Title:

Room reflections and constancy in speech-like sounds: Within-band effects

Running title:

Room reflections and constancy

P.A.C.S. numbers:

43.55.Hy, 43.71Gv, 43.71.Es, 43.71.An, 43.66.Ba

Authors:

Anthony J. Watkins, Andrew Raimond and Simon J. Makin

Correspondence to first author at:

Department of Psychology

The University of Reading

Reading RG6 6AL

UK

Phone: +44 (0)118-378-7559

Fax: +44 (0)118-378-6715

Email: syswatkn@reading.ac.uk

1. Introduction

Although the visual systems of organisms respond to the physical properties of light, such as its wavelengths or its luminance level, the information that they provide is about meaningful 'real-world' properties of things, such as their colour, or lightness. The physical properties of light from an object are affected by viewing conditions, such as the shade that affects objects' luminance in sunlight, but there is a compensation for effects of viewing conditions in visual processing, so that a light grey surface in dim shade can be distinguished from a black surface in full sunlight, even though the luminance of the surfaces might be the same. This ability to identify real-world properties in various viewing conditions is known as perceptual constancy, and vision uses information from the surrounding context in order to achieve this (Adelson 2000). Recently, visual researchers have investigated how perceptual grouping stands in relationship to constancy mechanisms, asking whether constancy precedes or follows grouping (Palmer, Brooks and Nelson 2003). The present research asks the same question of hearing.

The physical property of a sound that is studied here is the array of temporal envelopes that arises in the frequency channels of devices such as spectrographs or vocoders, and at the earliest stages of auditory processing (Glasberg and Moore 1990). When speech signals are played, the information that hearing provides is predominantly about the real-world phonetic content of the message. Constancy arises when a speech message is played several metres away from the listener in a room, where it is usually heard to have much the same phonetic content as it does when played nearby, even though the different amounts of reflected sound from the room's surfaces make the temporal envelopes of the two signals very different (Watkins and Makin 2007). This perceptual constancy would appear to result from the surrounding context being taken into account, as certain words recorded from the distant

message are heard as other words when they are played in a context that is recorded nearby. This constancy seems to arise from a perceptual compensation for the effects of room reverberation, which is seen in experiments where room reflections in context-speech have compensatory perceptual effects on adjacent test-words (Watkins 2005).

The speech-like sounds used in the present experiment were obtained by signal processing whereby the 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). This outcome seems to be a classic example of a grouping effect, 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.

Compensatory perceptual effects of context-speech on a test-word could happen either before or after the bands are grouped, as described in Fig. 1 for an 8-band signal. A before-grouping mechanism might act within each band so that the compensation is effected in a 'band-by-band' manner. This idea is tested here by applying distant (10-m) room-reflection patterns to only half (4) of the context's bands, while holding the reflection-pattern in the other bands at a nearby distance (0.32 m). In matched conditions, the test word has 10-m reflection-patterns in bands that the context does, while in mismatched conditions it does not. The experiment tests the band-by-band idea by looking for reduced compensation in mismatched conditions.

2. Method

2.1 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 booth, giving ‘dry’ speech. The context phrase was originally such a recording, of “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 that had a corner frequency of 50 Hz. 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.

2.2 Category boundaries

When room reflections obscure cues to the presence of a [t] in test words they oppose the amplitude modulation that formed the continuum, so more of the continuum's steps will be identified as "sir" in conditions where this happens. 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.

2.3 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 and Kjaer 4128 head and torso simulator, and a Bruel and 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. To obtain different amounts of reflected sound in a 'natural' way, different distances between the dummy-head transducers of the talker and listener were used, as the proportion of reflected sound, relative to direct sound energy, increases with source-to-receiver distance. 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. The amount of reflected sound at these distances is indicated by the time taken for the room's impulse-response energy to decay by 10 dB, 'EDT' (ISO 3382. 1997). At 10 m the A-weighted EDT was 0.14 s, while at 0.32 m this EDT was less than 0.01 s.

2.4 8-band speech

The individual bands were narrow-band 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, $n=1,2, \ldots 8$.

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 played through the auditory filter again, to eliminate frequencies outside the filter's band. The impulse response of this

second filter was reversed in time to correct for delays introduced by the operation of the first filter.

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. Room-reflection patterns were then added to each band, using distances appropriate for the experimental condition, and the 8 bands were summed.

2.5 Design

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 was divided between interleaved sets of odd-numbered and even-numbered bands, and the distance manipulation was applied to some of these bands while the others were held at 0.32-m. In a mismatched condition, the distance was varied in the odd-numbered bands of the context but in the even-numbered bands of the test word. In a matched condition for one of the listener groups, distance was varied in the even-numbered bands of both the context and the test-word. The other group had a different type of matched condition, where distance was varied in the odd-numbered bands of the context but in all 8 bands of the test word. These conditions are illustrated in Fig. 2.

The 12 listeners were evenly divided between the two groups. Each group identified test-word continua in unprocessed speech conditions as well in their matched and mismatched 8-band 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.

2.6 Procedure

Sounds were delivered to listeners at a peak level of 48 dB SPL through the left earpiece of Sennheiser HD480 headphones 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. The computer waited for the listener to respond before presenting the following trial.

3. Results

For each condition, category boundaries were pooled across the 6 listeners, and the resulting means are shown with their standard errors in Fig. 3.

Results with unprocessed speech replicate the compensation effects reported in earlier work (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, which is indicated by the arrows in the leftmost panels of Fig 3.

Results with 8-band speech depend on whether the test-word's bands are matched to those of the context in the crucial 10-m conditions. In mismatched conditions there is no compensation effect, as indicated by the absence of any arrows in the centre panels of Fig. 3. In matched conditions however, there is a compensation effect, which is indicated by the arrows in the rightmost panels of Fig. 3.

This pattern of results supports the band-by-band hypothesis, which was tested statistically with a 4-way analysis of variance, using the combination of 2-level factors that describe the conditions in the centre and rightmost panels of Fig. 3. The 4-way interaction was not significant, but the crucial interaction was among the three factors; 'test-word's distance', 'context's distance', and 'matched vs. mismatched', which was found to be significant with $F(1,10)=20.7$, and $p<0.0012$.

4. Discussion

Room-reflections in both the even-numbered and the odd-numbered frequency-bands seem to have some effect on the test word. Increasing the reflections' distance in test-words' even bands increases the number of "sir" responses in both matched and mismatched conditions, while for group 1, increasing the distance of test words' odd bands in the matched conditions also increases the number of "sir" responses for the near contexts. This last effect from the odd bands appears as an increase over effects from the even bands alone. Therefore, information about the [s] vs. [st] distinction needed to classify the test word seems to be distributed across more than one of the frequency bands in these 8-band stimuli.

Both sets of bands also seem to bring about compensation, as for group 1, this effect is from the odd-numbered bands in the context, while it is from the even-numbered bands in group 2.

However, this compensation effect is only seen in matched conditions, where the test-word and context share bands that have the 10-m room-reflection pattern.

Is perception governed by the less distorted, 0.32-m bands in these sounds? This might happen if hearing behaves like a 'missing data' speech recogniser, and bases its decisions on the less distorted parts of the signal (Palomäki, Brown and Barker 2002). Clearly, this could not be happening on a word by word basis with the listeners in this experiment, as there would be no effects of distance when only 4 of the test-word's bands are given the 10-m reflection patterns. A related idea is that the less distorted, 0.32-m bands are selected in the context and then 'tracked' through the test word, perhaps by an attention-like process, while the other bands are effectively ignored. Such a tracking could account for results with group 2, as it would give the effect of distance in mismatched conditions, where the tracked bands would be the ones that are at 10-m in the test word. This sort of tracking could also give the compensation effect in group 2's matched conditions, where the tracked bands would become the four that are also at 0.32 m in the test word. However, the compensation found with group 1 is not consistent with this 'band-tracking' idea, as compensation is obtained across conditions where all the test-word's bands are at 10 m.

A substantial part of the grouping that occurs when hearing these 8-band sounds is likely to be attentional in nature (Cooper and Roberts 2007). It is also possible that a more 'primitive', pre-attentive grouping process operates as well; perhaps one that is informed by amplitude modulation, and that exploits the correlations among the bands' temporal envelopes at the lower modulation frequencies in speech signals (Crouzet and Ainsworth 2001). Such a grouping may be 'obligatory' (Roberts, Glasberg and Moore 2002), so that an individual band is difficult to 'track' independently of the others.

Whatever processes bring about the cross-channel grouping in these 8-band sounds, it can be said from the present results that the constancy mechanism precedes the grouping. This is not to say that grouping at other levels is not involved, in particular, a sequential within-band grouping might well operate prior to the constancy operation. This view is consistent with observations about diverse types of grouping in vision, where it has been 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 (Palmer, Brooks and Nelson 2003).

Acknowledgements

This work was supported by a grant to the first author from EPSRC. We are grateful to Amy Beeston and Guy Brown for discussion.

References

Adelson EH (2000) Lightness Perception and Lightness Illusions. In: Gazzaniga M (ed) The New Cognitive Neurosciences, 2nd edn. MIT Press, Cambridge MA

Cooper HR and Roberts B (2007) Auditory stream segregation of tone sequences in cochlear implant listeners. Hear Res 225: 11–24

Crouzet O, Ainsworth WA (2001) On the various influences of envelope information on the perception of speech in adverse conditions: An analysis of between-channel envelope correlation. *CRAC Workshop (Consistent and Robust Acoustic Cues for sound analysis)* Aalborg, Scandinavia

Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47: 103-138

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

Palmer SE, Brooks JL, Nelson R (2003) When does grouping happen? Acta Psychologia, 114: 311-330.

Palomäki KJ, Brown GJ, Barker J (2002) Missing data speech recognition in reverberant conditions. In Proceedings of 2002 International Conference on Acoustics, Speech, and Signal Processing (ICASSP2002). pp 65-68

Roberts B, Glasberg BR, Moore BCJ (2002) Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. J Acoust Soc. Am 112: 2074-2085

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

Watkins AJ (2005) Perceptual compensation for effects of reverberation in speech identification. J Acoust Soc Am 118: 249-262

Watkins AJ, Makin, SJ (2007) Perceptual compensation for reverberation in speech identification: Effects of single-band, multiple-band and wideband contexts. Acta Acustica united with Acustica 93: 403-410

Figure captions

Figure 1. Two ideas about the way that grouping stands in relation to constancy in the 8-band speech used in this experiment.

Figure 2. Diagrammatic representations of the room-reflection content of the 8 bands of the speech in conditions where the context's distance and test-word's distance were both 10 m.

Figure 3. Means and standard errors of category boundaries in the experiment's conditions. The arrows indicate compensation effects.

Fig 1

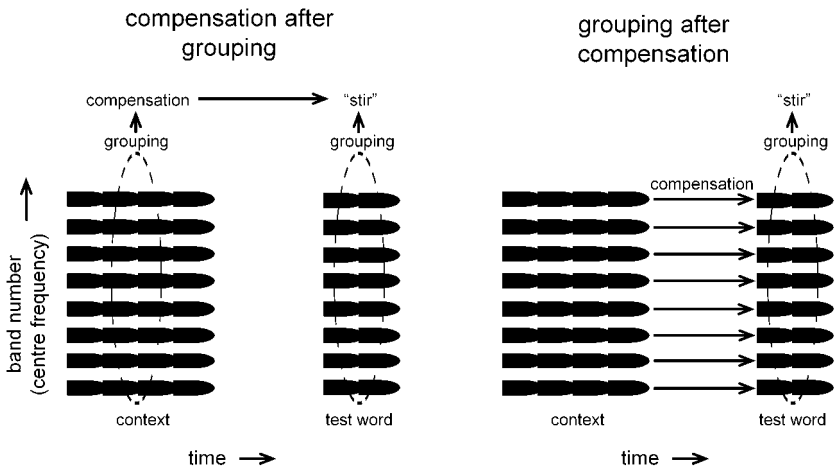

compensation after grouping
compensation
"stir"
grouping
grouping
band number (centre frequency)
context
test word
time
grouping after compensation
"stir"
grouping
compensation
context
test word
time

Fig. 2.

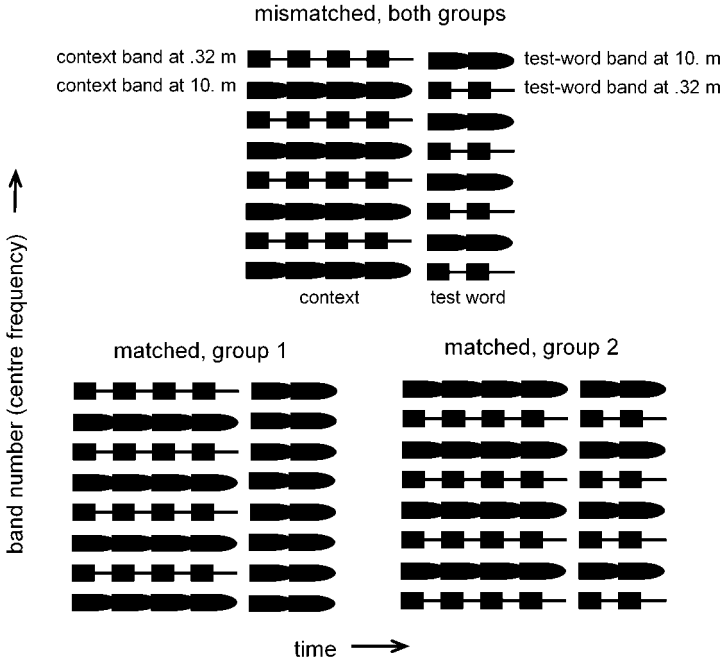
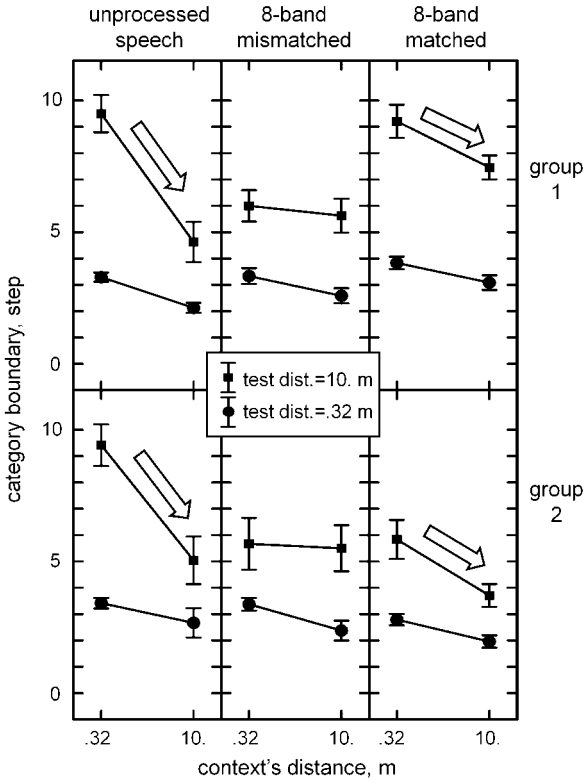

mismatched, both groups
context band at .32 m
context band at 10. m
test-word band at 10. m
test-word band at .32 m
context
test word
band number (centre frequency)
matched, group 1
matched, group 2
time

Fig. 3.


unprocessed speech
8-band mismatched
8-band matched
group 1
group 2
test dist.=10. m
test dist.=.32 m
category boundary, step
context's distance, m
10
5
0
.32
10.

Appended Abstract

Title: Room reflections and constancy in speech-like sounds: Within-band effects

Authors: Anthony J. Watkins, Andrew Raimond and Simon J. Makin

Correspondence to first author at:

Department of Psychology

The University of Reading

Reading RG6 6AL

UK

Phone: +44 (0)118-378-7559
Fax: +44 (0)118-378-6715
Email: syswatkn@reading.ac.uk

Abstract text

The experiment asks whether constancy in hearing precedes or follows grouping. Listeners heard speech-like sounds comprising 8 auditory-filter shaped noise-bands that had temporal envelopes corresponding to those arising in these filters when a speech message is played. The 'context' words in the message 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 that was formed by amplitude modulation. Listeners identified the test words appropriately and quite consistently, even though they had the 'robotic' quality typical of this type of 8-band speech. The speech-like effects of these sounds appears to be a consequence of auditory grouping. 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 reflections, or where it had a much lower level. Constancy effects were obtained with these 8-band sounds, but only in 'matched' conditions, where the room reflections were in the same bands in both the context and the test word. This was not the case in a comparison 'mismatched' condition, and here, no constancy effects were found. It would appear that this type of constancy in hearing precedes the across-channel grouping whose effects are so apparent in these sounds. This result is discussed in terms of the ubiquity of grouping across different levels of representation.

Keywords

Hearing, speech, constancy, room reflections, grouping