Hearing Research 354 (2017) 16–27

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research Paper

Stream segregation of concurrent speech and the verbal transformation effect: Influence of fundamental frequency and lateralization cues

Marcin Stachurski, Robert J. Summers, Brian Roberts*

Psychology, School of Life and Health Sciences, Aston University, Birmingham, B4 7ET, UK

ARTICLE INFO

Article history:
Received 4 May 2017
Received in revised form
25 July 2017
Accepted 31 July 2017
Available online 2 August 2017

Keywords:
Verbal transformation effect
Auditory grouping
Concurrent speech segregation
Fundamental frequency
Interaural time difference
Within-ear interaction

ABSTRACT

Repeating a recorded word produces verbal transformations (VTs); perceptual regrouping of acoustic-phonetic elements may contribute to this effect. The influence of fundamental frequency (F0) and lateralization grouping cues was explored by presenting two concurrent sequences of the same word resynthesized on different F0s (100 and 178 Hz). In experiment 1, listeners monitored both sequences simultaneously, reporting for each any change in stimulus identity. Three lateralization conditions were used – diotic, ±680-μs interaural time difference, and dichotic. Results were similar for the first two conditions, but fewer forms and later initial transformations were reported in the dichotic condition. This suggests that large lateralization differences *per se* have little effect – rather, there are more possibilities for regrouping when each ear receives both sequences. In the dichotic condition, VTs reported for one sequence were also more independent of those reported for the other. Experiment 2 used diotic stimuli and explored the effect of the number of sequences presented and monitored. The most forms and earliest transformations were reported when two sequences were presented but only one was monitored, indicating that high task demands decreased reporting of VTs for concurrent sequences. Overall, these findings support the idea that perceptual regrouping contributes to the VT effect.

1. Introduction

It has long been known that repeating aloud a word to oneself over and over leads to the sound of that word losing its meaning (e.g., Titchener, 1915, pp. 26–27); this lapse in meaning is called verbal satiation. A closely related phenomenon is the verbal transformation effect (VTE), in which listeners report changes in verbal form when a recording of a spoken word is repeated many times (Warren, 1961a; for reviews, see Warren, 1996, 2008). The VTE involves a series of abrupt changes in the perception of the speech signal, some to new forms and others back to forms previously reported. Notably, these alternative forms often involve complex phonetic distortion of the stimulus. The VTE is not simply a laboratory curiosity; it can provide insights into how the auditory system processes ambiguous sensory information and switches between alternative interpretations of that information.

The changes in verbal form which characterize the VTE were originally interpreted mainly in terms of linguistic processes. Specifically, it has been argued that verbal satiation (adaptation) of a given form occurs once that form has been perceived for a time and a new perceived form emerges from among competing lexical candidates (or sometimes phonologically plausible non-words) as a result of criterion shift (Warren, 1968). These processes continue and the new form itself undergoes satiation, replacement, and recovery from adaptation. More generally, the VTE has been seen as related to changes in the perception of connected discourse that may occur when the initial linguistic interpretation is not confirmed by subsequent context (Warren, 1968), and hence as related to mechanisms normally used to resolve ambiguities and correct errors when listening to speech (Warren and Warren, 1970; Obusek and Warren, 1973; Kaminska et al., 2000; Basirat et al., 2012). In addition, the profound changes across the lifespan observed for the frequency and type of transformations reported by listeners are consistent with age-related changes in linguistic skills and experience (Warren, 1961b; Warren and Warren, 1966).

It has also long been recognized that the VTE shares some

Abbreviations: F0, fundamental frequency; ITD, interaural time difference; PSOLA, Pitch Synchronous Overlap and Add method; VTE, verbal transformation effect

* Corresponding author.
E-mail address: b.roberts@aston.ac.uk (B. Roberts).

http://dx.doi.org/10.1016/j.heares.2017.07.016

common features with the temporal characteristics of the shifts in perceptual organization associated with reversible and multistable visual figures (e.g., Warren and Gregory, 1958; Ditzinger et al., 1997); indeed, recent research using neuroimaging has provided evidence that common functional brain networks underlie perceptual switching in auditory streaming and in verbal transformations (Kashino and Kondo, 2012). However, it is only since the millennium that the relationship between the VTE and cues for auditory stream segregation (Bregman and Campbell, 1971; Bregman, 1990) has been explored in any detail. Pitt and Shoaf (2002) showed that the verbal transformations experienced by listeners are related to the acoustic cues that help bind together the rapidly changing and diverse sounds of speech (see, e.g., Darwin, 2008). In particular, acoustic-phonetic elements that are periodic and have a low-frequency centroid, such as nasals, cohere better with neighbouring voiced vowels than do acoustic-phonetic elements that are aperiodic and have a high-frequency centroid, such as unvoiced fricatives, affricates, and plosives. Hence, extended repetition tends to lead to segregation of unvoiced consonants from the core vocalic parts of the stimulus into one or more streams, with the reported verbal form corresponding to the foreground percept and the unreported segments corresponding to the background. Therefore, the VTE is influenced not only by linguistic processes, but also by the cleaving off and regrouping of acoustic-phonetic elements in a speech stimulus. Since the establishment of this relationship, the VTE has been used as a means of investigating the role of formant transitions and the continuity of the pitch contour in holding together the speech stream (Stachurski, 2012; Stachurski et al., 2015) and of the role of lexical knowledge in the formation of speech streams (Billig et al., 2013).

All of the VTE studies considered so far involved presenting listeners with only one repeating stimulus sequence at a time. Warren and Ackroff (1976) adapted the established methods for studying the VTE to examine the effects of presenting two identical sequences at once (see also Warren, 1996). Fig. 1 illustrates the dichotic stimulus configuration used; the two sequences were distinguished by ear of presentation and played half a cycle out of phase to prevent binaural fusion. One aim of that study was to establish whether or not the same transformations would be heard at the same time on the left and the right; another was to explore the effect of the task demands involved when listeners monitor both sequences at once. It was assumed that simultaneous and identical changes would indicate that a single set of linguistic units was involved in processing both sequences, whereas independent changes would indicate two (or more) sets of functionally separate units. Warren and Ackroff reported that changes occurred at different times at the two ears and that all listeners had periods of time during which they perceived two different forms. For example, a repeating sequence of the word “tress” might be heard at a particular time as “dress” in one ear but as “commence” in the other. That listeners heard independent changes at the two ears was taken to indicate the involvement of more than one set of linguistic units in processing the two sequences, suggesting that everyday listening under cocktail-party conditions (Cherry, 1953) typically involves the processing of speech arising from spatially distinct sources by independent lexical analysers. Note, however,

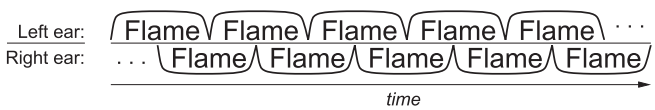

Fig. 1. Schematic illustrating the experimental setup used by Warren and Ackroff (1976) for the dichotic presentation of sequences of repeating stimulus words one half-cycle out of phase, using the example word “flame”.

that the method used did not include an accurate measure of time – the results for each trial consisted only of an ordinal list of transcribed responses flagged with the ear to which the listener was responding. Hence, the *degree* of independence in the responses to the two sequences was not quantified.

Warren and Ackroff (1976) also briefly reported a preliminary study in which five experienced listeners heard three concurrent sequences of the same stimulus; these sequences were each offset by one-third of a cycle – one to the left ear, one to both ears (centre), and one to the right ear. Monitoring all three sequences at once was challenging even for these experienced listeners, but all of them reported independent changes at the different spatial positions. To our knowledge, only one full-scale study has followed up on these observations (Zuck, 1992). That study used a similar configuration of three sequences but listeners were asked to monitor only one or other of them, and so the results did not provide any further insight into the independence of the transformations heard across the three sequences.

Our current understanding of the relationship between the VTE and auditory stream segregation, which is based on studies using single sequences (e.g., Pitt and Shoaf, 2002; Stachurski et al., 2015), suggests that stimulus configurations that increase the possibilities for perceptual regrouping of acoustic-phonetic elements should facilitate the VTE. Furthermore, it also suggests an alternative or additional explanation for the independent verbal transformations for two concurrent sequences of identical stimuli observed by Warren and Ackroff (1976). Specifically, different transformations at the two ears might be a consequence, at least in part, of independent streaming processes at the two ears that lead to independent patterns of segregation and regrouping for the acoustic-phonetic constituents of the stimulus. If this is the case, then using a stimulus configuration that lowers the likelihood of independent changes in the perceptual organization of the two sequences, relative to dichotic presentation, should decrease the independence of the verbal transformations reported for those sequences. Two experiments are reported here.[1] The first tested the hypothesis that allowing two concurrent sequences to interact in the auditory periphery would facilitate the VTE by increasing opportunities for perceptual reorganization but would also decrease the independence of the responses to the two sequences. The second experiment further explored the effects on the VTE of peripheral interaction between sequences and also extended Warren and Ackroff's research on the impact of task demands on listeners' responses to concurrent sequences.

2. Experiment 1

This experiment compared the patterns of verbal transformations for two concurrent sequences under dichotic presentation, in which the two sequences were isolated from one another at the auditory periphery, with those for conditions in which both sequences were present in both ears, in which the acoustic-phonetic constituents of the two sequences were able to interact in the same ear. The two voices were always distinguished using differences in fundamental frequency (F0), either with or without an additional lateralization cue based on ear of presentation or on interaural time difference (ITD) cues; ITD cues help listeners to track the speech of a particular talker across time (Darwin and Hukin, 1999). Differences in F0 provide a salient concurrent segregation cue known to be important in separating speech mixtures (Brokx and Nooteboom, 1982; Bird and Darwin, 1998;

[1] The experiments reported here correspond to reanalysed versions of experiments 1 and 2 in the doctoral thesis of Marcin Stachurski.

Summers et al., 2010, 2016). To our knowledge, this is the first time that the relationship between the VTE and simultaneous grouping has been investigated. By asking listeners to monitor both sequences and recording precisely the times at which they reported each transformation, it was possible to quantify and compare the extent of independence in the responses to the two sequences across conditions.

2.1. Method

2.1.1. Overview

The experiment was conducted using a modified version of the method used by Warren and Ackroff (1976). Listeners were presented with a series of trials, each comprising two concurrent sequences of continuously repeated tokens of a digitally modified natural utterance. Monosyllabic words were used because items with a large number of phonetic segments tend to evoke fewer verbal transformations (see, e.g., Warren, 1961a). The tokens for the concurrent sequences were derived from the same recording of a word, but differed in that one was resynthesized on a low F0 and the other on a high F0. The two versions were played with a half-cycle offset (i.e., half the duration of the token), in order to prevent across-sequence fusion of the identical parts of the tokens (i.e., the aperiodic segments). Each sequence was 3 min long; this choice was based on the observations of Pitt and Shoaf (2002), who found that listeners tended to stop reporting changes after that time owing to fatigue. On each trial, listeners were asked to monitor both sequences and to indicate throughout how they perceived each one of them.

There were three conditions—diotic, ITD, and dichotic—which differed only in the lateralization cue used to distinguish the concurrent low- and high-F0 sequences. In the diotic condition, the only cue available for listeners to segregate the concurrent sequences from one another was the difference in F0. The ITD condition differed from the diotic case only in that opposite ITDs were applied to the two sequences. The value used (±680 μs) is around the maximum natural cue available to distinguish two sources for a typical adult male listener, such that one sequence was heard as strongly left- and the other as strongly right-lateralized. In the dichotic condition, one sequence was presented in the left ear and the other in the right. This condition resembles that used by Warren and Ackroff (1976), except for the difference in F0 (ΔF0) between the two sequences. Note that the dichotic and ITD conditions both involve strong lateralization cues congruent with the ΔF0 cue, whereas the diotic and ITD conditions both involve the physical presence of the two sequences in both ears. This aspect of the design helped to distinguish the contributions of these factors to the transformations reported by listeners. It was hypothesized that the commonality of peripheral stimulation shared by the diotic and ITD conditions would allow perceptual regrouping of acoustic-phonetic elements across sequences, as well as within, which in turn would influence the number and type of VTs heard by listeners, and their degree of independence across sequences.

2.1.2. Listeners

This study was approved and overseen by the Aston University Ethics Committee. All listeners were native speakers of English who reported having no hearing problems. They received either cash or course credits for their participation; most listeners were Psychology students. Twelve listeners (2 males, mean age = 22.2 years, SD = 5.3, range = 18–37) successfully completed the experiment. Two listeners showed little or no tendency to experience VTs. Given that the VTE could not be used as a tool to explore the perceptual regrouping of acoustic-phonetic elements in these listeners, their data were discarded and they were substituted with new listeners.

2.1.3. Stimuli and conditions

The stimuli were derived from recordings of six monosyllabic words—“face”, “flame”, “noise”, “right”, “see”, and “sleep”—spoken by the same male talker (Summers). All these words have been used in a number of previous studies of the VTE, and were selected on the basis that recordings of them produced a variety of verbal transformations during pilot testing. To help achieve the target duration of 550 ms per token, the talker produced several examples of each utterance with the assistance of an on-screen metronome to pace speech production. Mono recordings with 16-bit resolution and a sampling rate of 22.05 kHz were made using a Sennheiser MD 918U-T microphone (Hannover, Germany) and Santa Cruz sound card (Turtle Beach, Valhalla, New York) in a single-walled sound-attenuating chamber (Industrial Acoustics 401A, Winchester, UK) housed within a quiet room. From a single recording session, instances were chosen that were clearly articulated, on a fairly flat F0 contour, and close to the target duration. Using Adobe Audition software, the target duration was matched exactly using small manual adjustments to the stimuli. This involved manipulations such as small changes to the duration of fricative bursts, by copying in or deleting a sample of steady fricative noise, or to the duration of intervals corresponding to closures in stop consonant production. For each stimulus, 5-ms linear ramps were applied at onset and offset.

The duration-adjusted and ramped version of each recording was then processed using PRAAT software (Boersma and Weenink, 2009). Each recording was monotonized and set to one of two constant F0 values using the Pitch Synchronous Overlap and Add method (PSOLA; Moulines and Charpentier, 1990). This time-domain manipulation identifies individual glottal pulses in the voiced segments and adjusts the time intervals between them, enabling each recording to be resynthesized on different F0s without associated changes in vocal-tract filtering. For one version, F0 = 100 Hz (low F0, male range); for the other, F0 = 178 Hz (high F0, female range), which corresponds to a ΔF0 of 10 semitones. A difference of this magnitude is known to provide a strong cue for the stream segregation of concurrent speech (e.g., Brokx and Nooteboom, 1982; Bird and Darwin, 1998); this difference also makes the two voices easily distinguishable and so minimizes the likelihood of confusing responses to the two sequences, even in the absence of supporting lateralization cues.

A terminal silence of 1 ms was added to each resynthesized stimulus. For the ITD condition, stereo versions of each stimulus were created using MITSYN (Henke, 2005) by introducing a delay of 15 samples on the appropriate channel (±680 μs for the sample rate of 22.05 kHz), such that the lateralization of the low- and high-F0 sequences was congruent with the dichotic condition. The delay for the ITD condition was accommodated within the terminal silence, so that the tokens used in all three conditions were precisely 551 ms long, corresponding to 327 repetitions per sequence in 3 min. In line with previous studies (e.g., Warren and Ackroff, 1976; Pitt and Shoaf, 2002), the interval between sequentially presented tokens was kept small to facilitate maximum re-segmentation and perceptual regrouping of acoustic-phonetic elements within the stimuli presented. An additional word, “train”, was recorded and transformed in the same way as described above for the experimental stimuli. This word was used in the practice trial for this experiment (see below).

The presentation software was custom written in VB.Net (Microsoft Visual Studio 2005) and run on a PC. Two looping stereo buffers, starting and finishing in synchrony but half a cycle out of phase, were used to create the repeating high- and low-F0 sequences; the outputs of these buffers were summed to create the concurrent sequences presented to listeners. All stimuli were presented at ~75 dB SPL (sound pressure level) using a Santa Cruz

sound card (Turtle Beach, Valhalla, New York) and Sennheiser HD480-13II headphones (Hannover, Germany); outputs were calibrated using a sound-level meter (Brüel and Kjaer, Type 2209, Nærum, Denmark) coupled to the earphones by an artificial ear (Type 4153). Each 3-min presentation was faded in and out using linear ramps corresponding to one complete stimulus token (i.e., 551 ms). Fading in/out is common practice in the VTE literature; e.g., Warren (1961b) increased the volume of his sequences at onset from zero to full in 1 s. Here, this tactic also disguised the fact that one sequence began and ended in the middle of the token, owing to the half-cycle offset. In the ITD and dichotic conditions, the standard and offset sequences were left and right lateralized, respectively. Stimulus allocation was balanced such that half the listeners always received low-F0 standard and high-F0 offset sequences; the opposite configuration was used for the other half.

2.1.4. Task and instructions

Listeners were told that they would hear a series of short verbal utterances spoken by two voices, one with a low pitch and one with a high pitch. They were also informed that the two voices would sometimes be heard as coming from different spatial locations. Listeners were asked to speak into the microphone (Sennheiser MD 918U-T) as soon as they were able to identify what each voice appeared to be saying, irrespective of whether it was perceived as a word, phrase, pseudo-word, or non-word. Listeners used a keyboard to indicate whether their response was to the high or low voice, using the 'up' or 'down' arrow key, respectively. For example, if a listener heard "book" spoken on the high F0, they were asked to press the 'up' arrow key (thereby displaying 'HIGH' on the screen), say the word "book" into the microphone, and then release the key. Thereafter, they were asked to continue monitoring both voices and to report any perceived changes to the verbal form of either voice, using the microphone and keyboard. Listeners were told that a change (VT) might involve the current percept either changing to a new form or reverting back to a previous form, both of which they must report. Note that, on any given trial, a listener could experience an indefinite number of VTs as long as at least two forms were reported. It was emphasized that a non-response was as important as a response so that listeners did not feel pressed to make a report if they did not hear a change in verbal form. Listeners were assured that there was no right or wrong answer to how the voices should be perceived and that in some cases they may hear few if any changes during a trial. Although this task allowed for continuous monitoring of both sequences, it is acknowledged that on any occasion when listeners heard near-simultaneous transformations on both voices, task constraints prevented them from responding to both at the same time.

2.1.5. Recording and transcription of responses

On any 3-min trial, stimulus presentation over the headphones, recording of verbal responses over the microphone, and recording of key presses, were time-locked; all started simultaneously. Listeners' verbal responses for each 3-min presentation were saved as 8-bit audio (.wav) files at a sampling rate of 11.025 kHz. The key presses indicating the F0 of the sequence on which the change occurred were stored as text files, where each response entry comprised precision timings of when the key was pressed and released, and the identity of the key pressed (i.e., 'up' or 'down' arrow). It was therefore possible accurately to assign verbal responses to individual key presses. For each text file entry, the experimenter searched the corresponding audio file for a verbal response occurring in the interval between the depression and release of the key and added a transcription; the identity of the key indicated the voice to which the response was made. The initial response to each voice was allocated a nominal time of 0 s. For subsequent responses to each voice, the time (since the start of the trial) at which the key was pressed was taken as the moment of perceptual change to that reported verbal form. On any trial for which there were no subsequent responses, the time to the first VT was assigned a nominal value of 180 s. On occasions when a listener accidentally reported the same verbal form twice in succession on the same voice, the second response was discounted.

2.1.6. Procedure

Listeners attended three testing sessions, each corresponding to one of the three lateralization conditions. Each session lasted ~30 min and took place on a different day. Listeners read the instructions at the start of the first session, after which the experimenter reiterated what was involved and answered any questions arising. The main part of the test session comprised six trials from the main experiment – each using a 3-min presentation consisting of two repeating versions of the same stimulus word, one on the low and the other on the high F0. Before the experimental trials began, listeners completed a practice trial comprising a 1-min presentation of the word "train", processed in the same way as for the ITD condition. Listeners were given a 1-min break between each trial. Subsequent test sessions comprised a brief recap of the instructions followed by the six experimental trials. Within each session, trials using particular words were presented in random order. The order of the three sessions (conditions) was fully counterbalanced across participants, requiring six people to complete a full set; the experiment comprised two sets of listeners.

2.1.7. Data analysis

For each combination of lateralization cue (none, 680-μs ITD, or ear of presentation), F0 (low or high) and stimulus word, four measures were calculated from the responses on each trial – the number of VTs, the number of forms (defined as cases where a given response had not occurred before on that trial – *including* the initial response, which is not a transformation), the time from the start of the trial to the first VT, and the percentage of time during the trial (the 'dwell time') for which the initial form was reported. The first three measures have been used widely in studies of the VTE since the pioneering research by Warren; to our knowledge, the fourth measure was first introduced by Ditzinger et al. (1997). We also devised three customized indices to explore the relatedness of responses across the concurrent sequences. To compute them, the responses were first divided into two classes – those where the preceding and/or following response was to the sequence on the other F0 (type 1) and the rest (type 2), for which the adjacent responses were to the same sequence as the current response. The initial responses to the two sequences were excluded as they are not transformations. The *dependency index* corresponds to the proportion of type 1 responses for which an adjacent response on the other F0 was of the same form as the current response and the *intervening-responses index* corresponds to the proportion of the total responses that were type 2. The *temporal-overlap index* corresponds to the proportion of time after the first VT during which the responses (irrespective of type) to the two sequences were of the same form as one another, including returns to the original form. Fig. 2 illustrates how the four VTE measures and the three indices were computed. The following two paragraphs explain in more detail these indices, their computation, and the reasoning behind them.

All three indices were computed only for those trials in which at least one VT occurred $\leq$150 s from stimulus onset; this approach was made possible by collapsing across stimulus word. By ensuring that there was at least 30 s from the first VT to the end of the trial, we ensured that the estimates obtained were meaningful (e.g., avoiding undefined values for nil responses) and more stable. The

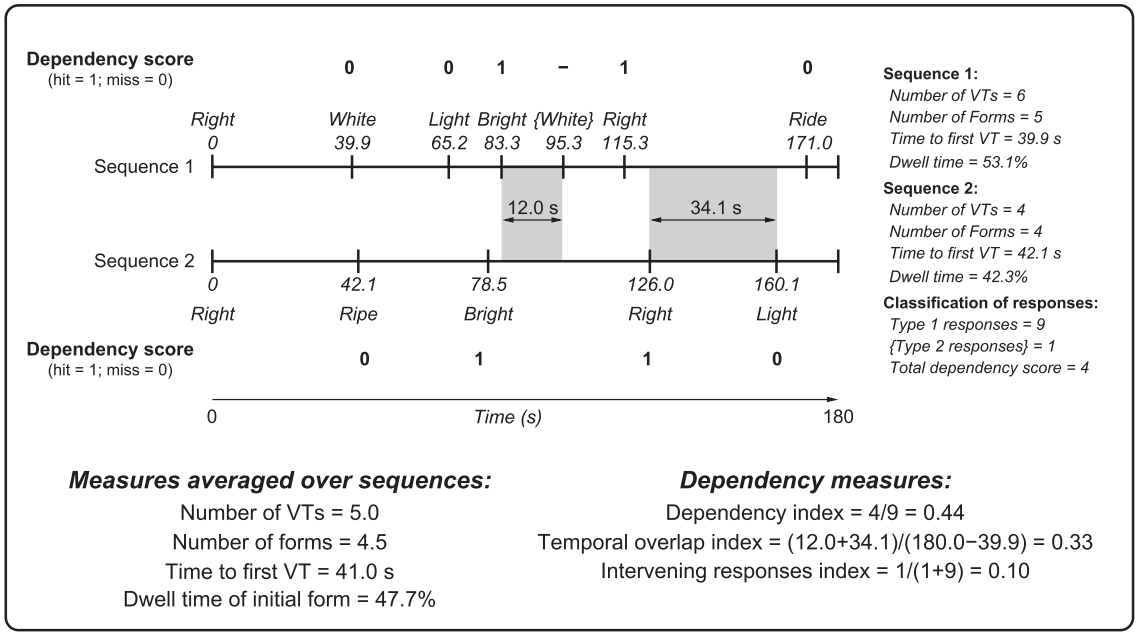

Fig. 2. Schematic illustrating how the four VTE measures were computed for each sequence and how the three indices comparing responses across two concurrent sequences were computed, using the example stimulus word "right" (see text for full details). Responses were classified as type 1 when the preceding and/or following response was on the other F0; the rest were classified as type 2. The dependency index is defined as the proportion of type 1 responses for which an adjacent response on the other F0 was of the same form as the current response (i.e., total hits divided by total number of type 1 responses). The temporal-overlap index is defined as the proportion of time after the first VT for which the same form was reported for both sequences; the time periods (after the first VT) during which this was the case are shown in this figure as shaded areas. The intervening-responses index is defined as the proportion of the total responses that were type 2.

dependency index was computed from the set of dependency scores, one for each type 1 response. Type 2 responses were excluded from the calculations because they were cases where the form reported on a given sequence had already changed before the next response on the other sequence. For a given type 1 response (R), the score was set to 1 (hit) if either the preceding or subsequent response on the other sequence was the same; otherwise, the score for R was set to 0 (miss). Given that each across-sequence relationship involved two responses, the total dependency score (if > 0) for any pair of sequences was always an even number. The scores were summed across the two sequences and the total number of hits was divided by the total number of type 1 responses to obtain the dependency index for that sequence pair. This method takes into account that every across-sequence relationship was scored as 2, so that possible values for the dependency index ranged from 0 (VTs were fully independent/unrelated across sequences) to 1 (fully dependent/related). The focus here was on first-order dependencies; there was no *a priori* reason to expect higher-order dependencies.

The temporal-overlap index was important for interpreting any observed differences in the dependency index between lateralization conditions. This is because, in principle, there is a circumstance in which a fall in the dependency index could occur without a change in the degree to which responses to the two sequences are related. Specifically, this is where there is an increase in the proportion of anti-correlated responses to the two sequences (i.e., a relationship where a response of a given form to one sequence decreases the likelihood of the same response to the other sequence). Such a change would, however, also result in a fall in the temporal-overlap index. Hence, a substantial fall in the dependency index in the context of relative constancy in the extent of temporal overlap was interpreted as indicating that the responses to the concurrent sequences were indeed more independent of one another, and was not simply an artefact of a change from more correlated towards more anti-correlated responses. The intervening-responses index was included to indicate the extent (if any) to which the proportion of type 2 responses varied across conditions.

The results were analysed in SPSS (IBM, version 21) using within-subjects analysis of variance (ANOVA). The Greenhouse-Geisser correction was applied to the degrees of freedom for all terms where Mauchly's test indicated a significant departure from sphericity. Pairwise comparisons were performed using Fisher's least significant difference (LSD) test, with the restriction that the factor being explored must be associated with a significant effect in the ANOVA (the restricted LSD test; Snedecor and Cochran, 1967; Keppel, 1991). The measure of effect size reported here is partial eta squared (η^2_p).

2.2. Results and discussion

2.2.1. Effects of lateralization condition, F0 value, and stimulus word on VTE responses to each sequence

The panels of Fig. 3 summarize the results for the four VTE measures used—number of VTs and forms (per 3-min trial), time to the first VT, and dwell time of the initial form—when averaged across stimulus words. A three-way ANOVA (3 lateralizations, 2 F0 values, 6 words) was performed on each measure and the statistical outcomes are presented in Table 1; significant pairwise comparisons between conditions are also shown in Fig. 3. Given the skewed distribution of times to the first VT and the substantial overall proportion of nil responses (25.0%; all assigned a nominal value of 180 s), which cannot be corrected using a simple transformation, the median times to the first VT are also shown.

According to the main hypothesis, presenting both sequences in both ears (diotic and ITD conditions) should *increase* the number of VTs and forms and *decrease* the time to the first VT and the dwell time of the initial form, relative to the dichotic condition. There was a significant main effect of lateralization condition for number of forms ($p = 0.003$), time to first VT ($p < 0.001$), and dwell time of the

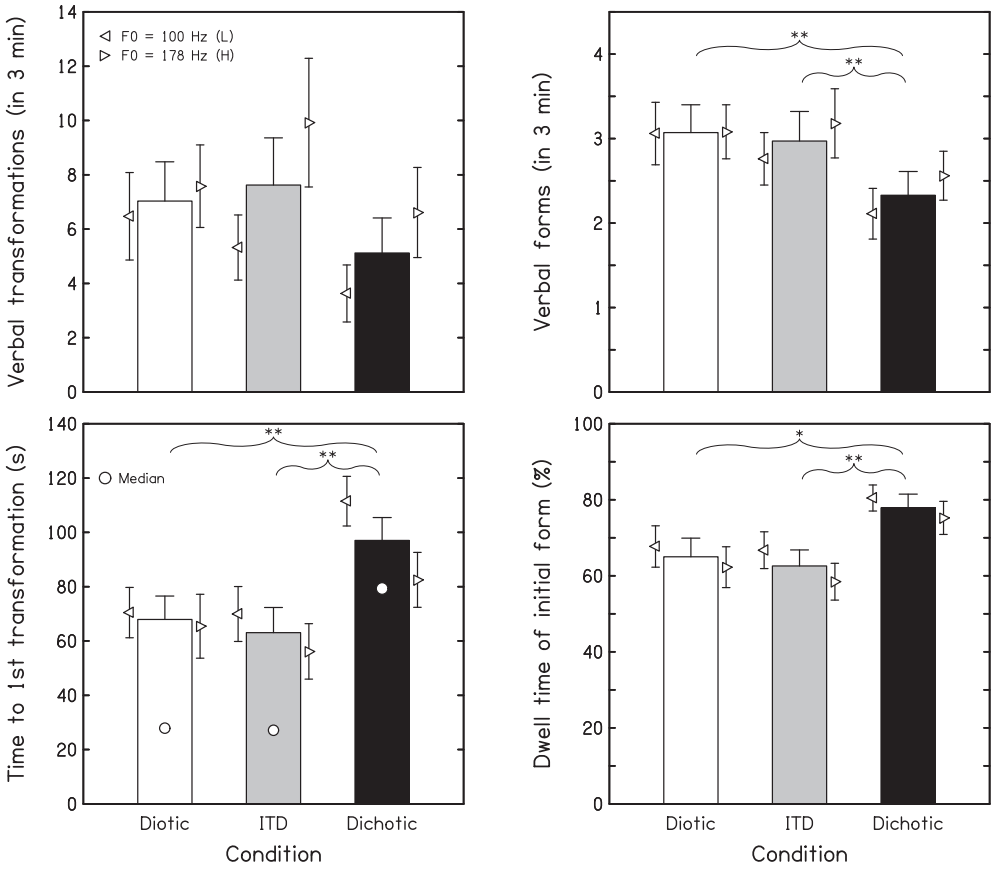

Fig. 3. Results for experiment 1 – Effects of differences in lateralization condition between two concurrent sequences of repeating stimulus words on the verbal transformation effect. These sequences were always distinguished by a difference in fundamental frequency. The results are shown in four panels: the number of verbal transformations (top left) and verbal forms (top right) reported in 3 min, the time to the first verbal transformation reported (bottom left), and the dwell time of the initial form (bottom right). Means (n = 12) and inter-subject standard errors are shown for each condition; significant pairwise comparisons are indicated using brackets and significance levels using asterisks (* = <5%, ** = <1%). For each of the four measures used, the results for the low-F0 sequence (L) and the high-F0 sequence (H) are shown separately on the left-hand and right-hand sides of the overall mean using leftward- and rightward-pointing triangles, respectively. The bottom left panel also shows the median times to the first verbal transformation (embedded open circles).

initial form ($p = 0.008$). For all these measures, the effect was in the predicted direction – pairwise comparisons showed that dichotic presentation resulted in significantly fewer forms, later first VTs, and a greater dwell time of the initial form than when both sequences were presented to both ears (range: $p = 0.039$ – $p = 0.001$); the results for the diotic and ITD conditions did not differ from one another. Consistent with our interpretation of the effect of lateralization condition on the VTE, the difference between the medians for the dichotic condition and the other cases was even more pronounced than that for the means. The direction of the effect of lateralization on the number of VTs was in accord with the main hypothesis, but was not significant ($p = 0.139$). This result is in line with earlier observations that there is often more variability across stimuli and listeners for the number of VTs reported than for the number of forms (e.g., Lass et al., 1973; Warren, 1996; Stachurski et al., 2015).

There was no *a priori* hypothesis about the effect of F0 value; the ΔF0 between the two sequences was intended simply to provide a concurrent segregation cue that would be effective even in the absence of lateralization cues. Across all four VTE measures, there was a significant main effect of F0 value (range: $p = 0.043$ – $p = 0.003$). Responses to the high-F0 sequence were typically associated with more VTs, more forms, a shorter time to the first VT, and a reduced dwell time of the initial form. The reason for this outcome is unclear, but may be related to the need for listeners to divide their attention between the two voices. There was also a significant main effect of stimulus word across all four measures ($p \leq 0.003$); this outcome is unsurprising because the likelihood of perceptual reorganization when a particular speech stimulus is presented under extended repetition depends on the acoustic diversity of the phonetic segments, such as differences in source excitation and spectral centroid, and the acoustic cues binding them together into a single stream, such as formant transitions and the F0 contour (Cole and Scott, 1973; Dorman et al., 1975; Darwin and Bethell-Fox, 1977; Stachurski et al., 2015; see also David et al., 2017). For example, a voiceless fricative is typically bound less strongly than a voiced nasal or approximant to the core vowel of a syllable. Appendix A shows the results separately for each stimulus word when collapsed across lateralization condition and F0 value. Note that, as might be expected, stimulus words associated with a greater number of VTs and forms were typically associated with shorter times to the first VT and dwell times of the initial form.

None of the interaction terms were significant for the measures number of VTs, time to first VT, and dwell time of the initial form. The only significant interaction was between lateralization and word for the number of forms ($p = 0.049$); this mainly arose because a particularly high number of forms was reported for the words “face” and “sleep” in the diotic and ITD conditions compared with the dichotic case. This pattern suggests that the impact on forms of whether or not the two sequences can interact within the same ear depends to some extent on the acoustic properties of

Table 1
Results for experiment 1. Summary of the three-way repeated-measures ANOVAs for the four response measures. Where necessary, the Greenhouse-Geisser correction was applied to the degrees of freedom. All significant terms are shown in bold. Where there is a significant main effect of lateralization, pairwise comparisons within that factor are also shown.

Part (a): Results for VTs				
Source	df	F	p	η^2_p
Lateralization (L)	(2, 22)	2.162	0.139	–
F0 value (F0)	**(1, 11)**	**14.262**	**0.003**	**0.565**
Word (W)	**(5, 55)**	**4.150**	**0.003**	**0.274**
L x F0	(2, 22)	2.257	0.128	–
L x W	(4.417, 48.589)	0.846	0.512	–
F0 x W	(2.229, 24.524)	0.608	0.570	–
L x F0 x W	(4.219, 46.414)	2.220	0.078	–
Part (b): Results for forms				
Source	df	F	p	η^2_p
Lateralization (L)	**(2, 22)**	**7.402**	**0.003**	**0.402**
F0 value (F0)	**(1, 11)**	**5.452**	**0.040**	**0.331**
Word (W)	**(5, 55)**	**4.581**	**0.001**	**0.294**
L x F0	(2, 22)	1.310	0.290	–
L x W	**(3.923, 43.157)**	**2.615**	**0.049**	**0.192**
F0 x W	(5, 55)	1.212	0.316	–
L x F0 x W	(4.198, 46.182)	2.344	0.066	–
Pairwise comparisons within factor L		t(11)		p
Dichotic vs. diotic		**3.747**		**0.003**
Dichotic vs. ITD		**4.196**		**0.001**
ITD vs. diotic		0.373		0.716
Part (c): Results for time to first VT				
Source	df	F	p	η^2_p
Lateralization (L)	**(2, 22)**	**13.509**	**<0.001**	**0.551**
F0 value (F0)	**(1, 11)**	**5.971**	**0.033**	**0.352**
Word (W)	**(5, 55)**	**4.530**	**0.002**	**0.292**
L x F0	(2, 22)	1.627	0.219	–
L x W	(10, 110)	0.662	0.757	–
F0 x W	(2.983, 32.813)	0.281	0.838	–
L x F0 x W	(10, 110)	1.250	0.268	–
Pairwise comparisons within factor L		t(11)		p
Dichotic vs. diotic		**3.839**		**0.003**
Dichotic vs. ITD		**4.370**		**0.001**
ITD vs. diotic		0.865		0.405
Part (d): Results for dwell time of initial form				
Source	df	F	p	η^2_p
Lateralization (L)	**(2, 22)**	**6.074**	**0.008**	**0.356**
F0 value (F0)	**(1, 11)**	**5.212**	**0.043**	**0.321**
Word (W)	**(5, 55)**	**4.168**	**0.003**	**0.275**
L x F0	(2, 22)	0.193	0.826	–
L x W	(10, 110)	1.650	0.102	–
F0 x W	(5, 55)	1.507	0.203	–
L x F0 x W	(10, 110)	0.577	0.830	–
Pairwise comparisons within factor L		t(11)		p
Dichotic vs. diotic		**2.344**		**0.039**
Dichotic vs. ITD		**3.265**		**0.008**
ITD vs. diotic		0.628		0.543

individual stimulus words.

Overall, the results support the hypothesis that the physical presence of the two sequences in both ears (diotic and ITD conditions) provided extra opportunities for the regrouping of acoustic-phonetic elements compared with dichotic presentation. Note, however, that the dual-monitoring method used here required listeners to divide their attention between the two voices and so is likely to have underestimated the transformations associated with each sequence (cf. Warren and Ackroff, 1976). This issue is considered further in experiment 2.

2.2.2. Comparison of VTE responses across sequences within concurrent pairs

The panels of Fig. 4 summarize the results for the three indices (dependency, temporal overlap, intervening responses), which together describe the relationship between the set of responses to one sequence and to the other. Computing these indices involved comparing responses across corresponding low- and high-F0 sequences and collapsing across stimulus word, and so neither F0 value nor word were factors in the statistical analysis. 33.8% of trials did not meet the inclusion criterion (first VT $\leq$ 150 s from stimulus onset) and so were excluded from the computation. To assess the effect of lateralization condition, a one-way ANOVA was performed on each measure and the statistical outcomes are presented in Table 2; significant effects are also shown in Fig. 4.

The results of Warren and Ackroff (1976) suggest that the VTs reported for the two sequences should be relatively independent of each other in the dichotic condition. It was predicted that presenting both sequences in both ears (diotic and ITD conditions), allowing within-ear interactions between them, would *decrease* the extent of this independence. There was a significant effect of lateralization condition for the dependency index ($p < 0.05$), but not for the temporal-overlap ($p = 0.376$) or intervening-response ($p = 0.926$) indices. As predicted, pairwise comparisons with the dichotic condition showed that responses to the two sequences gave rise to a higher dependency index in the diotic ($p = 0.047$) and ITD ($p = 0.035$) conditions, for which the value of this index was nearly double that for the dichotic case. The results for the diotic and ITD conditions did not differ from one another. Given that there were no significant differences in the temporal-overlap index across lateralization conditions, it seems likely that the significantly lower dependency index found for the dichotic condition was not a spurious consequence of a greater anti-correlation in the responses to the two sequences but instead shows a genuinely greater independence of those responses. Also, the high degree of similarity across conditions in the proportion of type 2 responses rules out any possibility that the dependency index might have been affected in some unanticipated way by the proportion of intervening responses. Notwithstanding the significant difference between the dichotic and the other lateralization conditions, it is noteworthy that most VTs were fairly independent across sequences in a concurrent pair even when both sequences were physically present in both ears (dependency index was typically <0.5). In this regard, note that for a finite set of forms, the dependency index would be non-zero even when there is no underlying connection between the responses to the two sequences.

3. Experiment 2

Warren and Ackroff's (1976) study included a comparison of responses to dichotic presentation of identical concurrent sequences when both sequences were monitored at once with when only one or other ear was monitored. There were significantly fewer responses for the report-both condition than for the combination of report-left and report-right conditions. The experiment reported here used diotic presentation to examine how verbal transformations for a sequence were influenced by the increased opportunities for perceptual regrouping offered by the presence of another sequence in the same ear and by the impact of the task demands of monitoring both at once. This was achieved by comparing responses in three circumstances – one sequence presented and reported, two sequences presented but only one reported, and two sequences presented and both reported. It was anticipated that presenting both sequences concurrently but asking listeners to monitor only one or other of them would best reveal the extent to which verbal transformations were increased by the

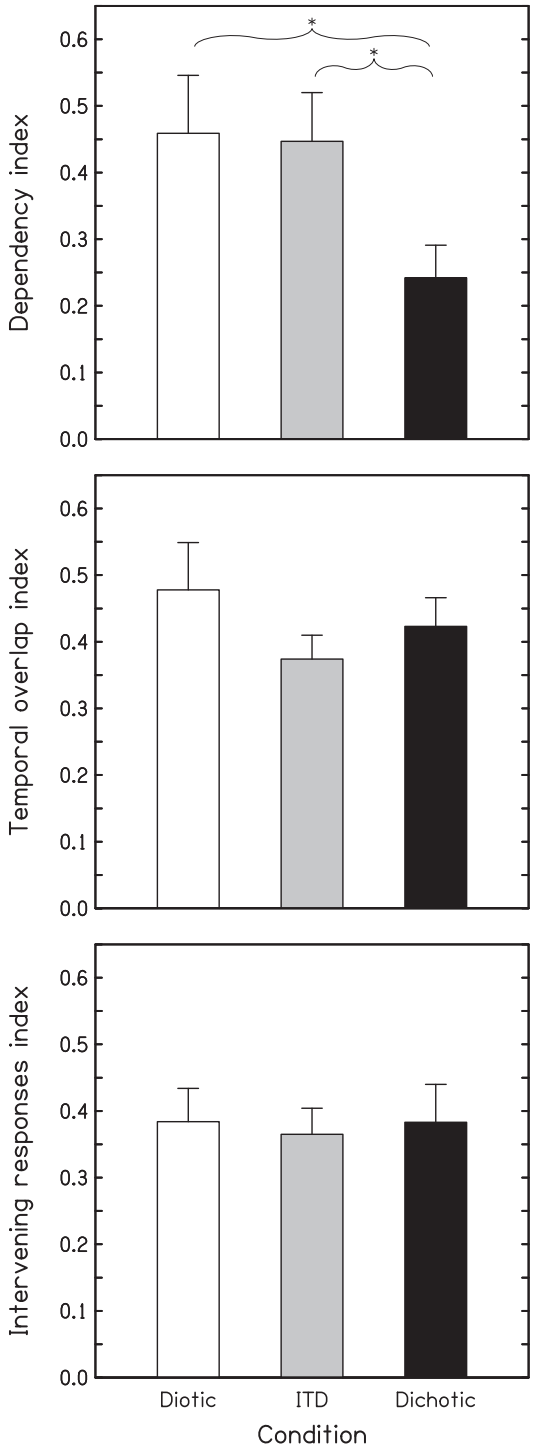

Fig. 4. Results for experiment 1 – Effects of differences in lateralization condition on the relatedness of the forms reported across the two sequences. The results are shown in three panels for the three indices: the dependency index (top), the temporal-overlap index (middle), and the intervening-responses index (bottom). Means (n = 12) and inter-subject standard errors are shown for each condition; significant pairwise comparisons are indicated using brackets and significance levels using asterisks (* = <5%, ** = <1%).

possibility of within-ear interactions between the two sequences. The two sequences were always distinguished only by a difference in F0. Note that an outcome indicating a high cost of dual monitoring in the absence of lateralization cues would favour an account in which the task demands arise from the need for listeners to divide their attention between two streams, rather than specifically between two spatial positions.

Table 2
Results for experiment 1. Summary of the one-way repeated-measures ANOVAs for the three indices, exploring the effects of lateralization condition. Where necessary, the Greenhouse-Geisser correction was applied to the degrees of freedom. All significant terms are shown in bold. Where there is a significant effect of lateralization, pairwise comparisons within that factor are also shown.

Part (a): Effects of lateralization condition (L) on the three indices				
Index	df	F	p	η^2_p
Dependency	**(2, 22)**	**3.456**	**0.0495**	**0.239**
Temporal overlap	(2, 22)	1.022	0.376	-
Intervening responses	(2, 22)	0.077	0.926	-

Part (b): Pairwise comparisons within factor L for the dependency index		
Condition pair	t(11)	p
Dichotic vs. diotic	**2.241**	**0.047**
Dichotic vs. ITD	**2.403**	**0.035**
ITD vs. diotic	0.128	0.900

3.1. Method

Except where stated, the same method was used as for experiment 1. Fifteen listeners (3 males, mean age = 22.5 years, SD = 3.0, range = 18–30) successfully completed the experiment; none of them took part in experiment 1. Test sequences were generated using the same set of stimulus words as for experiment 1; all sequences were presented diotically (i.e., without lateralization cues). There were three conditions, which differed only in terms of the number of sequences presented (S1 or S2) and the number of sequences to which listeners had to attend and respond (R1 or R2). In Condition S1R1, each 3-min presentation consisted of an on-going repetition of a single stimulus word, either on the low or the high F0. In Condition S2R1, each 3-min presentation consisted of two concurrent sequences played half a cycle out of phase, one on the high and the other on the low F0, and listeners were required to monitor and respond only to one or other of them (as instructed). Condition S2R2 differed only in that listeners were required to monitor and respond to both sequences concurrently. Hence, Condition S2R2 was identical to the diotic condition in experiment 1. The same response protocol was used in all three conditions – i.e., listeners were asked to respond by pressing the key corresponding to the pitch of the appropriate voice, speaking the verbal form that they heard on that voice into the microphone, and releasing the key.

Each listener completed the full set of conditions in five sessions by completing one test block per session. These test blocks, identified by condition label and the sequence F0 (or F0s) attended, were as follows: 1 = S1R1(L), 2 = S1R1(H), 3 = S2R1(L), 4 = S2R1(H), 5 = S2R2(L+H). Test blocks were counterbalanced across listeners using a simple five-cycle rotation, such that the set order for the first listener was 1-2-3-4-5, the order for the second was 2-3-4-5-1, and so forth. The first and subsequent sessions all included a 1-min practice trial, using the stimulus word “train”. The stimulus and response characteristics for this trial were configured in the same way as for the experimental trials in that block. VTs, verbal forms, the time to the first VT, and the dwell time of the initial form were transcribed and calculated as before.

3.2. Results and discussion

The results for the four VTE measures are summarized in Fig. 5. A three-way ANOVA (3 conditions, 2 F0 values, 6 words) was

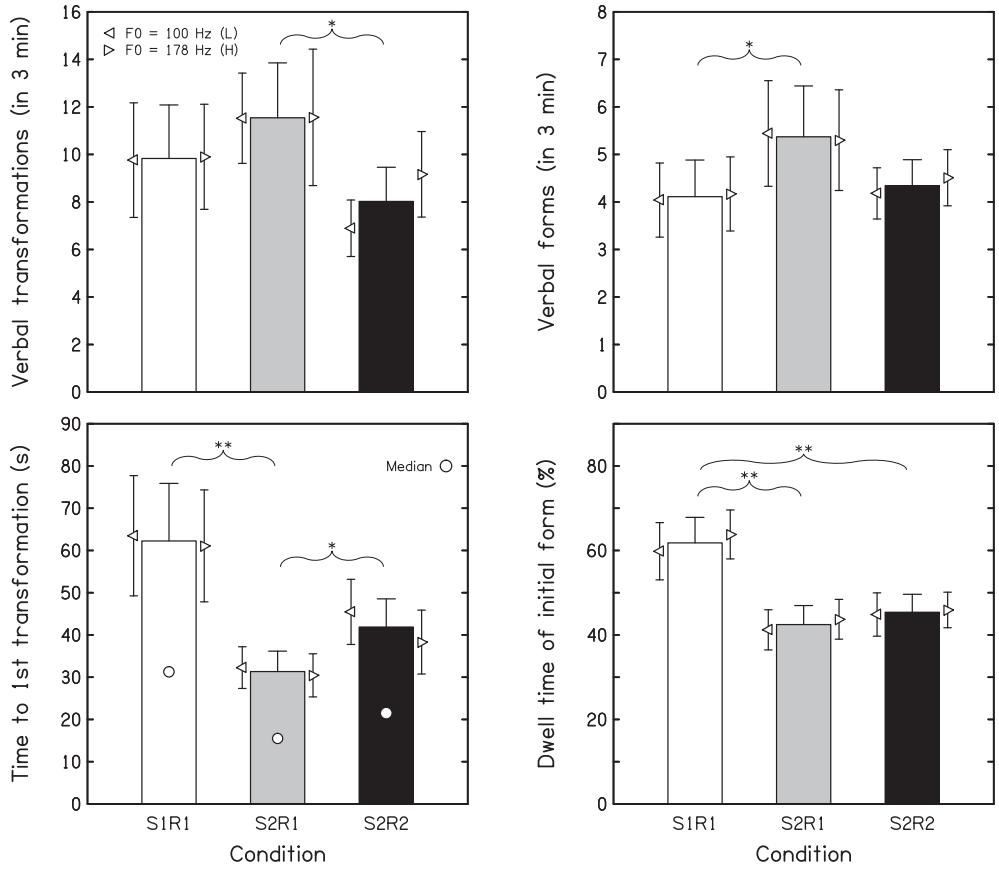

Fig. 5. Results for experiment 2 – Effects of the number of sequences presented concurrently (S = 1 or 2) and the number of sequences to which the listeners must respond (R = 1 or 2) on the verbal transformation effect. All stimuli were presented diotically; when there were two sequences, they were distinguished by a difference in fundamental frequency. The results are shown in four panels: the number of verbal transformations (top left) and verbal forms (top right) reported in 3 min, the time to the first verbal transformation reported (bottom left), and the dwell time of the initial form (bottom right). Means (n = 15) and inter-subject standard errors are shown for each condition; significant pairwise comparisons are indicated using brackets and significance levels using asterisks (* = <5%, ** = <1%). For each of the four measures used, the results for the low-F0 sequence (L) and the high-F0 sequence (H) are shown separately on the left-hand and right-hand sides of the overall mean using leftward- and rightward-pointing triangles, respectively. The bottom left panel also shows the median times to the first verbal transformation (embedded open circles).

performed on each measure and the statistical outcomes are presented in Table 3; significant pairwise comparisons between conditions are also shown in Fig. 5. The listeners in this experiment (none of whom took part in experiment 1) showed a greater overall tendency to produce VTs in Condition S2R2 than for its direct counterpart in experiment 1 (the diotic condition). Hence, the proportion of nil responses for the measure time to first VT was less than half that observed in experiment 1. Nonetheless, this proportion remained considerable (10.9%), and so as before the median times to the first VT are also shown.

The results of experiment 1 are in accord with the idea that presenting two sequences of repeating words in both ears facilitates perceptual regrouping, which for both sequences tends to increase the number of VTs and forms reported and to decrease the time to the first VT and the dwell time of the initial form. Here, it was predicted that this pattern would be most evident for the comparison between Conditions S1R1 and S2R1, because in the latter case the presence of the unmonitored sequence should increase opportunities for VTs to be heard on the monitored sequence. It was also predicted that the divided attention necessary to monitor both sequences concurrently (Condition S2R2) would inevitably tend to reduce the number of VTs reported, with attendant changes in the other VTE measures, despite an assumption of no change in the underlying tendency to experience VTs.

There were significant main effects of condition for all four measures – number of VTs ($p = 0.048$), number of forms ($p = 0.043$), time to first VT ($p = 0.018$), and dwell time of the initial form ($p < 0.001$). The origin of these effects was explored using pairwise comparisons, which indicated patterns consistent with the predictions. When listeners monitored only one sequence (S1R1 vs. S2R1), the presence of a concurrent but unattended sequence resulted in significantly more forms ($p = 0.012$), a shorter time to the first VT ($p = 0.008$), and a reduced dwell time of the initial form ($p < 0.001$); the outcome for number of VTs was in the predicted direction but was not significant. When listeners received both sequences at once (S2R1 vs. S2R2), monitoring only one or other of them resulted in significantly more reports of VTs ($p = 0.013$) and a shorter time to the first VT ($p = 0.036$), compared with when attention was divided across the two sequences. This outcome suggests that the task demands of monitoring both sequences arise mainly from the need to divide attention across streams (high vs. low F0) rather than specifically across spatial positions (left vs. right). Note that an account for the greater number of VTs reported in S2R1 than S2R2 based on confusions with VTs heard on the unattended sequence seems implausible given the large difference in F0 distinguishing the two voices. The result for forms was in the expected direction but was not significant; the mean dwell time of the initial form was almost identical for these two conditions. Consistent with these outcomes, the only significant pairwise comparison for S1R1 vs. S2R2 was the shorter

Table 3
Results for experiment 2. Summary of the three-way repeated-measures ANOVAs for the four response measures. Where necessary, the Greenhouse-Geisser correction was applied to the degrees of freedom. All significant terms are shown in bold. Where there is a significant main effect of condition, pairwise comparisons within that factor are also shown.

Part (a): Results for VTs				
Source	df	F	p	η^2_p
Condition (C)	**(2, 28)**	**3.388**	**0.048**	**0.195**
F0 value (F0)	(1, 14)	0.819	0.381	–
Word (W)	(1.598, 22.373)	3.541	0.055	–
C x F0	(1.300, 18.206)	1.474	0.248	–
C x W	(2.211, 30.949)	1.538	0.230	–
F0 x W	(5, 70)	0.136	0.983	–
C x F0 x W	(2.362, 33.062)	0.942	0.413	–
Pairwise comparisons within factor C		t(14)		p
S1R1 vs. S2R1		1.289		0.218
S1R1 vs. S2R2		1.217		0.244
S2R1 vs. S2R2		**2.858**		**0.013**
Part (b): Results for forms				
Source	df	F	p	η^2_p
Condition (C)	**(1.321, 18.493)**	**4.314**	**0.043**	**0.236**
F0 value (F0)	(1, 14)	0.334	0.573	–
Word (W)	**(2.434, 34.082)**	**6.082**	**0.003**	**0.303**
C x F0	(2, 28)	0.756	0.479	–
C x W	(4.795, 67.136)	2.337	0.054	–
F0 x W	(2.625, 36.748)	0.776	0.499	–
C x F0 x W	(2.954, 41.357)	1.188	0.326	–
Pairwise comparisons within factor C		t(14)		p
S1R1 vs. S2R1		**2.880**		**0.012**
S1R1 vs. S2R2		0.802		0.436
S2R1 vs. S2R2		1.743		0.103
Part (c): Results for time to first VT				
Source	df	F	p	η^2_p
Condition (C)	**(1.274, 17.837)**	**6.143**	**0.018**	**0.305**
F0 value (F0)	(1, 14)	1.179	0.296	–
Word (W)	**(2.394, 33.516)**	**6.554**	**0.002**	**0.319**
C x F0	(2, 28)	0.406	0.670	–
C x W	(4.514, 63.193)	1.714	0.151	–
F0 x W	(5, 70)	2.156	0.069	–
C x F0 x W	(4.102, 57.423)	2.077	0.094	–
Pairwise comparisons within factor C		t(14)		p
S1R1 vs. S2R1		**3.063**		**0.008**
S1R1 vs. S2R2		1.870		0.082
S2R1 vs. S2R2		**2.322**		**0.036**
Part (d): Results for dwell time of initial form				
Source	df	F	p	η^2_p
Condition (C)	**(2, 28)**	**18.094**	**<0.001**	**0.564**
F0 value (F0)	(1, 14)	0.941	0.349	–
Word (W)	**(5, 70)**	**4.908**	**0.001**	**0.260**
C x F0	(2, 28)	0.238	0.790	–
C x W	(10, 140)	1.630	0.104	–
F0 x W	**(5, 70)**	**5.479**	**<0.001**	**0.281**
C x F0 x W	**(10, 140)**	**2.842**	**0.003**	**0.169**
Pairwise comparisons within factor C		t(14)		p
S1R1 vs. S2R1		**5.772**		**<0.001**
S1R1 vs. S2R2		**4.181**		**0.001**
S2R1 vs. S2R2		0.953		0.357

dwell time of the initial form in the latter case ($p = 0.001$).

Unlike the clear asymmetries evident for F0 value in the results of experiment 1, there was no main effect of F0 on any measure in this experiment. Note that neither of the direct counterparts from the two experiments (condition labels = diotic and S2R2) showed much asymmetry between the low and high F0 cases (despite the lack of a significant interaction between lateralization condition and F0 on any VTE measure in experiment 1). Rather, the asymmetries seen in experiment 1 were associated mainly with the ITD and dichotic conditions, for which monitoring both sequences involved on-going shifts of attention between different spatial positions as well as between different pitch ranges. Although speculative, this pattern suggests that the prevailing high task demands of divided spatial attention tend to have a more adverse effect on reporting VTs for the voice with the low F0 than for the voice with the high F0. Presumably, the high-F0 voice is more salient and so gains a preferential focus for monitoring when the costs of dividing attention between the two F0s are reinforced by a congruent lateralization cue.

Consistent with the results for experiment 1, there was a significant main effect of stimulus word ($p \leq 0.003$) for three of the four measures; there was also a trend towards significance for the number of VTs reported ($p = 0.055$). Appendix B shows the results for each stimulus word when collapsed across condition and F0 value. The only significant interactions occurred for the dwell time of the initial form. The interaction between F0 and word mainly arose because the dwell time was greater on the low F0 for the words “face”, “flame”, and “sleep”, but greater on the high F0 for the words “right”, “noise”, and “sleep”. The three-way interaction arose mainly because the dwell time of the initial form showed little dependence on either F0 or word when only one sequence was present, but when two sequences were present at once the dwell time was strongly dependent on the stimulus word for the high F0 but not for the low F0.

4. General discussion

The results support the idea that perceptual segregation and regrouping contribute to the VTE (Pitt and Shoaf, 2002; Stachurski et al., 2015). Specifically, when two sequences are present in both ears, the greater opportunity this affords for the regrouping of acoustic-phonetic elements typically increases the number of forms heard, speeds up the time to the first VT, and reduces the dwell time of the initial form. This outcome was observed regardless of whether the two voices were distinguished by F0 alone or also by ITD cues for different lateralization. Although the experiments reported here did not explicitly examine constraints on the regroupings possible when two sequences are present in the same ear at the same time, it seems likely that the substantial ΔF0 between the two voices would have limited the opportunities for combining together vocalic parts from both sequences. Using dichotic presentation of concurrent sequences, Warren and Ackroff (1976) found that the cost of dividing attention between two sequences when required to monitor both at once led to an underestimation of the transformations that occurred for each sequence; the current study has shown that this finding extends to cases where the sequences are distinguished only by F0. The effects of F0 on the VTE observed in experiment 1 were most probably a consequence of the task demands of dual monitoring observed in experiment 2.

To our knowledge, hitherto there has been no quantitative analysis of the degree of independence between the transformations reported on the two sequences. The analysis used here supports the notion that responses across the two sequences are relatively independent of one another, but shows that the degree of independence is modulated to some extent by the stimulus configuration used. Specifically, presenting both sequences in both ears (the diotic and ITD conditions) significantly lowered the degree of independence relative to the dichotic case. Nonetheless, listeners still heard the same form for both voices only ~35% of the

time after the first transformation. Note that the finding of fairly independent transformations for concurrent sequences that are not presented to separate ears indicates that the locus for the processes of perceptual organization and linguistic interpretation driving the VTE is the auditory object (i.e., hearing two voices), not the ear of presentation receiving the input. For the concurrent presentation of two sequences, our assumption is that the specific verbal form heard for a given sequence at a particular time arises from an interaction between the neural representation of that sequence in its current state of perceptual organization and the relative extents of adaptation of the linguistic units activated by that neural representation. Although speculative, it seems likely that the partial loss of independence associated with concurrent presentation in the same ear indicates fewer independent changes in the perceptual organization of the two sequences, rather than less independence of the satiation and recovery of the linguistic units responding to each voice.

The finding that the perceptual organization of concurrent sequences is not entirely independent even when dichotic presentation precludes their interaction in the peripheral auditory system is not surprising given the range of central factors known to influence perceptual organization. For general auditory grouping, these factors include attention and switching attention (e.g., Cusack et al., 2004), and the effects of pattern regularity on stream formation and stabilization (e.g., Bendixen et al., 2010; Devergie et al., 2010). For the perceptual organization of speech, additional factors include articulatory constraints (e.g., Basirat et al., 2012) and lexical constraints (e.g., Billig et al., 2013). Listening to speech in the presence of other speech is commonplace and doing so involves concurrent segregation, as in the VTE study reported here. These circumstances typically lead to even more degraded and ambiguous acoustic cues to a spoken message than occur for speech in quiet, owing to energetic and informational masking between voices. The results of the study reported here are in accord with the notion that a comprehensive account of spoken word recognition must involve reciprocal interactions between auditory stream segregation and linguistic knowledge, whereby each affects the other (Billig et al., 2013).

Future research might apply the approach taken here to stimuli like those used in the VTE study by Stachurski et al. (2015). They removed some of the continuity cues from a sequence comprising a repeating word by deleting formant transitions between adjacent phonetic segments or by inserting abrupt changes into the F0 contour. The effect of these manipulations was to increase the number of forms heard and to decrease the time to the first VT, which is in accord with the idea that the perceptual cleaving and regrouping of acoustic-phonetic elements plays an important role in the VTE (Pitt and Shoaf, 2002). One might predict a particularly large number of forms and a short time to the first VT if concurrent sequences are presented following the removal of acoustic continuity cues linking adjacent phonetic segments. It may also prove informative to repeat the experiments reported here but using a smaller ΔF0. A difference of only 3–4 semitones is sufficient to provide an effective cue for segregating concurrent speech (e.g., Brokx and Nooteboom, 1982; Bird and Darwin, 1998), but using sequences less distinct in voice pitch may affect listeners' responses in two ways. Specifically, there is a greater likelihood of confusing responses to the two sequences and also of perceptual reorganizations that involve combining together vocalic parts from both sequences. Separating the contributions of these factors may prove challenging, but one might expect to see differences emerging between the diotic and ITD conditions if confusions between sequences become an important factor when ΔF0 is small. This is because, unlike the diotic condition, the ITD condition involves a salient lateralization cue that helps distinguish between the two sequences even when the difference in voice pitch is small.

In conclusion, the VTE provides a rich paradigm for investigating the processes underlying the perceptual organization and recognition of speech. Here, we used concurrent sequences of a repeating word on different F0s to extend this investigation to circumstances more akin to those where cocktail-party listening conditions prevail (Cherry, 1953). The main finding is that the VTE is facilitated when two sequences are able to interact in the peripheral auditory system, which we interpret as arising from the greater opportunities this provides for the perceptual regrouping of acoustic-phonetic elements (Pitt and Shoaf, 2002). The other result of within-ear interaction is a fall in the degree of independence of responses to the two sequences, which we interpret as arising from less independent perceptual organization of the two sequences rather than from less independence of the linguistic units activated by those organizations. Finally, the high task demands of monitoring two sequences at once lead to fewer reports of transformations. The results for diotic presentation show that this outcome arises mainly from dividing attention across auditory objects rather than across spatial positions.

Acknowledgement

Correspondence concerning this article should be addressed to Brian Roberts, Psychology, School of Life and Health Sciences, Aston University, Birmingham B4 7ET, UK, Email: b.roberts@aston.ac.uk, ORCID: 0000-0002-4232-9459. To access the research data underlying this publication, see http://doi.org/10.17036/researchdata.aston.ac.uk.00000278. This research was supported by Research Grant EP/F016484/1 from the Engineering and Physical Sciences Research Council (UK), which provided a Ph.D. studentship for Marcin Stachurski under the supervision of Brian Roberts. We are grateful to Peter Bailey, Mark Georgeson, Denis McKeown, and Mark Pitt for their helpful comments on this research. Preliminary presentations on this research were given at the Annual Conference of the British Society of Audiology (Southampton, United Kingdom, September 2009), and the 167th Meeting of the Acoustical Society of America (Providence, Rhode Island, May 2014).

Appendix A

Results for experiment 1. Mean (and inter-subject standard error) per stimulus word for each response measure, when collapsed across F0 and lateralization condition.

Stimulus word	VTs/3 min (#)	Forms/3 min (#)	Time to 1st VT (s)	Dwell time of initial form (%)
“Face”	5.24 (1.12)	3.21 (0.44)	74.01 (9.77)	57.11 (6.19)
“Flame”	8.50 (2.15)	2.92 (0.42)	70.02 (9.50)	70.69 (4.73)
“Noise”	4.44 (1.33)	2.06 (0.28)	95.53 (15.15)	78.18 (4.87)
“Right”	4.57 (1.23)	2.08 (0.24)	108.89 (11.93)	80.13 (3.68)
“See”	8.82 (1.90)	3.19 (0.40)	50.20 (13.26)	60.55 (6.64)
“Sleep”	7.96 (1.65)	3.29 (0.47)	57.33 (13.61)	64.20 (5.47)

Appendix B

Results for experiment 2. Mean (and inter-subject standard error) per stimulus word for each response measure, when collapsed across F0 and S/R condition.

Stimulus word	VTs/3 min (#)	Forms/3 min (#)	Time to 1st VT (s)	Dwell time of initial form (%)
"Face"	9.79 (1.44)	5.12 (0.91)	40.19 (7.66)	43.62 (4.03)
"Flame"	11.36 (3.00)	4.88 (0.73)	40.18 (6.25)	51.55 (5.96)
"Noise"	6.36 (1.55)	3.71 (0.93)	76.93 (16.10)	62.50 (7.15)
"Right"	9.31 (1.98)	4.11 (0.76)	46.09 (11.12)	49.71 (5.19)
"See"	11.84 (2.47)	4.88 (0.76)	31.94 (5.27)	47.42 (5.24)
"Sleep"	10.12 (1.76)	4.94 (0.73)	35.66 (6.19)	44.45 (4.03)

References

Basirat, A., Schwartz, J.L., Sato, M., 2012. Perceptuo-motor interactions in the perceptual organization of speech: evidence from the verbal transformation effect. Philos. Trans. R. Soc. B Biol. Sci. 367, 965–976.

Bendixen, A., Denham, S.L., Gyimesi, K., Winkler, I., 2010. Regular patterns stabilize auditory streams. J. Acoust. Soc. Am. 128, 3658–3666.

Billig, A.J., Davis, M.H., Deeks, J.M., Monstrey, J., Carlyon, R.P., 2013. Lexical influences on auditory streaming. Curr. Biol. 23, 1585–1589.

Bird, J., Darwin, C.J., 1998. Effects of a difference in fundamental frequency in separating two sentences. In: Palmer, A.R., Rees, A., Summerfield, A.Q., Meddis, R. (Eds.), Psychophysical and Physiological Advances in Hearing. Whurr, London, pp. 263–269.

Boersma, P., Weenink, D., 2009. PRAAT: doing phonetics by computer (Version 5.1.02) [Computer program]. Retrieved from. http://www.praat.org/.

Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. MIT Press, Cambridge, MA.

Bregman, A.S., Campbell, J., 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89, 244–249.

Brokx, J.P.L., Nooteboom, S.G., 1982. Intonation and the perceptual separation of simultaneous voices. J. Phon. 10, 23–36.

Cherry, E.C., 1953. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25, 975–979.

Cole, R.C., Scott, B., 1973. Perception of temporal order in speech: the role of vowel transitions. Can. J. Psychol. 27, 441–449.

Cusack, R., Deeks, J., Aikman, G., Carlyon, R.P., 2004. Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J. Exp. Psychol. Hum. Percept. Perform. 30, 643–656.

Darwin, C.J., 2008. Listening to speech in the presence of other sounds. Philos. Trans. R. Soc. B Biol. Sci. 363, 1011–1021.

Darwin, C.J., Bethell-Fox, C.E., 1977. Pitch continuity and speech source attribution. J. Exp. Psychol. Hum. Percept. Perform. 3, 665–672.

Darwin, C.J., Hukin, R.W., 1999. Auditory objects of attention: the role of interaural time differences. J. Exp. Psychol. Hum. Percept. Perform. 25, 617–629.

David, M., Lavandier, M., Grimault, N., Oxenham, A.J., 2017. Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency. Hear. Res. 344, 235–243.

Devergie, A., Grimault, N., Tillmann, B., Berthommier, F., 2010. Effect of rhythmic attention on the segregation of interleaved melodies. J. Acoust. Soc. Am. 128, EL1–EL7.

Ditzinger, T., Tuller, B., Kelso, J.A.S., 1997. Temporal patterning in an auditory illusion: the verbal transformation effect. Biol. Cybern. 77, 23–30.

Dorman, M.F., Cutting, J.E., Raphael, L.J., 1975. Perception of temporal order in vowel sequences with and without formant transitions. J. Exp. Psychol. Hum. Percept. Perform. 104, 121–129.

Henke, W.L., 2005. Mitsyn: a Coherent Family of High-level Languages for Time Signal Processing [Computer Program]. Belmont, MA.

Kaminska, Z., Pool, M., Mayer, P., 2000. Verbal transformation: habituation or spreading activation? Brain Lang. 71, 285–298.

Kashino, M., Kondo, H.M., 2012. Functional brain networks underlying perceptual switching: auditory streaming and verbal transformations. Philos. Trans. R. Soc. B Biol. Sci. 367, 977–987.

Keppel, G., 1991. Design and Analysis: a Researcher's Handbook, third ed. Prentice-Hall, Upper Saddle River, NJ.

Lass, N.J., West, L.K., Taft, D.D., 1973. A non-verbal analogue of the verbal transformation effect. Can. J. Psychol. 27, 273–279.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453–467.

Obusek, C.J., Warren, R.M., 1973. Relation of the verbal transformation and the phonemic restoration effects. Cogn. Psychol. 5, 97–107.

Pitt, M.A., Shoaf, L., 2002. Linking verbal transformations to their causes. J. Exp. Psychol. Hum. Percept. Perform. 28, 150–162.

Snedecor, G.W., Cochran, W.G., 1967. Statistical Methods, sixth ed. Iowa University Press, Ames, IA.

Stachurski, M., 2012. The Verbal Transformation Effect: an Exploration of the Perceptual Organization of Speech. Doctoral thesis. Aston University, Birmingham, UK.

Stachurski, M., Summers, R.J., Roberts, B., 2015. The verbal transformation effect and the perceptual organization of speech: influence of formant transitions and F0-contour continuity. Hear. Res. 323, 22–31.

Summers, R.J., Bailey, P.J., Roberts, B., 2010. Effects of differences in fundamental frequency on across-formant grouping in speech perception. J. Acoust. Soc. Am. 128, 3667–3677.

Summers, R.J., Bailey, P.J., Roberts, B., 2016. Informational masking and the effects of differences in fundamental frequency and fundamental-frequency contour on phonetic integration in a formant ensemble. Hear. Res. 344, 295–303.

Titchener, E.B., 1915. A Beginner's Psychology. Macmillan, New York.

Warren, R.M., 1961a. Illusory changes of distinct speech upon repetition – the verbal transformation effect. Brit. J. Psychol. 52, 249–258.

Warren, R.M., 1961b. Illusory changes in repeated words: differences between young adults and the aged. Am. J. Psychol. 74, 506–516.

Warren, R.M., 1968. Verbal transformation effect and auditory perceptual mechanisms. Psychol. Bull. 70, 261–270.

Warren, R.M., 1996. Auditory illusions and perceptual processing of speech. In: Lass, N.J. (Ed.), Principles of Experimental Phonetics. Mosby, St. Louis, MO, pp. 435–466.

Warren, R.M., 2008. Auditory Perception: an Analysis and Synthesis, third ed. Cambridge University Press, New York.

Warren, R.M., Ackroff, J.M., 1976. Dichotic verbal transformations and evidence for separate processors for identical stimuli. Nature 259, 475–477.

Warren, R.M., Gregory, R.L., 1958. An auditory analogue of the visual reversible figure. Am. J. Psychol. 71, 612–613.

Warren, R.M., Warren, R.P., 1966. A comparison of speech perception in childhood, maturity, and old age by means of the verbal transformation effect. J. Verbal Learn. Verbal Behav. 5, 142–146.

Warren, R.M., Warren, R.P., 1970. Auditory illusions and confusions. Sci. Am. 223 (Dec.), 30–36.

Zuck, D., 1992. The verbal transformation effect: auditory illusions as an index of lexical processing and homolog activation. Brain Lang. 43, 323–335.