Grouping and the Verbal Transformation Effect: The influence of fundamental frequency, ear of presentation, and interaural time-difference cues

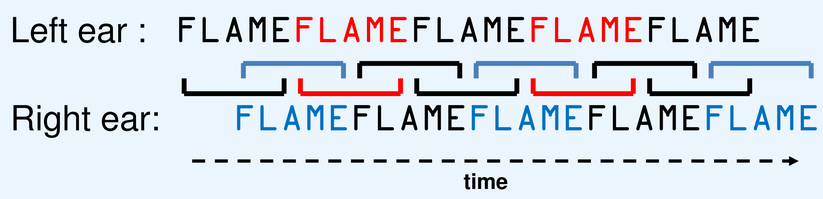
Marcin Stachurski, Robert J. Summers, and Brian Roberts (Psychology, School of Life & Health Sciences, Aston University, UK)

1.Introduction

- Our auditory system needs to separate out the sounds that come from different events in the environment and group together sound streams originating from the same source. This *auditory scene analysis* (Bregman, 1990), is governed by a set of general principles for grouping sound elements; however, these do not seem to account sufficiently for the perceptual coherence of the rapidly changing and diverse acoustic elements of speech (Remez et al., 1994).

Verbal Transformation Effect

- One phenomenon believed to reflect the operation of perceptual mechanisms under difficult listening conditions is the Verbal Transformation Effect (VTE; Warren, 1961). Upon listening to a recycled word, participants report hearing changes to the initial stimulus. For example, a 3-minute 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*.
- The paradigm for the VTE seems to involve two general principles of *verbal satiation* and the emergence of a different form resulting from a *shift in perceptual criteria*. Next, due to the lack of a normal linguistic context, these processes continue and the new form undergoes satiation and replacement.
- The possible role of perceptual re-grouping in the VTE was largely neglected until Ditzinger et al (1997), who observed that different forms in the VTE often occur as alternating pairs. Pitt and Shoaf (2002) have since shown that the perceptual re-grouping of speech sounds plays a key role in the VTE. Specifically, streaming-based verbal transformations (VTs) depend on the acoustic properties of the stimuli. In particular, phonetic elements such as fricatives, affricates, or plosives (stops) cohere less well with adjacent phonemes and therefore are more prone to stream segregation.
- In 1976, Warren and Ackroff explored the effect of stimulating each ear with the same repeated word, but offset by half a cycle to prevent the word being heard as a single fused image. The VTs heard on the two sequences were independent of one another.



- The study is of some interest in relation to the issue of the perceptual re-grouping of acoustic elements in the speech signal. Namely, as the two repeating words are in competition with each other, systematic investigation of the effect of factors such as differences in fundamental (F0) frequency and interaural time-difference cues could inform us about their respective roles in the perceptual re-groupings of the verbal transformations.

Aims

- To use stimulus arrangements designed to encourage competition between different perceptual organizations as an effective means of identifying and characterizing the grouping factors, i.e. to investigate the influence of F0 frequency and ITD cues on the type and pattern of verbal transformations.
- Over the conditions which include differences in F0 frequency and ITD cues, we would expect to find different frequencies and patterns of VTs. For example, as ITD cues are introduced to create a left-right disparity, an increase in perceived lateralization should increase the stream segregation of the two sequences. Even when there is no spatial disparity, F0 frequency differences alone may still enable streaming to occur and listeners to experience VTs.

2. Methods – Stimuli, Conditions, and Procedure

- 6 words: 'face', 'right', 'sleep', 'see', 'noise', 'flame'. All 550 ms long, monotonized, resynthesized at F0=100 Hz (low pitch) and 178 Hz (high pitch)
- Modified version of Warren and Ackroff's (1976) experimental design. However, the two sequences were resynthesized on different F0 frequencies (10-semitone difference)
- 3 lateralization cues: 0-µs ITD (diotic, centralized image), ±680-µs ITD (max natural ITD cue), Dichotic (two sequences in opposite ears)
- 3 sessions, each session had 6 presentations of 3 min each. Listeners (n=12) were asked to report every change in word identity and to indicate on which pitch they occurred.
- Measures taken: number of *verbal transformations* (any change to the reported stimulus) and number of *forms* (any transition that has not occurred before); e.g. train, plane, tray, plane, train, plane, train – 6 VTs and 3 Forms

3. Results – Number of Verbal Transformations and Forms

Average no. of VTs for each location cue and stimulus word across all listeners (each word duration was 550 ms, giving 327 repetitions in 3 min).

	Verbal Transformations reported (in 3 min)	Different Forms (in 3 min)
0 µs ITD	**14.04**	**4.04**
680 µs ITD	**15.24**	**3.90**
Dichotic	**10.24**	**2.63**
Noise	**8.86**	**2.08**
Flame	**17.00**	**3.78**
Face	**10.47**	**4.33**
Sleep	**15.92**	**4.50**
See	**17.64**	**4.31**
Right	**9.14**	**2.14**

VTs

pitch: High (8.03 VTs/3 min) > Low (5.14 VTs/3 min)
word: 'noise' < 'see'

Forms

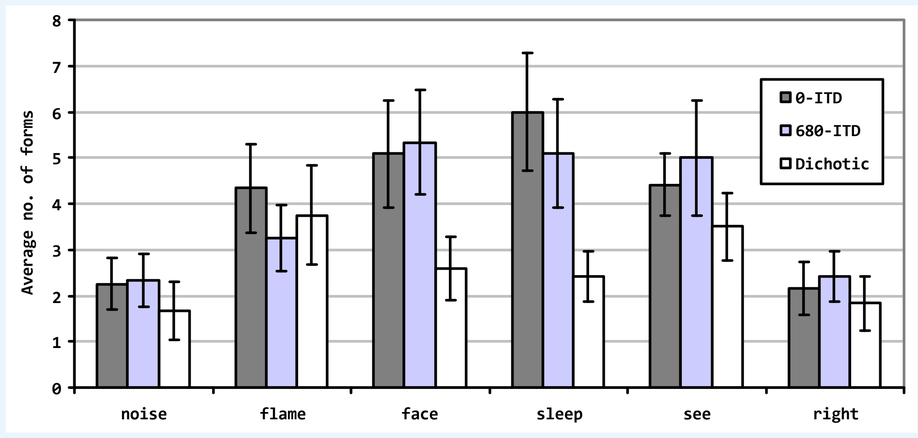
pitch: High (1.90 Forms/3 min) > Low (1.62 Forms/3 min)
word: 'noise' < 'face', 'noise' < 'sleep' 'sleep' > 'right'
lateralization cue:
Dichotic < 0-µs ITD (diotic) & ±680-µs ITD

References

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
Ditzinger, T., Tuller, B., & Kelso, J.A. (1997). Temporal patterning in an auditory illusion: the verbal transformation effect. *Biological cybernetics* 77, 23-30.
Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., Lang, J.M. (1994). On the perceptual organization of speech. *Psychological Review* 101, 129-156.
Pitt, M.A., & Shoaf, L. (2002). Linking verbal transformations to their causes. *Journal of Experimental Psychology: Human Perception & Performance* 28, 150-162.
Warren, R.M. (1961). Illusory changes of distinct speech upon repetition - The verbal transformation effect. *British Journal of Psychology* 52, 249-258.
Warren, R.M., & Ackroff, J.M. (1976). Dichotic verbal transformations and evidence for separate processors for identical stimuli. *Nature* 259, 475-477.

4. Results – Number of Verbal Transformations and Forms – cont.

The Location Cue x Word interaction indicates that, for words 'face' and 'sleep', significantly more forms were reported for the 0-µs ITD (diotic) and ±680-µs ITD conditions than for the dichotic case (see Figure below).



- Suggests that the impact of whether or not the two sequences can interact within the same ear depends on the acoustic properties of individual stimulus words (see below).

5. Results – Timing of the first Verbal Transformation

Average times of the first VT for each lateralization cue and word (nil responses taken as 180 s)

lateralization cue:
Dichotic > 0-µs ITD (diotic) & ±680-µs ITD

pitch: Low (83.93 s) > High (67.40 s)

word: 'sleep' < 'right', 'see' < 'right'

	Time of the first VT (sec)
0 µs ITD	**67.90**
680 µs ITD	**62.09**
Dichotic	**97.01**
Noise	**95.55**
Flame	**70.09**
Face	**74.03**
Sleep	**57.22**
See	**50.24**
Right	**106.88**

- Differences between words: Effect of particular phonetic segments and the likelihood of them 'cleaving off' perceptually from the rest of the stimulus word? Voiceless fricatives 'f' and 's' show a greater tendency for stream segregation than do voiced approximant 'r' or voiced nasal 'n'.

6. Discussion

• Fewer responses in dichotic condition.
As the difference in apparent lateralization between the two ears increased the number of forms was reduced, with the fewest forms in the dichotic condition. For the number of VTs reported, the pattern was similar to that for number of forms (i.e., fewer VTs for the dichotic case). This suggests that if the two sequences can interact with each other by being physically present in the same ear then there will be additional re-grouping taking place that cannot happen when they are sent to different ears.

•Later first verbal transformation in dichotic condition.
Average first VT occurred significantly later for the dichotic case than for the 0-µs ITD (diotic) or ±680-µs ITD conditions, with no difference between the latter two cases. Having the words physically present in the same ear, as is the case with 0-µs and ±680-µs ITD conditions, facilitates the first VT, which suggests that perceptual re-groupings of the phonetic components of words occur more quickly.

•Differences between words.
Word differences probably reflect differences in their phonetic content. This implies that the number of ways in which a given word can recombine depends on the acoustic variation of its elements and the lateralization cue present. For example, as the word 'noise' includes only voiced elements of speech, it is less inclined to transform or produce different forms. It constitutes phonetic elements that appear to be more resistant to perceptual re-grouping. These characteristics include speech sounds with most of their energy in similar frequency regions and which are linked with clear formant transitions. On the other hand, the frequency centroids for fricatives, affricates, and plosive stops are typically higher than for the formants of vowels, nasals, and approximants; also frication and plosion are unvoiced excitation sources. Therefore, these speech sounds may show a greater tendency for stream segregation, and hence words containing them are more likely to transform.

•Pitch difference.
Listeners tended to report more VTs and forms on the high rather than on the low pitch. Additionally, the timing of the first VT occurred significantly earlier for the high pitch. Preliminary data suggests that this difference is due to the task demands, i.e., it is a consequence of listeners monitoring both sequences simultaneously.

•0-µs ITD (diotic) vs. ±680-µs ITD.
There was no significant difference between these conditions. The effect of ITD cues may only be apparent for much smaller delta-F0s than were tested here.

•Additional analyses: VTE independence measure.
Two measures were used to assess the relatedness of responses to the high- and low-pitched sequences. The main measure, the *dependency index*, compared each response to one sequence with the previous and subsequent response to the other. Scores ranged from 0 (independent/unrelated VTs) to 1 (fully dependent/related VTs). Most VTs were found to be independent even when both sequences were present in both ears (0-µs ITD = 0.23, ±680-µs ITD = 0.26), but they were significantly less independent than for the dichotic case (0.09).

The other measure, the *temporal overlap index*, gave the proportion of time after the first VT for which responses to both sequences were the same. The values obtained (0-µs ITD = 0.41, ±680-µs ITD = 0.34, dichotic = 0.36) did not differ significantly. This indicates that the lower dependency index for the dichotic case was not a spurious consequence of more anticorrelation in the responses to the two sequences.

Acknowledgements

Supported by EPSRC. Grant Reference EP/F016484/1 (Roberts & Bailey).
We thank Peter Bailey for his suggestions and advice in relation to this research.

Email: stachurm@aston.ac.uk

Engineering and Physical Sciences Research Council