vertical arrows show the shift in the category boundary (S_{cond}) effected as the test-word's nominal distance is increased from 0.32 to 10 m.

This result-pattern is fairly well captured in the pooled perceptual-weighting function shown in Fig. 8, although its symmetry is a straightforward consequence of the choice of symmetrical coefficient-patterns. Nevertheless, given the size of the error bars and the separations of the means, the monotonically increasing nature of the pattern is clear. This is in agreement with prediction; i.e., for distinguishing these "sir"-"stir" test words, the perceptual weighting will increase towards higher frequency bands, as this is where there are bigger differences between the two words' temporal envelopes.

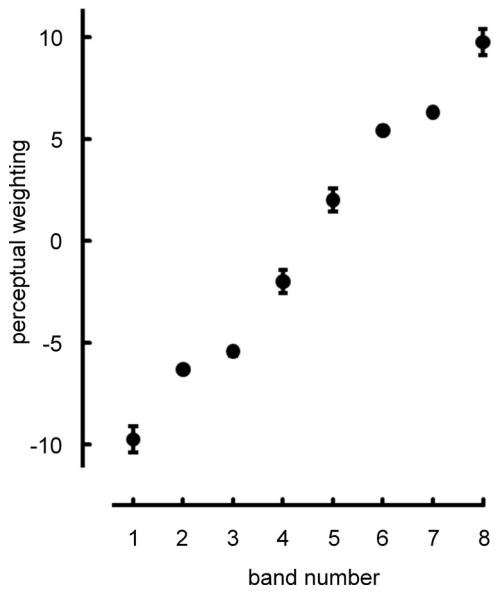

FIG. 8. The perceptual weighting of each of the vocoder's bands in experiment 2, which measured the shifts in category boundaries that are brought about as the test-word's distance is increased while the context's distance is held at 0.32 m. For each band (n), the perceptual weighting is calculated by multiplying the shift in a condition (S_{cond}), shown in Fig. 7, by the coefficient that indicates whether the distance of the band is varied ($V_{n,\text{cond}} = \pm 1$, shown in Fig. 6), and then summing these products over the six conditions, as indicated in Eq. (1). Bars are the mean ± two standard errors, and are shown where they extend beyond the data point.

IV. EXPERIMENT 3: PERCEPTUAL WEIGHTING OF BANDS IN THE CONTEXT

A. Rationale

The perceptual weightings of bands in test words show a high-pass characteristic, but for bands in the context, it is possible that there are different factors influencing the frequency-pattern of their perceptual weights. The influence of other factors might come about if compensation is effected through a gathering of information about the reflections' level across frequencies. To test this idea, experiment 3 measures effects brought about by the context's bands.

The experiment again uses the three subset-pair patterns (Fig. 6), but here, these refer to bands in the context, while the dependent variable is now the perceptual-constancy effect. To assess this effect, shifts brought about by increasing the BRIR's distance in all the test-word's bands are measured, and then these shifts are compared across conditions where the context is near (at 0.32 m) or far (at 10 m). A constancy effect from the context will result in a difference between these shifts.

The test-word's distance in this experiment now refers to the distances of all its bands, while the contexts' distances are "nominal" in that the BRIRs' distances are only varied in the subset halves of its bands, while the distances of the remaining bands are fixed. Results are again pooled to assess the degree of association between variation of reflections' level in a particular band, and the attendant variation of category-boundary differences across the subsets. The assumption is that this association will be stronger for the more heavily-weighted bands in the context. The experiment asks whether the resulting perceptual-weighting function is similar to that found for the test-word's bands in experiment 2.

B. Method

Listeners heard test words and contexts with reflection patterns (BRIRs) whose nominal distances were 0.32 or 10 m, giving all four combinations. The BRIRs on the test-word's bands were all at the same distance. The BRIRs on certain bands of the context were held at 0.32 m at both its nominal distances. This was done in the same way as for experiment 2's test-words, using the patterns shown in Fig. 6. This gives the six conditions (cond = 1,2,…,6) with the different patterns of coefficients ($V_{n,\text{cond}}$). Here, these coefficients indicate whether the BRIR on a context band (n) should be held at 0.32 m ($V_{n,\text{cond}} = -1$), or varied with the condition's nominal distance between 0.32 and 10 m ($V_{n,\text{cond}} = +1$). This experiment measures the perceptual compensation for effects of room reflections in each condition, which is calculated from a difference between two category boundary shifts. One of these shifts, $S_{a,\text{cond}}$ is the difference between category boundaries in conditions with the test-word's distances at 0.32 and 10 m when the context's nominal distance is 0.32 m. The other shift, $S_{b,\text{cond}}$ is the corresponding category-boundary difference when the context's nominal distance is 10 m. The

perceptual weighting of the context's bands was then calculated from

$$\text{perceptual weighting of band } n = \sum_{\text{cond}=1}^{\text{cond}=6} \left(S_{a,\text{cond}} - S_{b,\text{cond}}\right) V_{n,\text{cond}}. \quad (2)$$

There were three groups of six listeners (a, b, and c) who each heard conditions with an odd and an even value of condition, which were paired; 1 with 2, 3 with 4, or 5 with 6, for groups a–c, respectively. Over an experimental session, listeners heard eight-band speech at 2 test-word distances × 2 context distances × 2 odd-or-even cond × 11 continuum steps × 3 repeats = 396 trials. These trials were again randomly interspersed with unprocessed-speech fillers that comprised 2 test-word distances × 2 context distances × 11 continuum steps × 3 repeats = 132 trials. Overall this gives 528 trials for listeners, who each heard different randomized orderings. Other aspects of the method were the same as for experiments 1 and 2.

C. Results

Figure 9 shows means of category boundaries in conditions with eight-band speech for the three groups of six listeners. A four-way analysis of variance tested for differences among these category boundaries. There was a three-level between-subject factor, subject group (a, b, or c) along with a two-level within-subject factor indicating whether the value of the condition for the group was odd or even. A further two within-subject factors assessed the compensation effect and had two levels each; context distance and test-word distance.

The general pattern of results is indicated by three significant three-way interactions. The interaction among test-word's distance, context's distance and odd-or-even ($F_{(1,15)} = 37.91$, $p < 0.0001$), shows that there is generally more compensation from higher frequency (even numbered) bands. The other two three-way interactions show that differences between the odd and even numbered conditions tend to increase from group a to group c. This gives interactions among these two factors together with context's distance ($F_{(2,15)} = 5.72$, $p < 0.015$) and together with test-word's distance ($F_{(2,15)} = 10.07$, $p < 0.002$).

There are significant two-way interactions and main effects that seem to arise through the pattern of the higher-order interactions. The significant two-way interactions are test distance with context distance ($F_{(1,15)} = 180.45$, $p < 0.0001$), odd-or-even with context's distance $F_{(1,15)} = 71.63$, $p < 0.0001$), odd-or-even with test- word's distance ($F_{(1,15)} = 9.25$, $p < 0.009$), context's distance with group ($F_{(2,15)} = 4.67$, $p < 0.027$), and odd-or-even with group ($F_{(2,15)} = 19.55$, $p < 0.0001$). All four main effects are significant; context's distance ($F_{(1,15)} = 323.4$, $p < 0.0001$), test distance ($F_{(1,15)} = 710.09$, $p < 0.0001$), odd-or-even ($F_{(1,15)} = 90.37$, $p < 0.0001$), and group ($F_{(2,15)} = 7.79$, $p < 0.005$).

Figure 10 shows the perceptual weights of the context's bands, obtained with Eq. (2) from the means and standard errors of listeners' category-boundary shifts. These values are compared with values obtained for the test-word's bands in experiment 2 by plotting the two values for each band against each other. In both experiments there seems to be a similar monotonic increase in the band's perceptual weighting as its center frequency increases.

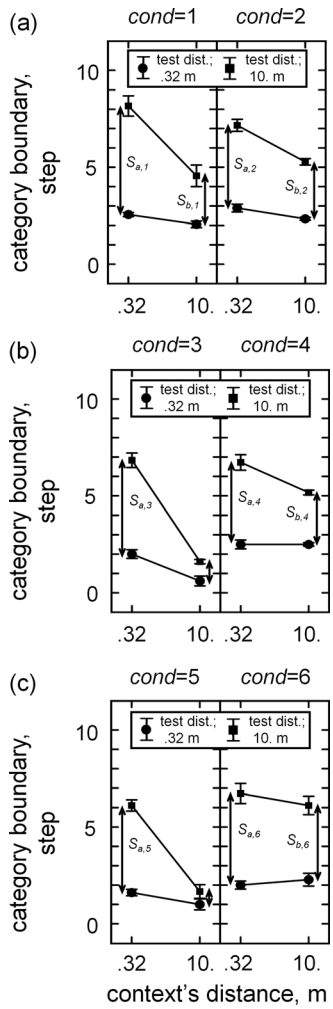

FIG. 9. The graphs show means and standard errors (bars) of category boundaries from experiment 3, where three groups of six listeners identified test words whose distance varied between 0.32 and 10 m in contexts whose nominal distance was also varied between 0.32 and 10 m. Graphs (a), (b), and (c) show results for the different listener-groups, who heard sounds from a condition (cond) where selections of the context's bands were held at 0.32 m, as given by the coefficients for these condition numbers in Fig. 6. Vertical arrows on the left-side of the panels show the shift in the category boundary effected when the test-word's distance is increased from 0.32 to 10 m and when the context's nominal distance is 0.32 m ($S_{a,\text{cond}}$). Vertical arrows on the right-side of panels show the corresponding shift that occurs when the context's nominal distance is 10 m ($S_{b,\text{cond}}$). The size of the perceptual compensation effect is indicated by the difference $S_{a,\text{cond}} - S_{b,\text{cond}}$.

D. Discussion

The high-pass pattern of perceptual weights for the context's bands in this experiment is also apparent from Fig. 9, which shows that in several respects, it is similar to the pattern found for the test-word's bands in experiment 2. The higher-frequency members of subset pairs are shown in the left-side panels of the figure, and in each case there are clear constancy effects, indicated by the differences between shifts at the two

test-word distances. The corresponding differences are much smaller in the lower-frequency subset pairs that are plotted in the right-side panels. This comparison shows that the constancy comes primarily from the bands of the higher-frequency members of a pair, while there is much less influence of the level of reflections in lower-frequency members. In one case, with the low frequency bands in condition 6, there is effectively no effect of constancy at all, even though these bands have high levels of room reflections when their nominal distance is at 10 m (Fig. 4). Hence, constancy seems to be confined to the upper 4 bands, and it is unaffected by the level of reflections in the context's lower-frequency bands. From this result it would seem that bands are largely independent in terms of their contribution to constancy effects.

This result-pattern is well captured in the pooled perceptual-weighting function shown in Fig. 10, where its similarity to the monotonically increasing pattern of perceptual weightings in experiment 2 is apparent. These results suggest that the factors operating to determine the perceptual weightings of test-word bands in experiment 2 are similar to those that bring about the pattern of perceptual weightings found here for the context's bands.

V. GENERAL DISCUSSION

A. Constancy and fine-structure effects

Experiment 1 assessed the possibility that the level of reflections in speech is indicated in perceptual constancy by aspects of the sound's fine structure. Possible fine-structure indicators that vary with the level of the reflections include the strength of harmonicity in the speech as well as the characteristic shaping of a sound's spectrum when reflections are present. Speech from a noise-band vocoder was used to give conditions where sounds had aspects of their fine-structure scrambled. In some of these conditions only the speech fine-structure was scrambled, while in others, the reflection pattern's fine-structure was scrambled as well. Results indicate that constancy effects are robust to both types of scrambling, as they are just as substantial in each of these conditions as they are in conditions using unprocessed speech. It would appear that this constancy is brought about through a processing of the temporal envelope, in a way that is independent of the fine-grained temporal structure of the signal. These results are therefore consistent with a perceptual mechanism that assesses the level of reflections from the prominence of tails in the temporal envelope (Stecker and Hafter, 2000; Watkins, 2005a,b).

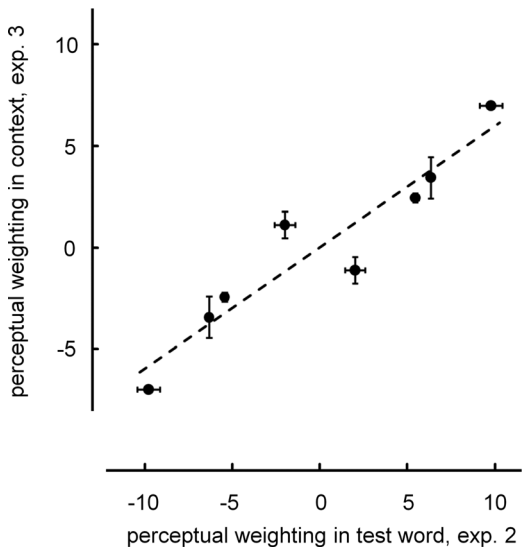

FIG. 10. The perceptual weighting of each of the vocoder's bands in experiment 3, which measured the shifts in category boundaries effected as the test-words' and the contexts' distances are varied. For each band (n), the perceptual weighting is calculated firstly by multiplying the difference between shifts in a condition ($S_{a,\mathrm{cond}} - S_{b,\mathrm{cond}}$, shown in Fig. 9), by the coefficient that indicates whether the distance of the band is varied ($V_{n,\mathrm{cond}} = \pm 1$, shown in Fig. 6). These products are then summed over the six conditions, as indicated in Eq. (2). Results from experiment 2 are also shown for comparison by plotting a band's perceptual weightings in the two experiments against one another. Bars are the mean ± two standard errors, and are shown where they extend beyond the data point. The linear regression line (dashed) is the least-squares fit and has a gradient of 0.60 with an intercept at zero. The corresponding correlation coefficient is significant (Pearson's $r_{(6)} = 0.94$, $p < 0.01$ two-tailed).

Conversely, there are other types of experiment where effects of preceding speech-contexts might plausibly have come about through some form of fine-structure processing (Brandewie and Zahorik, 2010). These experiments measured effects of speech-contexts on the intelligibility of subsequent target-words. The presence of contexts improved intelligibility in conditions with real-room binaural listening, but there was no such improvement in anechoic or monaural conditions. These authors characterized the mechanism responsible for these improvements as a form of "echo suppression," which is related to the binaural spatial effects seen in experiments using click stimuli and single, isolated reflections (e.g., Clifton, 1987).

It seems rather unlikely that results in the present experiments arise from this sort of echo suppression, as here and elsewhere (e.g., Watkins, 2005a) substantial temporal-envelope constancy is found in monaural conditions. Moreover, the lack of sensitivity to fine structure found here with temporal-envelope constancy seems unlike effects of an echo-suppression mechanism. Those effects tend to break down when there are small changes in the reflection pattern's fine structure, whereas, experiment 1 finds that temporal-envelope constancy is hardly affected when the reflection-pattern's fine structure is scrambled differently in the context and the test-word. Consequently, experiment 1 is a further indication of important differences between echo suppression and the constancy observed in the present type of experiment. Other evidence of such differences includes the observation that single reflection-clusters give little temporal-envelope constancy (Watkins, 2005b), and the observation that temporal-envelope constancy persists when the room is switched between the context and the test word (Watkins, 2005a).

It seems likely that there are at least two mechanisms responsible for perceptual constancy effects in speech, as suggested by Brandewie and Zahorik (2010) and by Watkins (2005a). One is essentially monaural, which is the temporal-envelope constancy that seems to operate in the present experiments, while the other mechanism is one that relies on inter-aural processing. However, it is not presently clear whether the difference between monaural and dichotic conditions observed by Watkins (2005a, experiment 3) comes

from the same perceptual mechanism as the intelligibility advantages in binaural listening observed by Brandewie and Zahorik (2010).

B. Constancy and perceptual weightings

The effect of increasing the reflections' level on test words was assessed in experiment 2, where the perceptual weightings of bands tended to follow the frequency characteristics of the frication in the [s] or [st] at the start of these words. This result is consistent with listeners giving more weight to the region of the spectrum where there are larger differences between the sounds' temporal envelopes.

A comparable scale of measurement was then used to assess the perceptual weightings of the context's bands in experiment 3, which asked whether different factors influence the frequency-pattern of perceptual weights when the constancy effect is the dependent variable. Such factors could have effects through a compensation that gathers information about the reflections' level across frequencies. However, experiment 3 shows that constancy effects from a band are little influenced by the reflections' level in other bands. Hence, the bands appear to be largely independent in terms of their influences on constancy.

The present results are consistent with the idea that constancy is effected by a within-band ("band-by-band") process that precedes the perceptual grouping involved in identification of the test word (Watkins *et al*., 2010a,b). This grouping stage is thought to be the primary determinant of the perceptual weightings, and so the same set of perceptual weightings that apply when reflections are added to test-word's bands will also apply for constancy effects from the context's bands. Watkins *et al*. suggest that a constancy that comes before grouping in this way indicates similarities with certain processes in visual perception (Palmer, Brooks, and Nelson, 2003), and is consistent with the difficulties listeners have with separating sounds in multi-source environments when they are hearing noise-vocoder sounds or when they are listening with cochlear implants (Nelson *et al*., 2003; Qin and Oxenham, 2003; Stickney *et al*. 2004; Poissant *et al*., 2006).

Both of the perceptual weighting functions obtained here seem to be determined by characteristics of the part of speech being tested, and in this respect, they resemble other assessments of the perceptual weightings of vocoder-bands (Apoux and Healy, 2010). However, other methods of assessing these weightings indicate the involvement of diverse factors. For example, Apoux and Bacon (2004) found that perceptual frequency-weightings could vary among different listening-conditions. Also, with a very different method again, Ming and Holt (2009), find that the importance function measured using the articulation index (Studebaker *et al*., 1987; ANSI, 1997), gives corresponding perceptual weightings in vocoder listening. This ANSI function is essentially an inverted U-shape, with a peak at 2 kHz, and so is quite unlike the monotonically increasing functions found here. So although the methods used in the present research give weightings that are sufficient for comparisons between its experiments, the broader implications of the particular weighting-values found are probably rather limited.

VI. CONCLUSIONS

(1) Although aspects of the signal's fine structure are affected by the level of room reflections, perceptual constancy from preceding speech-contexts is still apparent when the fine-structure is scrambled in noise-vocoded speech. Hence, the level of reflections in speech is only signaled to the constancy mechanism by aspects of the context's temporal-envelope.
(2) The perceptual weighting of bands in these experiments' "sir" vs "stir" test-sounds is a monotonically increasing function of their center-frequency, consistent with listeners giving more weight to frequency regions where there are larger differences between the two sounds' temporal envelopes.
(3) A similar pattern of perceptual weightings is found when constancy effects from the context's bands are measured, suggesting that the factors operating to determine the perceptual weightings of test-word bands are similar to those that bring about the pattern of perceptual weightings found here for the context's bands.
(4) The constancy effect from a band in the context seems largely independent of the reflections' level in other bands, consistent with the idea that constancy is brought about by a within-band process that precedes the perceptual grouping involved in identification of the test word.

ACKNOWLEDGMENTS

This work was supported by a grant to the first author from EPSRC. We are grateful to Amy Beeston, Guy Brown, Peter Derleth, Kalle Palomaki, and Hynek Hermansky for discussions.

ANSI. (**1997**). "Methods for calculation of the speech intelligibility index," American National Standards Institute, New York.
Apoux, F., and Bacon, S. P. (**2004**). "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," J. Acoust. Soc. Am. **116**, 1671–1680.
Apoux, F., and Healy, E. W. (**2010**). "Auditory channel weights for consonant recognition in normal-hearing listeners," J. Acoust. Soc. Am. **127**, 1991.
Bilsen, F. A. (**1967/68**). "Thresholds of perception of repetition pitch. Conclusions concerning coloration in room acoustics and correlation in the hearing organ," Acustica **19**, 27–32.
Bilsen, F. A., and Ritsma, R. J. (**1967/68**). "Repetition pitch mediated by temporal fine structure at dominant spectral regions," Acustica **19**, 114–115.
Brandewie, E., and Zahorik, P. (**2010**). "Prior listening in rooms improves speech intelligibility," J. Acoust. Soc. Am. **128**, 291–299.
Clifton, R. K. (**1987**). "Breakdown of echo suppression in the precedence effect," J. Acoust. Soc. Am. **82**, 1834–1835.
Culling, J. F., Summerfield, Q., and Marshall, D. H. (**1994**). "Effects of simulated reverberation on the use of binaural cues and fundamental frequency differences for separating concurrent vowels," Speech Commun. **14**, 71–96.
Culling, J. F., Hodder, K. I., and Toh, C. Y. (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.
Dorman, M. F., and Loizou, P. C. (**1998**). "The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels," Ear Hear. **19**, 162–166.
Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.
Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (**2001**). "Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants," J. Acoust. Soc. Am. **110**, 1150–1163.

Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Houtgast, T., and Steeneken, H. J. M. (**1973**). "The modulation transfer function in acoustics as a predictor of speech intelligibility," Acustica **28**, 66–73.
ISO 3382, (**1997**). "Acoustics—Measurement of the reverberation time of rooms with reference to other acoustical parameters," International Organization for Standardization, Geneva.
Longworth-Reed, L., Brandewie, E., and Zahorik, P. (**2009**). "Time-forward speech intelligibility in time-reversed rooms," J. Acoust. Soc. Am. **125**, EL13–EL19.
Ming, V. L., and Holt, L. L. (**2009**). "Efficient coding in human auditory perception," J. Acoust. Soc. Am. **126**, 1312–1320.
Nelson, P. B., Jin, S. H., Carney, A. E., and Nelson, D. A., (**2003**). "Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners," J. Acoust. Soc. Am. **113**, 961–968.
Palmer, S. E., Brooks, J. L., and Nelson, R. (**2003**). "When does grouping happen?" Acta Psychol. **114**, 311–330.
Poissant, S. F., Whitmal, N. A., and Freyman, R. L. (**2006**). "Effects of reverberation and masking on speech intelligibility in cochlear implant simulations," J. Acoust. Soc. Am. **119**, 1606–1615.
Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Roman, N., and Wang, D. L. (**2006**). "Pitch-based monaural segregation of reverberant speech," J. Acoust. Soc. Am. **120**, 458–469.
Schroeder, M. R. (**1968**). "Reference signal for signal quality studies," J. Acoust. Soc. Am. **44**, 1735–1736.
Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nat. (London) **416**, 87–90.
Stecker, G. C., and Hafter, E. R. (**2000**). "An effect of temporal asymmetry on loudness," J. Acoust. Soc. Am. **107**, 3358–3368.
Stickney, G. S., Zeng, F-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1987**). "A frequency importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130–1138.
Watkins, A. J. (**2005a**). "Perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **118**, 249–262.
Watkins, A. J. (**2005b**). "Perceptual compensation for effects of echo and of reverberation in speech identification," Acust. Acta Acust. **91**, 892–901.
Watkins, A. J., and Makin, S. J. (**2007a**). "Perceptual compensation for reverberation in speech identification: Effects of single-band, multiple-band and wideband contexts," Acta Acust. United Acust. **93**, 403–410.
Watkins, A. J., and Makin, S. J. (**2007b**). "Steady-spectrum contexts and perceptual compensation for reverberation in speech identification," Acoust. Soc. Am. **121**, 257–266.
Watkins, A. J., Raimond, A. P., and Makin, S. J. (**2010a**). "Room reflections and constancy in speech-like sounds: Within-band effects," in *The Neurophysiological Bases of Auditory Perception*, edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer, New York), pp. 439–447.
Watkins, A. J., Raimond, A. P., and Makin, S. J. (**2010b**). "Constancy in the perception of speech when the level of room-reflections varies," in *Binaural Processing and Spatial Hearing. ISAAR—International Symposium on Auditory and Audiological Research*, edited by J. Buchholz, T. Dau, J. Dalsgaard, and T. Poulsen (The Danavox Jubilee Foundation, Ballerup, Denmark), pp. 371–380.