Hearing Research 323 (2015) 22–31

Contents lists available at ScienceDirect

Hearing Research

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

Research paper

The verbal transformation effect and the perceptual organization of speech: Influence of formant transitions and F0-contour continuity

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 13 August 2014
Received in revised form
9 January 2015
Accepted 12 January 2015
Available online 22 January 2015

ABSTRACT

This study explored the role of formant transitions and F0-contour continuity in binding together speech sounds into a coherent stream. Listening to a repeating recorded word produces verbal transformations to different forms; stream segregation contributes to this effect and so it can be used to measure changes in perceptual coherence. In experiment 1, monosyllables with strong formant transitions between the initial consonant and following vowel were monotonized; each monosyllable was paired with a weak-transitions counterpart. Further stimuli were derived by replacing the consonant-vowel transitions with samples from adjacent steady portions. Each stimulus was concatenated into a 3-min-long sequence. Listeners only reported more forms in the transitions-removed condition for strong-transitions words, for which formant-frequency discontinuities were substantial. In experiment 2, the F0 contour of all-voiced monosyllables was shaped to follow a rising or falling pattern, spanning one octave. Consecutive tokens either had the same contour, giving an abrupt F0 change between each token, or alternated, giving a continuous contour. Discontinuous sequences caused more transformations and forms, and shorter times to the first transformation. Overall, these findings support the notion that continuity cues provided by formant transitions and the F0 contour play an important role in maintaining the perceptual coherence of speech.

1. Introduction

1.1. Temporal order judgments and the perceptual coherence of speech

The intelligibility of speech depends not only on the ability of the listener to identify the constituent phonetic segments, but also to perceive them in the correct order. Our ability to do this is remarkable – speech rate in everyday speech is ~12 phonemes/s (Efron, 1963) and intelligibility remains fairly robust when the rate is artificially increased up to ~40–50 phonemes/s (Foulke and Sticht, 1969). However, perception of the temporal order of speech sounds is a non-trivial problem, even when heard in quiet. This is because the speech signal consists of rapidly changing and diverse acoustic elements – a signal that has been described as a patchwork of buzzes conjoined with hisses, whistles, and clicks (Remez and Rubin, 1992). The problem becomes apparent when one considers the phenomenon of auditory stream segregation (Bregman and Campbell, 1971; see Bregman, 1990). There exists a large body of evidence from studies using rapidly repeating sequences of simple sounds that the auditory system segregates sounds into separate streams based on differences in their spectral and temporal characteristics (for reviews, see Moore and Gockel, 2002, 2012). Critically, it is much harder to judge the relative timing of sounds that form part of separate streams rather than the same stream (Bregman and Campbell, 1971; Roberts et al., 2002). This makes sense from the perspective of auditory scene analysis (Bregman, 1990), because sounds falling in different streams are interpreted as arising from independent sources and so their relative timing is seen as accidental rather than a meaningful property of a stimulus.

A striking demonstration of the effect of stream segregation on temporal order judgments is provided by a study in which listeners heard a repeating cycle of four disparate sounds drawn from the set high tone, hiss, low tone, buzz, and the vowel [i] (Warren et al., 1969). Listeners generally found it easy to identify which items

Abbreviations: CVC, consonant-vowel-consonant; F0, fundamental frequency; PSOLA, Pitch Synchronous Overlap and Add method; VTE, verbal transformation effect

* Corresponding author. Tel.: +44 121 204 3887.
E-mail address: b.roberts@aston.ac.uk (B. Roberts).

http://dx.doi.org/10.1016/j.heares.2015.01.007

were present, but the accuracy of temporal order judgments was at chance when each item was 200 ms long (equivalent to 5 phonemes/s) and the duration per item had to be increased to 700 ms before half the listeners could judge the order correctly. Even experienced listeners required items ~300 ms long for accurate performance. Thomas et al. (1970) extended this approach using repeating sequences consisting only of four steady vowels [i, ɛ, ɑ, u] synthesized on the same fundamental frequency (F0). Despite the greater similarity between sounds, the accuracy of order judgments fell to near chance for items ≤100 ms long (i.e., ≥10 phonemes/s).

1.2. Role of transitional acoustic cues in holding the speech stream together

Why then does speech not normally segregate into its heterogeneous parts, impairing the ability to perceive the order of its constituent phonemes? Some researchers have argued that the perceptual organization of speech depends primarily on speech-specific constraints rather than on general principles of auditory grouping (Remez and Rubin, 1992; Remez et al., 1994). However, the evidence base for such constraints is limited (Roberts et al., 2014), and it remains unclear what might constitute phonetic coherence in acoustical terms (Darwin, 2008). Perhaps the most likely explanation is that the speech signal contains transitional cues that help to bind its elements together. This possibility is suggested by the finding that smoothly changing pure-tone sequences, in which adjacent high and low tones are linked by frequency glides, are less prone to stream segregation than sequences with abrupt changes (Bregman and Dannenbring, 1973). Transitional cues in speech arise from two aspects of production. First, the continuous movements of the articulators generate smoothly changing formant patterns; these formant transitions are characteristic of the speech signal (see, e.g., Stevens, 1998). Second, the F0 contour typically changes in a smooth and continuous way, such that there are no sudden changes between adjacent voiced segments. The experiments reported here examine further the role of formant transitions and the F0 contour in maintaining the perceptual coherence of speech. Without this coherence, the speech signal would fragment into separate perceptual streams, leading to a failure of accurate temporal order perception that would render the speech unintelligible.

Relatively few studies have investigated the role of formant transitions and of the F0 contour in the perceptual coherence of speech, and those studies were conducted before the advent of digital editing and signal processing techniques. Cole and Scott (1973) proposed that, in addition to carrying phonetic information, formant transitions are critically important in holding together the acoustically diverse elements of speech, such as binding the fricative noise to the adjacent voiced vowel in the syllables “sa” and “sha”. To test this hypothesis, listeners were asked to judge the order of four stimuli derived from recorded speech and presented in a repeating tape loop. The key comparison was between consonant-vowel (CV) syllables comprising consonant frication followed either by the first 75 ms of the vowel (i.e., including formant transitions) or by 75 ms of steady vowel (i.e., no transitions). The two versions of each syllable were almost indistinguishable when played individually. However, when played on a tape loop, the consonant and vowel portions of the transitionless versions were far more prone to stream segregation, leading to impaired judgments of syllable order. This outcome supports the notion that continuity cues provided by formant transitions are important in holding together the speech stream (see also Lackner and Goldstein, 1974).

Darwin (1975) proposed that another important factor contributing to the perceptual coherence of speech is continuity of the F0 contour. Darwin and Bethell-Fox (1977) tested this idea using three-formant patterns whose frequencies changed smoothly back and forth between two vowel positions (duration/cycle = 240 ms, comprising 2 × 60-ms steady portions and 2 × 60 ms linear glides). When these stimuli were synthesized on a monotone (F0 = 130 Hz), a single voice was heard and listeners reported approximants (associated with the gliding portions). However, when an abrupt discontinuity between two F0 values (101 Hz and 178 Hz) was introduced midway through each gliding portion (i.e., every 120 ms), two voices were heard – one on each pitch – and stop-consonant percepts became dominant. The main cues for the perception of intervocalic stops in natural utterances are a closure (silent interval) preceded and followed by rapid formant transitions. Hence, Darwin and Bethell-Fox (1977) interpreted the effect of the F0-contour discontinuities as arising from a perceptual division of the continuous formant pattern at the points of the step-change into two streams. Each stream was presumed to be silent during the other, such that each stream would be heard as an abrupt alternation between a silent interval and a vocalic portion fringed by short glides. Further support for the idea that the F0 contour contributes to the coherence of the speech stream is provided by evidence that the phonetic interpretation of a silent gap signalling stop or affricate manner is altered when there is a large discontinuity in F0 across the gap. For example, introducing a brief silent gap between “say” and “shop” usually causes listeners to perceive “say chop.” However, the change from the perception of a fricative to an affricate does not occur if the first word is spoken by a female talker and the second by a male (Raphael et al., 1976; Dorman et al., 1979), presumably because the different source attribution for the two words ensures that the silent gap is no longer interpreted as a meaningful stimulus feature.

1.3. The verbal transformation effect and its relationship to stream segregation

One reason why repeating sequences of sounds are so commonly used in studies of streaming is because repetition increases the tendency for stream segregation to super-normal levels (Bregman, 1990). Another striking phenomenon associated with extended repetition is the verbal transformation effect (VTE). When a recording of a spoken word is repeated many times, listeners begin to report changes in its verbal form (Warren, 1961; for reviews, see Warren, 1996, 2008). Typically, a series of abrupt changes occurs, some to new forms and others back to forms previously reported. For example, a 3-min presentation of a repeated word “ripe” may include the following responses: *ripe*, *right*, *white*, *white–light*, *right*, *right–light*, *ripe*, *right*, *ripe*, *bright–light*, *right*, *ripe*, *bright–light*, *right*, *bright–light* (after Warren, 1961). Classically, the VTE was regarded as the result of two underlying principles. After perceiving a particular form for a time, verbal satiation (adaptation) of that form occurs and a new perceived form emerges from among competing lexical candidates as a result of criterion shift (Warren, 1968). These processes continue and the new form itself undergoes satiation, replacement, and recovery.

Although the VTE was originally interpreted primarily in terms of linguistic processes, and continues to be investigated in that context (e.g., Bashford et al., 2006, 2009), more recent research has often emphasized the key similarities the VTE shares with a range of other organizational phenomena. Notably, Ditzinger et al. (1997) found that the VTE exhibits perceptual switching and alternations that have the temporal dynamics characteristic of bi-/multi-stability first observed for ambiguous visual figures (Schwartz et al., 2012). There is also evidence that common functional brain networks underlie perceptual switching in auditory streaming and in verbal transformations (Kashino and Kondo, 2012).

Pitt and Shoaf (2002) have shown that listeners exposed to a repeating utterance often hear multiple streams, with the verbal transformation corresponding to the foreground percept and the unreported segments corresponding to the background. This re-grouping of phonetic segments influences the verbal form perceived in the foreground. Furthermore, the nature and extent of the re-grouping depends on the acoustic properties of the stimuli. For example, fricative hiss and plosive bursts (noise excitation, high-frequency centroid) are less similar perceptually to the vocalic portions of the stimulus (buzz excitation, low-frequency centroid) and so are more prone to segregate, cleaving off into the background. Recently, the VTE has itself been used as a tool to investigate the influence of the acoustic properties of speech (Stachurski, 2012) and of lexical knowledge (Billig et al., 2013) on the formation of auditory streams.

1.4. The current study

Our principal aim in the experiments reported here was to apply a complementary approach to previous studies of the role of transitional cues, all of which used temporal order judgments as an indirect measure of stream segregation. To our knowledge, this is the first time that the VTE has been used to assess changes in the perceptual coherence of speech arising from manipulations of the continuity of formant transitions and of the F0 contour. Other than Cole and Scott (1973), all the pioneering studies of the role of acoustic transitions in the coherence of speech used synthetic-formant patterns rather than stimuli derived by limited and precision modification of natural utterances. Also, the advent of digital editing and signal processing has changed the ways in which natural utterances can be manipulated. First, portions of a speech signal can be removed and replaced more seamlessly than is usually achieved with analogue tape splicing. Second, the F0 of a speech signal can be modified to conform to a precise constant value or to a pre-defined contour using “pitch warping” techniques (Moulines and Charpentier, 1990).

Given the nature of the VTE, a wide range of measures can be derived from the response patterns of listeners, no one of which can capture all aspects of the phenomenon. Warren (1961) used two measures – the number of verbal transformations (VTs) and the number of unique verbal forms reported – and these measures have been used in most subsequent studies of the VTE. As noted by Warren (1961), the number of VTs heard is not necessarily related to the number of different forms heard, because an indefinite number of VTs might be experienced so long as at least two forms are heard. Another longstanding and widely used measure of the VTE is the time to the first VT, which is based on the observation that some words begin to transform more quickly than others on repetition (e.g., Natsoulas, 1965; Kaminska and Mayer, 2002). Other measures have been used in some VTE studies – for example, Shoaf and Pitt (2002) used the number of VTs back to the original form as an indicator of perceptual coherence and Billig et al. (2013) computed the proportion of time during the trial for which the initial form was reported. However, no consensus has emerged clearly favouring one measure over another. In the experiments reported here, which explore further the role of continuity cues in the perceptual organization of speech, we have used the three most common measures of the VTE – i.e., number of VTs, number of forms, and time to the first VT.

2. Experiment 1

This experiment explored the impact on the VTE of manipulating the continuity of the formant transitions linking the initial unvoiced consonant with the following voiced vowel in repeating sequences of monosyllabic words. We hypothesized that removing transitions involving substantial changes in formant frequency would increase stream segregation and the re-grouping of phonetic elements, leading to a greater tendency to experience VTs.

2.1. Method

2.1.1. Overview

The experiment was conducted using a modified version of the protocol devised by Warren (1961). Listeners received a series of trials, each of which comprised a sequence of continuously repeated tokens of a digitally modified natural utterance, presented diotically. Monosyllabic words were used because items with a large number of phonetic elements tend to evoke fewer verbal transformations (e.g., Warren, 1961). Each sequence was 3 min long; this choice was based on the observations of Pitt and Shoaf (2002), who found that listeners tend to stop reporting changes after that time owing to fatigue. On each trial, listeners were asked to monitor the sequence carefully and to indicate throughout how they perceived it. All stimulus words began with an unvoiced consonant followed by a voiced vowel (cf. Cole and Scott, 1973). There were four conditions, distinguished by a combination of two binary-state variables: (a) whether the extent of frequency change in the formant transitions linking the consonant and vowel segments was large or small (‘strong’ vs. ‘weak’ transitions); (b) whether the portions containing these formant transitions were removed and replaced using samples taken from the adjacent steady portions (transitions-intact vs. transitions-removed versions). For the strong-transitions stimuli, this manipulation introduces substantial discontinuities into the formant-frequency contours. These discontinuities should increase the tendency for stream segregation of the initial consonant from the following vowel, thus compromising the perceptual coherence of the speech signal and increasing the tendency for listeners to experience verbal transformations.

2.1.2. Participants

This study was approved and overseen by the Aston University Ethics Committee. All listeners gave informed consent; they were native talkers of English (mostly British) who reported having normal hearing. Listeners were drawn from a mixed undergraduate and postgraduate university population and received either cash or course credit for participating. Twelve listeners (3 males, mean age = 26.4 years, SD = 4.8, range = 18.9–31.1) successfully completed the experiment.

2.1.3. Stimuli and conditions

Stimuli were derived from recordings of a set of monosyllabic words, carefully articulated by the first author (male, 32 years old, mean F0 ≈ 130 Hz) to obtain tokens with clear formant transitions. All words were CVCs with non-centralized and tense (long) vowels [i, ɑ, ɔ], for which the longer duration of the steady portion made it easier to replace the transitions between the initial consonant and the vowel. In each case, the initial consonant was an unvoiced fricative [f, θ, s, ʃ], plosive [p, t, k], or affricate [tʃ]. The final consonant was unvoiced and involved a closure (plosive or affricate), but the formant transitions associated with this segment were not manipulated. Before the experiment proper, it was established in a pilot study that the chosen set of stimuli would generate a reasonable number of verbal transformations and forms. The stimulus set included 12 words; there were six with strong formant transitions between the initial consonant and the vowel – *short*, *chart*, *sharp*, *seek*, *thought*, and *torch*, which were paired with 6 words producing weak formant transitions – *fort, park, sheep, peak,*

caught, and *porch*. Words were paired such that the first pair was *short-fort*, the second was *chart-park*, and so forth.

To help achieve the target duration of 500 ms per token, an on-screen metronome was used to pace speech production and several examples of each utterance were spoken. Mono recordings with 16-bit resolution and a sampling rate of 22.05 kHz were made using a microphone (Sennheiser MD 918U-T) and Santa Cruz sound card (Turtle Beach) in a single-walled sound-attenuating chamber (Industrial Acoustics 401A) housed within a quiet room. Instances were chosen that were clearly enunciated, on a fairly flat F0 contour, and close to the target duration. Digital manipulation was carried out by the first author; modified tokens were accepted with reference to visual displays of their time waveforms and spectrograms, and to their perceived playback quality. An exact match to the target duration was achieved using the 'stretch' function in Adobe Audition, which shortens or extends audio files without changing their F0, after which 5-ms linear ramps were applied at stimulus onset and offset. PRAAT software (Boersma and Weenink, 2009) was used to set each recording to a constant F0 of 130 Hz 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, allowing source characteristics of the speech (F0 contour) to be changed without changing its filter characteristics. These tokens formed the reference set of stimuli, which constituted the transitions-intact condition.

The edited version of each stimulus word was created by digital manipulation of each token in the reference set, using Adobe Audition. This involved removing and replacing seamlessly the last part of the initial consonant (9–26 ms) and the first part of the following vowel (31–69 ms, corresponding to 4–9 glottal pulses). Within each word pair, the duration of editing required to remove and replace the formant transitions from the consonant and vowel portions of the 'strong-transitions' word was matched when the 'weak-transitions' word was edited. To replace the spliced-out portion of the vowel, a single glottal pulse copied from the steady portion was iterated several times in its place, using the appropriate inter-pulse interval. For the removed portion of the consonant, a segment of corresponding duration from the middle of the same consonant was copied and spliced in. Finally, a PRAAT script was used to apply the original amplitude envelope, extracted from the corresponding reference stimulus. All transitions-removed stimuli were checked to ensure that there were no audible discontinuities and that they sounded as similar as possible to their transitions-intact counterparts. Fig. 1 shows spectrograms of the stimuli for an example word pair, *chart-park*. The marked region encompasses the manipulated portions, for which the formant transitions have been removed and replaced in the edited versions. Note the presence of significant formant transitions in the top-left panel, which illustrates the transitions-intact case for the strong-transitions word (*chart*), and the consequent discontinuities in the transitions-removed version (bottom-left panel).

Each trial was a 3-min sequence comprising 360 repetitions of the same stimulus word, concatenated without silent intervals between tokens. The presentation software was custom written in VB.Net (Microsoft Visual Studio 2005) and run on a PC. All stimuli (16-bit resolution, 22.05-kHz sampling rate) were presented at a long term average of ~75 dB SPL (sound pressure level) using a Santa Cruz sound card (Turtle Beach) and Sennheiser HD480-13II headphones; outputs were calibrated using a sound-level meter (Brüel & Kjaer, type 2209) coupled to the earphones by an artificial ear (type 4153). Each sequence was faded in and out using 500-ms linear ramps, each corresponding to one complete token.

2.1.4. Task and instructions

Listeners were told that they would hear a series of short verbal utterances on each trial. They were asked to respond as soon as they could identify what the voice appeared to be saying, irrespective of whether it was perceived as a word, phrase, pseudo-word, or non-word. This involved pressing the 'down arrow' key,

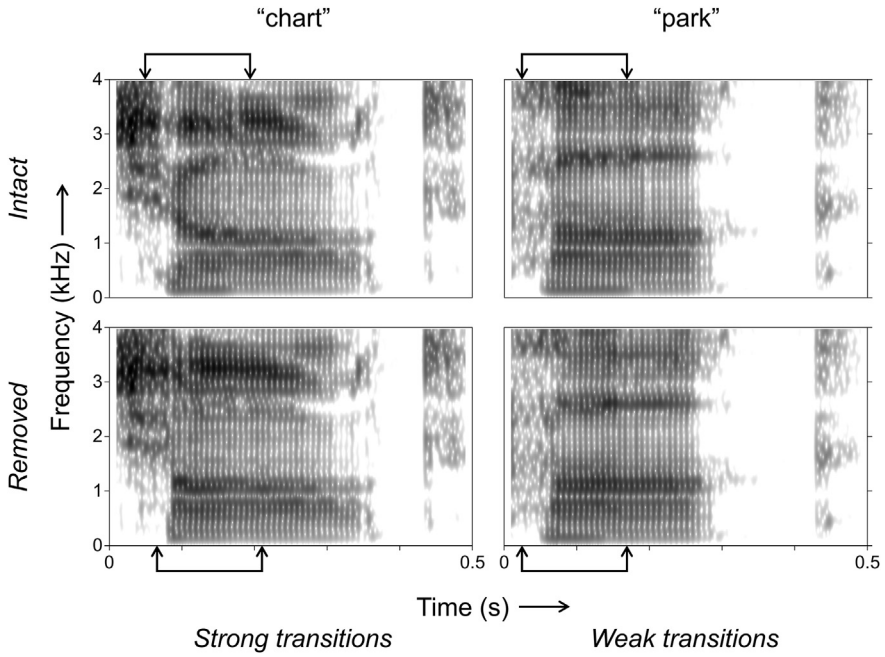

Fig. 1. Stimuli for experiment 1 – spectrograms for the word pair *chart-park*, illustrating the transitions-intact (reference) and transitions-removed (edited) versions of tokens with 'strong' and 'weak' formant transitions (see main text for definition). The regions indicated with brackets encompass those parts of the initial consonant and following vowel that were subject to the editing manipulation, in which the formant transitions were removed and replaced using samples from the neighbouring steady parts of the tokens.

speaking into the microphone (Sennheiser MD 918U-T), and releasing the key. Thereafter, they were asked to continue monitoring the voice and to report any changes in their perception of the verbal form, 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 to a previous form, both of which they must report. It was emphasized that a non-response was as important as a response so that listeners did not feel under pressure 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 regarding how the voice should be perceived and that in some cases they may hear few, if any, changes during a trial.

2.1.5. Recording and transcription of responses

Stimulus presentation over the headphones, recording of verbal responses over the microphone, and recording of key presses were all time-locked. Listeners' verbal responses for each 3-min presentation were saved as 8-bit audio (.wav) files at a sampling rate of 11.025 kHz. Each response entry comprised precision timings of when the key was pressed and released, so that it was possible to assign accurately verbal responses to individual key presses. For each text file entry, the first author 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 initial response to the stimulus was allocated a nominal time of 0 s. For subsequent responses, the time (since the start of the trial) at which the key was pressed was taken as the moment of perceptual change to that verbal form. On a small number of trials there were no subsequent responses, in which case 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, the second response was discounted.

2.1.6. Procedure

Listeners attended four testing sessions in total, each corresponding to one of the four conditions. Each session lasted ~30 min and was taken 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. In order to familiarize them with the task, listeners then completed a practice trial comprising a 1-min presentation of the word *rose*, processed and set to the same F0 as for the reference stimuli. The rest of the test session comprised six 3-min presentations from the main experiment, each consisting of a repeating version of a particular stimulus word (see above) on a constant pitch. Listeners were given a 1-min break between each trial. Subsequent test sessions comprised a recap of the instructions followed by the experimental trials. In any one testing session, there were always six experimental trials; trials using particular words were presented in random order.

Appropriate counterbalancing of conditions across listeners was achieved as follows. First, the 12 stimulus words were arbitrarily divided into two groups of six – the first three (1) and last three (2) word pairs. This has the advantage that both groups included an equal number of strong- and weak-transitions words and, within each session, every experimental stimulus (strong transitions) was always accompanied by its own matched-editing control (weak transitions). The allocation of word-pair groups across sessions was always in the order 1, 2, 1, and 2. However, the editing factor – whether the stimulus was in its transitions-intact (I) or -removed (R) form – was set such that the order across sessions for odd- and even-numbered listeners was I1-R2-R1-I2 and R1-I2-I1-R2, respectively. Hence, four people were needed to complete a balanced set of conditions; the experiment comprised three sets of listeners.

2.1.7. Data analysis

For each combination of transitions type (strong or weak), transitions editing (intact or removed), and word pair, three 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), and the time from the start of the trial to the first VT. Data were analysed in SPSS (IBM, version 20) using within-subjects analysis of variance (ANOVA). The measure of effect size reported here is partial eta squared (η^2_p).

2.2. Results and discussion

Fig. 2 summarizes the results for the three measures used – number of VTs and forms (per 3-min trial), and time to the first VT – when averaged across stimulus words. A two-way ANOVA (transitions type, transitions editing) was performed on each measure and the statistical outcomes are presented in Table 1; significant effects are also indicated on Fig. 2. According to the experimental hypothesis, the critical outcome is the interaction term. This is because removing formant transitions should cause a greater reduction in perceptual coherence for strong- than weak-transitions words, which should manifest as a greater rise in the number of VTs and forms and a greater reduction in the time to the first VT. Appendix A shows the results separately for each stimulus word in the transitions-intact condition.

The results for forms indicates a significant interaction ($p = 0.007$) in the predicted direction – compared with the weak-transitions condition, the discontinuities introduced into strong-transitions words by replacing their formant transitions with samples from the steady portions led to a greater increase in the number of verbal forms reported. Indeed, removing the transitions had little or no effect on the number of forms reported in the weak-transitions case. However, neither of the other VTE measures gave rise to a significant interaction or trend. It is not entirely clear why forms proved to be the most effective measure in this experiment. Some researchers have considered changes in verbal form to be the essence of the phenomenon and have used it as their sole or primary measure of the VTE (e.g., Ohde and Sharf, 1979; MacKay et al., 1993; Kaminska et al., 2000), and others have reported less variability for forms than for VTs across stimuli and listeners (e.g., Lass et al., 1973; Warren, 1996; Stachurski, 2012). Hence, there are at least some grounds for supposing that changes in forms may provide a better measure of changes in perceptual coherence. Nonetheless, VTs and time to first VT are widely used measures and both proved to be as effective as forms in the second experiment reported here (see below).

The only other significant outcome was the main effect of transition type ($p = 0.004$) on the number of VTs reported. This reflects the greater number of VTs reported overall for strong- than

Table 1
Results of experiment 1. Summary of the two-way repeated-measures ANOVAs for the three response measures. Significant terms are shown in bold.

Source	df	F	p	η^2_p
Part (a): Results for VTs				
Transitions type (T)	**(1,11)**	**12.993**	**0.004**	**0.542**
Transitions editing (E)	(1,11)	3.845	0.076	–
T × E interaction	(1,11)	0.598	0.456	–
Part (b): Results for forms				
Transitions type (T)	(1,11)	0.922	0.357	–
Transitions editing (E)	(1,11)	3.606	0.084	–
T × E interaction	**(1,11)**	**11.039**	**0.007**	**0.501**
Part (c): Results for time to first VT				
Transitions type (T)	(1,11)	2.065	0.179	–
Transitions editing (E)	(1,11)	0.278	0.609	–
T × E interaction	(1,11)	0.117	0.739	–

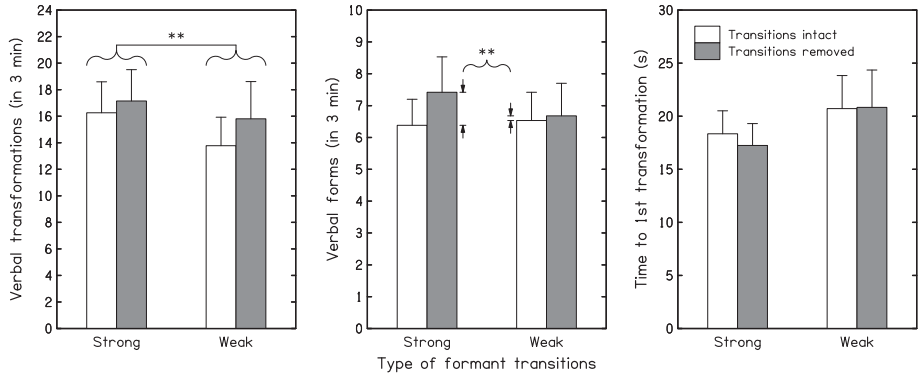

Fig. 2. Results of experiment 1 – the effects of transitions type (strong vs. weak, see main text for definition) and transitions editing status (intact vs. removed) on the number of verbal transformations reported (left panel), the number of verbal forms reported (middle panel), and the time to the first verbal transformation reported (right panel). Means (n = 12) and inter-subject standard errors are shown. The significant main effect for transformations (left panel) and interaction for forms (centre panel) are indicated using brackets and asterisks. The mean differences between the intact and removed cases contributing to the interaction correspond to the distances between the arrowheads, as marked by pairs of short horizontal lines.

for weak-transitions words, but this outcome is uninformative given that different words were used in the two conditions and therefore it is not possible to match them precisely. Hence, this outcome is really a surrogate effect of stimulus word, masquerading as an effect of transitions type. The main effect of transitions editing was not significant, but there are signs of a trend ($p < 0.1$) towards reporting a greater number of VTs and forms in the transitions-removed case. Note, however, that such a trend cannot account for the critical interaction observed for forms.

Overall, albeit with the caveat of the discrepancies between different measures, the results suggest that the effect on stream segregation of removing formant transitions from a repeating sequence of CV syllables observed by Cole and Scott (1973) was not merely an artefact of analogue tape splicing. Rather, the significant interaction found here for the number of verbal forms reported supports the notion that continuity of formant tracks facilitates integration of the heterogeneous acoustic elements comprising speech into a single perceptual stream. This in turn helps to maintain the ability of listeners to perceive the correct temporal order of the underlying phonemes.

3. Experiment 2

This experiment explored the impact on the VTE of manipulating the continuity of the F0 contour in repeating sequences of monosyllabic words with continuous voicing. We hypothesized that a large, abrupt, and repeated discontinuity in the F0 contour would lead to greater stream segregation and re-grouping of phonetic elements than for the continuous case, leading to a greater tendency to experience VTs.

3.1. Method

Twenty four listeners (6 males, mean age = 24.1 years, SD = 5.6, range = 18.5–34.8) completed the experiment, five of whom also took part in experiment 1. The same response protocol was used as before – i.e., listeners were asked to respond by pressing the 'down arrow' key, speaking the verbal form that they heard at that time into the microphone, and releasing the key. Except where stated, the same method and procedure was used as before.

The stimulus set consisted entirely of monosyllabic CVC words with continuous voicing – i.e., there were no phonemes involving closures or unvoiced frication. The words used were: *lathe, maze, nose, vows, wave,* and *writhe*. They were spoken by the same talker and processed in a similar way to the reference stimuli used in experiment 1, except that here the experimental manipulation involved shaping the F0 contour rather than removing and replacing the CV formant transitions. Hence, an important factor in selecting recorded tokens to generate the stimulus set was the extent to which the natural F0 contour could be identified automatically in PRAAT (the algorithm was not always successful at extracting a complete and continuous F0 contour). After the duration of the selected tokens was adjusted to 500 ms, a PRAAT script used the PSOLA algorithm to monotonize the stimulus words and apply the desired F0 contours. The contour shape was a half-sine trajectory on a linear frequency scale, spanning one octave within the normal range of variation for a human male voice. Two versions of each stimulus word were generated – one with a rising F0 contour (100 Hz – 200 Hz) and the other with a falling contour (200 Hz – 100 Hz).

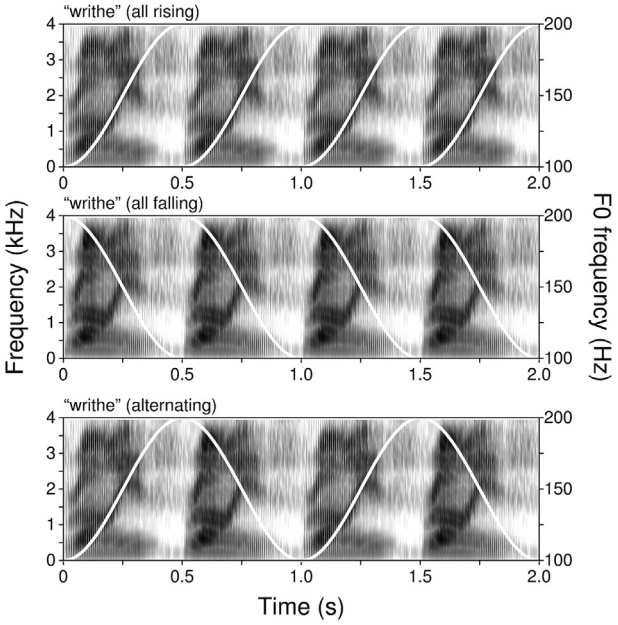

Fig. 3. Stimuli for experiment 2 – spectrograms and F0 contours illustrating the structure of the stimulus sequences used. Each panel shows four cycles for the example word *writhe*. The F0 contours imposed on these tokens (see F0 frequency axis on the right-hand side) have one of three configurations – all rising (RR, top panel), all falling (FF, middle panel), or alternating (RF, bottom panel). Note the large and abrupt discontinuity in the F0 contour at token boundaries in the RR and FF sequences, which is not present in the RF sequence.

Three types of sequence were constructed – one in which all tokens had rising F0 contours (RR), one in which all tokens had falling F0 contours (FF), and one in which successive tokens had alternating F0 contours (RF). Fig. 3 illustrates the differences between the three types of sequence. For the RR and FF sequences, the F0 contour is discontinuous – there is an abrupt change in F0 of one octave at each boundary between adjacent tokens. In contrast, the contour for the RF sequence is smooth and continuous across tokens, as well as within. Pilot work indicated that the *direction* of the pitch change within each token may itself influence the VTE. The reason why this might occur is unclear, but one possibility is that it relates to the linguistic function of intonation – in English, a falling F0 contour is most often used and a rising contour generally signifies a question. To avoid the possibility of any confound arising from contour direction, the experiment was framed as a comparison of two cases – the *continuous* (RF) and *discontinuous* conditions (pooled RR & FF). This framing is balanced in that listeners are exposed to an equal proportion of tokens with rising and falling contours (50% each).

Each listener attended three testing sessions, one for each type of sequence (RR, FF, and RF). As for experiment 1, each session included a 1-min practice trial using the stimulus word *rose* (F0 = 130 Hz) prior to six experimental trials. Full counterbalancing was used. Hence, six people were needed to complete a balanced set of sessions; the experiment comprised four sets of listeners. As before, responses were transcribed and VTs, verbal forms, and times to first VT were calculated.

3.2. Results and discussion

Fig. 4 summarizes the results for the three measures obtained – number of VTs, number of forms, and time to the first VT. Appendix B shows the results separately for each stimulus word when collapsed across condition. A two-way ANOVA (F0 continuity, word) was performed on each measure and the statistical outcomes are presented in Table 2. The critical outcome here is the main effect of contour continuity. Making the F0 contour discontinuous should result in a greater loss of perceptual coherence in the repeating sequence (i.e., increased stream segregation), leading to a rise in the number of VTs and forms reported and a shorter time to the first VT. Significant main effects in the predicted direction ($p < 0.05$) were obtained for all three measures. There were also highly significant effects of stimulus word ($p \leq 0.001$) on all three measures, but there were no significant interactions between contour continuity and word ($p > 0.5$). The results support Darwin and Bethell-Fox's (1977) conclusion that F0-contour continuity plays an important role in binding together the vocalic portions of speech across time, extending their findings for synthetic three-formant patterns to stimuli more similar to natural utterances.

Table 2
Results of experiment 2. Summary of the two-way repeated-measures ANOVAs for the three response measures. Significant terms are shown in bold.

Source	df	F	p	η^2_p
Part (a): Results for VTs				
F0 continuity (C)	**(1,23)**	**4.476**	**0.045**	**0.163**
Word (W)	**(5,115)**	**4.658**	**0.001**	**0.168**
C x W interaction	(5,115)	0.805	0.549	–
Part (b): Results for forms				
F0 continuity (C)	**(1,23)**	**4.352**	**0.048**	**0.159**
Word (W)	**(5,115)**	**16.496**	**<0.001**	**0.418**
C x W interaction	(5,115)	0.206	0.959	–
Part (c): Results for time to first VT				
F0 continuity (C)	**(1,23)**	**6.279**	**0.020**	**0.214**
Word (W)	**(5,115)**	**5.234**	**<0.001**	**0.185**
C x W interaction	(5,115)	0.288	0.919	–

The consistency of outcomes across all three measures observed here provides stronger support for the F0-continuity hypothesis than was obtained for formant transitions in experiment 1. Nonetheless, it would not be appropriate to conclude that continuity of the F0 contour is more important than transitions continuity in maintaining the perceptual coherence of speech. Whilst there is no simple way of equating the extent of the discontinuities involved in the two experiments, the one-octave discontinuity in the F0 contour used in experiment 2 is arguably larger than the sum of the smaller discontinuities in individual formant-frequency contours for the edited versions of the strong-transitions stimuli used in experiment 1. Hence, a direct comparison of outcomes does not allow a fair judgment to be made of the relative importance of the two factors in holding together the speech stream.

4. General discussion

Taken together, the results of the experiments reported here support the notion that acoustic cues for smooth and continuous

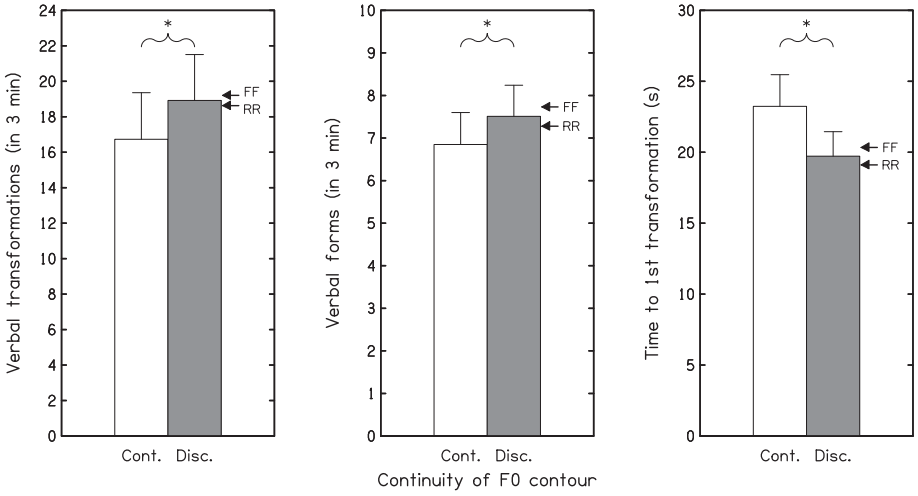

Fig. 4. Results of experiment 2 – the effects of F0-contour continuity (continuous [RF] vs. discontinuous [RR & FF, pooled]) on the number of verbal transformations reported (left panel), the number of verbal forms reported (middle panel), and the time to the first verbal transformation reported (right panel). Means (n = 24) and inter-subject standard errors are shown. For the pooled discontinuous condition, the means of the contributing RR and FF cases are also shown separately using arrows. The significant effects of F0 continuity are indicated using brackets and asterisks.

change, particularly those provided by formant transitions between adjacent phonetic segments and by continuity of the F0 contour, increase the perceptual coherence of the speech signal (Cole and Scott, 1973; Darwin and Bethell-Fox, 1977). This in turn makes the speech signal more resistant to the increased tendency for stream segregation usually evoked by rapid stimulus repetition, reducing the likelihood of perceptual re-grouping of phonetic segments that contributes to the VTE (Pitt and Shoaf, 2002). The results reported here also extend the use of the VTE as an experimental method for investigating the perceptual organization of speech (e.g., Stachurski, 2012; Billig et al., 2013).

The implicit assumption made thus far is that the impact of transitional cues on the perceptual coherence of the speech stream arises purely through acoustic continuity. This may not always be the case. In a recent study, listeners were presented with a repeating sequence of syllables and asked to detect occasional target syllables in which a brief silent gap was inserted between the initial [s] and the rest of the (vocalic) syllable (Billig et al., 2013). The repetition of the syllable was intended to evoke the VTE through segregation of the [s] into a separate stream from the vocalic portion, thus reducing the ability of listeners to judge the relative timing of these elements and hence to detect a temporal gap inserted into the syllable (cf. Bregman and Campbell, 1971). Listeners performed better if segregation of the initial [s] from the vocalic portion caused the syllables presented before the target to transform from a familiar word to a non-word (e.g., from “stone” to “dohne”) rather than from an acoustically similar non-word to a familiar word (e.g., from “stome” to “dome”). This asymmetry indicates that the perceptual organization of speech is influenced not only by the acoustic properties of the speech signal, but also by high-level lexical constraints.

Clarke et al. (2014) have recently provided further evidence that linguistic constraints, as well as acoustic cues, influence the perceptual coherence of speech. Their study used the phenomenon of phonemic restoration (Warren, 1970; Warren and Sherman, 1974), in which top-down constraints are used by listeners to reconstruct phonetic segments that have been obscured or replaced by a broadband masker. The extent of phonemic restoration is indexed by measuring how much the intelligibility of speech interrupted by short silent gaps is increased when those gaps are filled with a noise burst intense enough to have acted as a masker (e.g., Bashford et al., 1992). Clarke et al. (2014) used sentence-length materials, for which the linguistic cues for completion are potentially strong, and measured the effect on phonemic restoration of introducing discontinuities in voice characteristics – specifically, F0 and/or vocal-tract length – across the interruptions. They found that both manipulations significantly reduced intelligibility, but nonetheless the relative benefit of phonemic restoration was maintained. This suggests that there are circumstances in which linguistic cues assist in holding together the speech stream despite the presence of substantial discontinuities in acoustic properties of the voice.

In addition to evidence of the operation of linguistic constraints, there is at least some suggestion that articulatory constraints on plausible acoustic transitions may influence the coherence of the speech stream (Dorman et al., 1975). Cole and Scott's (1973) investigation of the role of formant transitions in binding together sequences of speech sounds was limited to fricative-vowel syllables, for which extended repetition leads quite easily to segregation of the noisy and vocalic portions. Their approach was extended by Dorman et al. (1975) using all-voiced sequences (F0 = 110 Hz) of synthetic formant patterns. Listeners heard repeating four-vowel sequences and judged their temporal order in the presence or absence of transition cues. Fig. 5 shows schematic spectrograms of the five types of sequence used – long and short vowels (V_L, V_S), vowels connected by interpolated formant transitions like those seen in natural speech with coarticulation (V_T), vowels connected by rising initial and falling final formant transitions designed to support the perception of /b/-vowel-/b/ syllables (CVC), and vowels connected to form pseudo-syllables by an inverted pattern of formant transitions implausible for the production of any consonant (C′VC′). Note that, in the steady conditions, the various transitions are replaced by the appropriate steady portions (V_L) or silence (V_S).

As predicted, the accuracy of temporal order judgments was much better for the connected V_T and CVC conditions than for the transitionless V_L and V_S conditions. Indeed, the inferior accuracy observed for the V_S case indicates that transitions between vowel nuclei are more effective than silence at reducing streaming. Accuracy was worst of all for the pseudo-syllables condition, despite the presence of acoustic continuity cues. One interpretation of this outcome is that only phonetically relevant transitions reduce auditory streaming of sequences of speech sounds. Such a conclusion would be consistent with recent neuroimaging evidence that articulatory-based representations play an important role in the

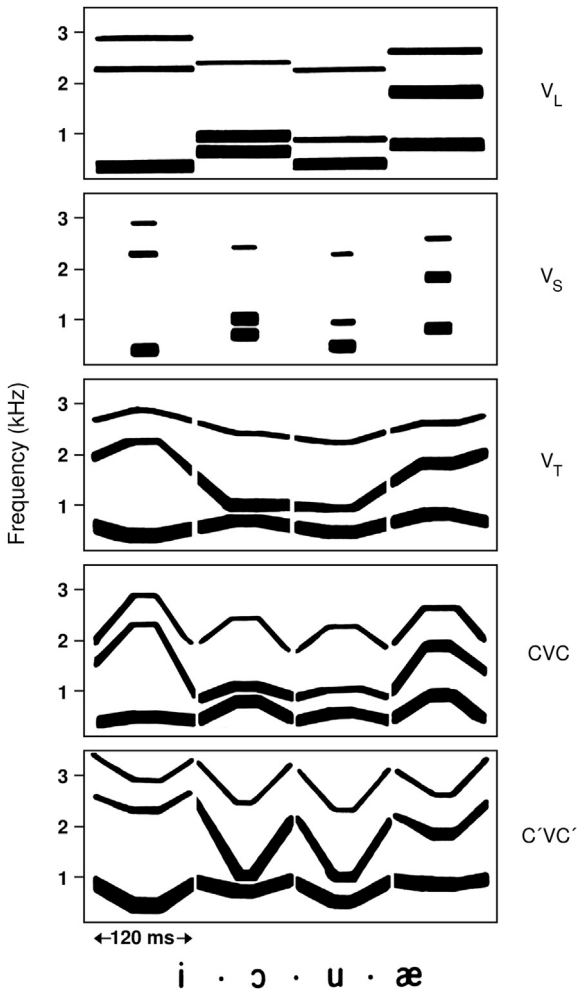

Fig. 5. Schematic spectrograms illustrating the types of sequence used by Dorman et al. (1975). The sequences shown are: long-vowel (V_L), short-vowel (V_S), vowel-with-transition (V_T), consonant-vowel-consonant (CVC), and pseudo-syllables comprising phonetically impossible sequences (C′VC′). Adapted with permission from Fig. 1 (p. 122) of “Perception of temporal order in vowel sequences with and without formant transitions,” by M.F. Dorman, J.E. Cutting, and L.J. Raphael, 1975, *Journal of Experimental Psychology: Human Perception & Performance*, *1*, 121–129. Copyright 1975 by the American Psychological Association.

emergence and stabilization of speech percepts during a VT task (Basirat et al., 2012). However, as noted by Dorman et al. (1975), this conclusion cannot be confirmed because the implausible transitions made identification of the contributing vowels much more difficult, which would itself have compromised order judgments. A possible solution to this confound might be to include short silent gaps between the implausible transitions and the steady vowels, but to our knowledge no subsequent study has pursued this line of enquiry.

In conclusion, the impact of discontinuities on the VTE observed in the experiments reported here supports the proposal that formant transitions and the F0 contour, aside from their linguistic functions, provide continuity cues that help hold together the speech stream, maintaining its perceptual coherence. Without this coherence, supported by acoustic cues and higher-level speech cues, it would be difficult or impossible to judge the temporal order of phonetic segments, rendering speech unintelligible.

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. 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, and Denis McKeown for their helpful comments on this research. We also thank Deniz Başkent and an anonymous reviewer for their comments on an earlier version of this manuscript.

Preliminary presentations on this research were given at the 161st Meeting of the Acoustical Society of America (Seattle, WA, May 2011), and the Annual Conference of the British Society of Audiology (Nottingham, UK, September 2011). The experiments reported here correspond to experiments 3 and 4 in the doctoral thesis of Marcin Stachurski. A number of errors, omissions, and inconsistencies in the thesis report were identified and corrected during the reanalysis of the datasets and the preparation of this article. Most notably, the thesis is inconsistent in its use of the term "form", which sometimes includes and sometimes excludes the first form reported.

Appendix A

Results of experiment 1. Mean (and inter-subject standard error) per stimulus word on each response measure in the transitions-intact condition (unedited reference case)

Word pair (strong, weak)	VTs/3 min (#)	Forms/3 min (#)	Time to 1st VT (s)
short (S)	15.00 (1.88)	6.83 (1.21)	15.67 (2.46)
fort (W)	16.00 (2.20)	6.92 (0.47)	13.95 (5.28)
chart (S)	15.33 (2.29)	6.83 (0.91)	18.25 (3.64)
park (W)	9.75 (2.28)	5.92 (1.15)	36.88 (14.05)
sharp (S)	13.17 (2.58)	5.08 (0.70)	23.98 (3.84)
sheep (W)	12.33 (1.98)	5.33 (0.62)	23.54 (4.50)
seek (S)	16.42 (3.85)	5.00 (0.84)	22.53 (6.17)
peak (W)	15.42 (3.55)	6.42 (1.05)	22.63 (3.48)
thought (S)	20.67 (3.43)	6.92 (0.84)	14.05 (2.95)
caught (W)	14.17 (2.41)	7.58 (1.47)	12.88 (2.33)
torch (S)	17.00 (3.21)	7.58 (1.49)	15.55 (3.16)
porch (W)	15.00 (2.81)	7.00 (1.23)	14.44 (1.66)

Appendix B

Results of experiment 2. Mean (and inter-subject standard error) per stimulus word on each response measure, when collapsed with equal weighting across the continuous F0 (RF) and discontinuous F0 (pooled RR and FF) conditions

Stimulus word	VTs/3 min (#)	Forms/3 min (#)	Time to 1st VT (s)
lathe	18.69 (2.31)	8.70 (0.92)	17.88 (2.72)
maze	17.58 (2.65)	6.94 (0.71)	17.02 (1.61)
nose	15.45 (2.45)	6.02 (0.75)	23.17 (4.33)
vows	20.99 (3.55)	8.48 (0.76)	15.37 (2.05)
wave	15.02 (2.75)	4.96 (0.59)	36.25 (6.32)
writhe	19.21 (2.53)	7.97 (1.00)	19.17 (2.21)

References

Bashford, J.A., Riener, K.R., Warren, R.M., 1992. Increasing the intelligibility of speech through multiple phonemic restorations. Percept. Psychophys. 51, 211–217.
Bashford, J.A., Warren, R.M., Lenz, P.W., 2006. Polling the effective neighborhoods of spoken words with the verbal transformation effect. J. Acoust. Soc. Am. 119, EL55–EL59.
Bashford, J.A., Warren, R.M., Lenz, P.W., 2009. The spread and density of the phonological neighborhood can strongly influence the verbal transformation illusion. Proc. Meet. Acoust. 6, 060002 (8 pages).
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.
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.
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.
Bregman, A.S., Dannenbring, G.L., 1973. The effect of continuity on auditory stream segregation. Percept. Psychophys. 13, 308–312.
Clarke, J., Gaudrain, E., Chatterjee, M., Başkent, D., 2014. T'ain't the way you say it, it's what you say – perceptual continuity of voice and top-down restoration of speech. Hear Res. 315, 80–87.
Cole, R.C., Scott, B., 1973. Perception of temporal order in speech: the role of vowel transitions. Can. J. Psychol. 27, 441–449.
Darwin, C.J., 1975. On the dynamic use of prosody in speech perception. In: Cohen, A., Nooteboom, S.G. (Eds.), Structure and Process in Speech Perception. Springer-Verlag, Berlin, pp. 178–194.
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.
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. 1, 121–129.
Dorman, M.F., Raphael, L.J., Liberman, A.M., 1979. Some experiments on the sound of silence in phonetic perception. J. Acoust. Soc. Am. 65, 1518–1532.
Efron, R., 1963. Temporal perception, aphasia, and déjà vu. Brain 86, 403–424.
Foulke, E., Sticht, T.G., 1969. Review of research on the intelligibility and comprehension of accelerated speech. Psychol. Bull. 72, 50–62.
Kaminska, Z., Mayer, P., 2002. Changing words and changing sounds: a change of tune for verbal transformation theory? Eur. J. Cogn. Psychol. 14, 315–333.
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.
Lackner, J.R., Goldstein, L.M., 1974. Primary auditory stream segregation of repeated consonant-vowel sequences. J. Acoust. Soc. Am. 56, 1651–1652.
Lass, N.J., West, L.K., Taft, D.D., 1973. A non-verbal analogue to the verbal transformation effect. Can. J. Psychol. 27, 272–279.
MacKay, D.G., Wulf, G., Yin, C., Abrams, L., 1993. Relations between word perception and production: new theory and data on the verbal transformation effect. J. Mem. Lang. 32, 624–646.
Moore, B.C.J., Gockel, H., 2002. Factors influencing sequential stream segregation. Acta Acust. Acust. 88, 320–333.
Moore, B.C.J., Gockel, H., 2012. Properties of auditory stream formation. Philos. Trans. R. Soc. B: Biol. Sci. 367, 919–931.

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Commun. 9, 453–467.
Natsoulas, T., 1965. A study of the verbal-transformation effect. Am. J. Psychol. 78, 257–263.
Ohde, R.N., Sharf, D.J., 1979. Relationship between adaptation and the percept and transformations of stop consonant voicing: effects of the number of repetitions and intensity of adaptors. J. Acoust. Soc. Am. 66, 30–45.
Pitt, M.A., Shoaf, L., 2002. Linking verbal transformations to their causes. J. Exp. Psychol. Hum. Percept. Perform. 28, 150–162.
Raphael, L.J., Dorman, M.F., Liberman, A.M., 1976. Some ecological constraints on the perception of stops and affricates. J. Acoust. Soc. Am. 59, S25 (abstract).
Remez, R.E., Rubin, P.E., 1992. Acoustic shards, perceptual glue. Haskins Lab. Status Rep. Speech Res. SR-111/112, 1–10.
Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., Lang, J.M., 1994. On the perceptual organization of speech. Psychol. Rev. 101, 129–156.
Roberts, B., Glasberg, B.R., Moore, B.C.J., 2002. Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. J. Acoust. Soc. Am. 112, 2074–2085.
Roberts, B., Summers, R.J., Bailey, P.J., 2014. Formant-frequency variation and informational masking of speech by extraneous formants: evidence against dynamic and speech-specific acoustical constraints. J. Exp. Psychol. Hum. Percept. Perform. 40, 1507–1525.
Schwartz, J.-C., Grimault, N., Hupé, J.-M., Moore, B.C.J., Pressnitzer, D., 2012. Multistability in perception: binding sensory modalities, an overview. Philos. Trans. R. Soc. B: Biol. Sci. 367, 896–905.
Shoaf, L.C., Pitt, M.A., 2002. Does node stability underlie the verbal transformation effect? A test of node structure theory. Percept. Psychophys. 64, 795–803.
Stachurski, M., 2012. The Verbal Transformation Effect: an Exploration of the Perceptual Organization of Speech. Doctoral thesis. Aston University, Birmingham, UK.
Stevens, K.N., 1998. Acoustic Phonetics. MIT Press, Cambridge, MA.
Thomas, I.B., Hill, P.B., Carroll, F.S., Garcia, B., 1970. Temporal order in the perception of vowels. J. Acoust. Soc. Am. 48, 1010–1013.
Warren, R.M., 1961. Illusory changes of distinct speech upon repetition – the verbal transformation effect. Brit. J. Psychol. 52, 249–258.
Warren, R.M., 1968. Verbal transformation effect and auditory perceptual mechanisms. Psychol. Bull. 70, 261–270.
Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167, 392–393.
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., Obusek, C.J., Farmer, R.M., Warren, R.P., 1969. Auditory sequence: confusion of patterns other than speech or music. Science 164, 586–587.
Warren, R.M., Sherman, G.L., 1974. Phonemic restorations based on subsequent context. Percept. Psychophys. 16, 150–156.