# Constancy in the perception of speech when the level of room-reflections varies

Anthony Watkins, Simon Makin, and Andrew Raimond

*Department of Psychology, University of Reading, Reading RG6 6AL, United Kingdom*

A speech message played several metres from the listener in a room is usually heard to have much the same phonetic content as it does when played nearby, even though the different amounts of reflected sound make the temporal envelopes of these signals very different. To study this 'constancy' effect, listeners heard speech messages and speech-like sounds comprising 8 auditory-filter shaped noise-bands that had temporal envelopes corresponding to those in these filters when the speech message is played. The 'contexts' were "next you'll get _to click on", into which a "sir" or "stir" test word was inserted. These test words were from an 11-step continuum, formed by amplitude modulation. Listeners identified the test words appropriately, even in the 8-band conditions where the speech had a 'robotic' quality. Constancy was assessed by comparing the influence of room reflections on the test word across conditions where the context had either the same level of room reflections (i.e. from the same, far distance), or where it had a much lower level (i.e. from nearby). Constancy effects were obtained with both the natural- and the 8-band speech. Results are considered in terms of the degree of 'matching' between the context's and test-word's bands.

## INTRODUCTION

The speech-like sounds studied in the present experiment were produced by a 'sparse' noise-excited vocoder, in which a speech recording is passed through an 8-filter 'bank' before obtaining the temporal envelope in each of these channels. Each envelope is then applied to a narrowband noise that has the channel's centre-frequency and bandwidth. When the filter-bank's centre-frequencies span the speech range, signals obtained by adding the processed bands together are heard to be distinctly speech-like, and the original message is quite intelligible (Shannon *et al*., 1995). Such results indicate the importance of information from these narrow-band temporal-envelopes in speech perception.

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, 2007a). 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, it gives the perceptual compensation for effects of room reverberation that is seen in experiments

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation, 2010.

where room reflections in context-speech have compensatory perceptual effects on adjacent test-words (Watkins, 2005).

One index of the amount of reverberation present in these contexts is the degree of modulation attenuation at modulation-rates that are significant in speech (Houtgast and Steeneken, 1973). However, experiments reported by Watkins (2005), using contexts that have reversed reverberation, showed that the constancy mechanism is independent of modulation attenuation in important respects. The compensation mechanism was found to be sensitive to the characteristic time-direction of reverberation, as compensation varied with the prominence of the tails that reverberation adds at the ends of sounds and at spectral-transitions in auditory filters. Compensation was not obtained with reversed reverberation, where there are no prominent tails, even though the modulation attenuation is much the same as it is with forwards reverberation (Longworth-Reed, *et al*., 2009).

Further experiments, with steady-spectrum noise contexts, have indicated that while tails from reverberation are not sufficiently prominent when these sounds have smoothly-varying temporal envelopes, they become more prominent with noise signals that have sharply-varying temporal envelopes, such as those that arise in auditory filters when speech is heard. Consequently, compensation can be substantial with certain contexts that are unintelligible noises (Watkins and Makin, 2007b), and so it seems unlikely that phonetic perception is involved. These and other results support the idea that a more general, perceptual-constancy mechanism is involved in this compensation.

The experiments reported here ask whether compensation for the effects of room reverberation is brought about in a 'band-by-band' manner. Such compensation would give constancy effects within frequency bands, but not between them. This idea is tested with the 8-band vocoder speech and with reflection patterns containing either a high or a low level of reflected sound. The higher-level patterns are applied to only half (4) of the test word's bands, while the reflection pattern in the other bands is held at the lower level. In matched conditions, the context has the higher-level patterns in bands that the test-word does, while in mismatched conditions it does not. The experiments test the band-by-band idea by looking for reduced compensation in mismatched conditions.

## METHOD

### Speech contexts and the test-word continuum

The methods described by Watkins (2005) were used to obtain context phrases containing test-words from a continuum between "sir" and "stir". 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 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.

**Category boundaries**

When room reflections oppose the amplitude modulation that formed the continuum, they obscure cues to the presence of a [t] in test words, so more of the continuum's steps will be identified as "sir". To indicate differences between conditions in the number of steps that are identified as "sir", 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 4 presentations of each of the continuum's steps played in the context, and category boundaries here were found from the total of 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.

**Room reflections**

The methods described by Watkins (2005) were also used to introduce room reflections into the dry contexts and test words by convolution with room impulse-responses. This gives the effect of monaural real-room listening over headphones. The monaural impulse-responses were obtained in rooms using dummy-head transducers (a speaker in a Bruel & Kjaer 4128 head and torso simulator, and a Bruel & Kjaer 4134 microphone in the ear 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 room impulse responses 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 m 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), which was 8 dB at 0.32 m, descending to 2 dB at 10 m.

**8-band speech**

The individual 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 centre frequency, as given by the 'Cambridge ERB' (Glasberg and Moore, 1990). The 8 centre-frequencies were equally log-spaced across the speech range, starting at 250 Hz, and increasing by intervals of a musical fifth (7/12 octave). Bands were numbered from low to high centre-frequency, using a band number, n=1,2, … 8.

To obtain one of these bands, the speech was played through an auditory filter, followed by a 'signal correlated noise' operation, which involves reversing the polarity of a randomly selected half of the signal's samples. 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 then added to each band, and the bands were summed. The relative levels of the bands were adjusted with a 'speech-shaping' filter, whose frequency response was the long-term average spectrum of the original speech-context.

**Designs**

In experiment 1, the reflection pattern's distance was varied between 0.32 m and 10 m to give the different context-distance and test-word distance conditions. The 8-band speech of the test word was divided between interleaved sets of odd-numbered and even-numbered bands, and the distance manipulation was applied to the even-numbered bands while the others were held at 0.32 m. In a mismatched condition, the distance was varied in odd-numbered bands of the context while the other bands were omitted. In a matched condition, distance was varied in the even numbered bands of the context while the other bands were omitted. Experiment 2 was similar except that bands held at 0.32 m were substituted for the context bands that were omitted in experiment 1. The distances of bands in experiments 1 and 2 are shown in Table. 1.

| | | | Experiment 1 contexts | | | | Experiment 2 contexts | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Test word | | Matched | | Mismatched | | Matched | | Mismatched | |
| Nominal distance: | .32 | 10. | .32 | 10. | .32 | 10. | .32 | 10. | .32 | 10. |
| Distance of odd-numbered bands: | .32 | .32 | no band | no band | .32 | 10. | .32 | .32 | .32 | 10. |
| Distance of even-numbered bands: | .32 | 10. | .32 | 10. | no band | no band | .32 | 10. | .32 | .32 |

**Table 1**: Distances in metres of the room impulse-responses that were applied to the 8 bands of the sparse noise-excited vocoder in different experimental conditions.

The 6 listeners in each experiment identified test-word continua in unprocessed speech conditions as well in their matched and mismatched conditions. All combinations of the context and test-word distance were presented in each of these conditions, and each listener received the trials in a different randomized order.

### Procedure

Sounds were presented to listeners 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. Before the 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 8-band sounds as speech. Trials were administered to listeners in individual sessions 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 following trial.

## RESULTS

For each condition in both experiments, category boundaries were pooled across the 6 listeners, and the resulting means are shown with their standard errors in Fig. 1.

Results with unprocessed speech replicate the constancy effects reported in earlier work (e.g. Watkins, 2005). When the context is nearby, increasing the test word's distance causes more of the continuum's members to be heard as "sir", so there is a corresponding increase in the category boundary. However, when the context's distance is also increased to 10 m, there is a compensation effect, giving a reduction in the category boundary.

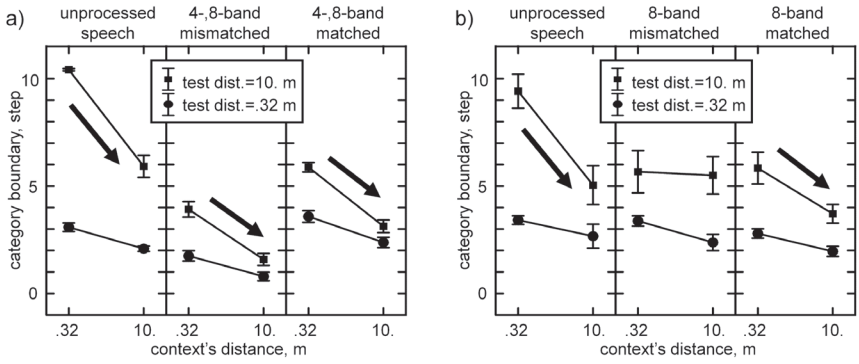


**Fig. 1: Panel a**: Category boundaries in experiment 1. **Panel b**: Category boundaries in experiment 2. Data points are means with standard-error bars. Constancy effects are arrowed.

With the 8-band speech in experiment 1, increasing the test word's distance causes more of the continuum's members to be heard as "sir". The corresponding increases in category boundaries are smaller than with unprocessed speech, which probably reflects the contribution of the test-word bands that were held at 0.32 m. However, in both matched and mismatched conditions there is a compensation effect. In experiment 2, increasing the test word's distance also increases category boundaries, but now the compensation effect is eliminated in mismatched conditions.

The pattern of results in experiment 2 supports the band-by-band hypothesis, which was tested statistically with a 3-way within-subject analysis of variance on data in the conditions with 8-band speech, using the 2-level factors 'test distance', 'context distance', and 'matched vs. mismatched'. The crucial 3-way interaction is significant with $F(1,5)=11.68$, and $p<0.02$.

The corresponding 3-way interaction in experiment 1 is not significant. However, the overall constancy effect, which is the 2-way 'test distance' and 'context distance' interaction, is substantial, with $F(1,5)=24.40$ and $p<0.005$. Additionally in experiment 1, there is a main effect of the 'matched vs. mismatched' factor where $F(1,5)=24.40$ and $p<0.005$, reflecting the higher category boundaries in matched conditions.

## Discussion

In experiment 2, the constancy effect is apparent in matched conditions, but is eliminated in mismatched conditions, so in this case, compensation from the context appears to arise in a band-by-band fashion.

The band-by-band result-pattern is not found in experiment 1, where compensation arises with both the matched and the mismatched contexts. A possible reason for this

is that tails from the reverberation will arise not only near the centre-frequencies of the vocoder's bands, but also away from these frequencies, in the skirts of the filters. These frequencies will extend to the putative 'mismatched' bands, and may thereby be sufficient to effect compensation. These off-frequency tails will be obscured by the added bands that are held at 0.32 m, but only in experiment 2's contexts.

These off-frequency tails are shown in the temporal envelopes that are plotted in Fig. 2. Here, the experiments' contexts in mismatched conditions have been passed through auditory (gammatone) filters centred at the upper four centre-frequencies of the vocoder (bands 5 through 8). The tails that have been found to be crucial for compensation can be seen in all these bands in experiment 1's 10-m stimuli, whereas in experiment 2's 10-m stimuli, the tails are only seen in the (mismatched) odd-numbered bands.

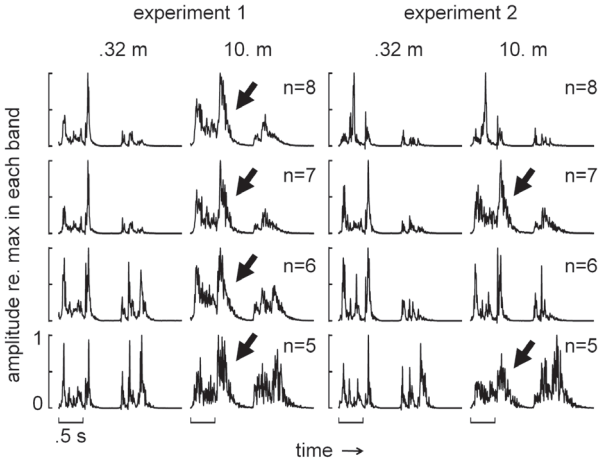


**Fig. 2**: Temporal-envelopes in the upper 4 bands of the context in mismatched conditions, with some of the more prominent tails arrowed. Amplitudes are relative to the band's maximum, so the lower level in experiment 1's even-numbered bands is not shown.

The higher category boundaries in experiment 1's matched conditions seem to arise from a combination of two factors. One of these concerns the long-term spectrum of the 4-band contexts, which has a pattern of peaks and dips. The peaks are at the centre-frequencies of each of the 4 bands, while the dips are near the centre frequencies of the omitted bands. Such long term spectral characteristics of speech contexts have been shown to have 'inverse filtering' effects on subsequent speech and other sounds (Watkins, 1991). It is as if the context's long-term spectrum is inverted and applied to the subsequent sound, and such inverse-filtering describes result-patterns across contexts with diverse long-term spectral characteristics (Watkins

and Makin, 1996). The effect of this type of filtering in experiment 1 will be to raise the relative level of the 4 test-word bands that were omitted from the context, and to lower the relative level of the other 4 test-word bands. Consequently, the effective level of the test-word's even-numbered bands will be lowered in matched conditions. The other factor concerns the relative importance of the different test-word bands for the perception of the [t] that distinguishes the test words. If the even-numbered bands are more important for hearing the [t] in 'stir', then the effect of attenuating them through inverse-filtering will be to increase 'sir' responses. This will give rise to higher category boundaries in matched conditions, which is the effect observed in experiment 1. Findings from some preliminary experiments have confirmed this pattern of importance across the test word's bands.

The 8-band vocoder-speech used in the present experiments is heard to be distinctly speech-like. This effect seems to be a classic example of perceptual grouping, as the speech-like quality of the summed-band 'whole' is not at all apparent when any of the individual-band 'parts' are played in isolation. Recently, visual researchers have investigated how perceptual grouping stands in relationship to constancy mechanisms, asking whether constancy precedes or follows grouping (Palmer *et al*., 2003). These authors concluded that perceptual grouping is ubiquitous, in that it occurs for each level of representation. Consequently, grouping in vision can occur before, after, and even during different types of constancy operation. The present experiments indicate that the form of constancy studied here is a band-by-band process, so it precedes the across-band grouping that gives the vocoder signals their speech-like quality.

**Acknowledgements**

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

## REFERENCES

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.

Palmer, S. E., Brooks, J. L., and Nelson, R. (**2003**). "When does grouping happen?" Acta Psychologia, **114**, 311-330.

Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science, **270**, 303–304.

Watkins, A. J. (**1991**). "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **90**, 2942–2955.

Watkins, A. J., and Makin, S. J. (**1996**). "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **99**, 3749-3757.

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

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 Acustica united with Acustica **93**, 403-410.

Watkins, A. J., and Makin, S. J. (**2007b**). "Steady-spectrum contexts and perceptual compensation for reverberation in speech identification," J. Acoust. Soc. Am. **121**, 257-266.