INTERSPEECH
2009 BRIGHTON

Perceptual grouping of alternating word pairs: Effect of pitch difference and presentation rate

Nandini Iyer, Douglas S. Brungart and Brian D. Simpson

Air Force Research Laboratory, Wright-Patterson AFB, Ohio

nandini.iyer@wpafb.af.mil, douglas.brungart@wpafb.af.mil, brian.simpson@wpafb.af.mil

10.21437/Interspeech.2009-133

Abstract

When listeners hear sequences of tones that slowly alternate between a low frequency and a slightly higher frequency, they tend to report hearing a single stream of alternating tones. However, when the alternation rate and/or the frequency difference increases, they often report hearing two distinct streams: a slowly pulsing high and low frequency stream. This experiment used repeating sequences of spondees to investigate whether a similar streaming phenomenon might occur for speech stimuli. The F0 difference between every other word was varied from 0 - 18 semitones. Each word was either 100 or 125 ms in duration. The inter-onset intervals (IOIs) of the individual words were varied from 100 - 300 ms. The spondees were selected in such a way that listeners who perceived a single stream of sequential words would report hearing a different set of spondees than ones who perceived two distinct streams grouped by frequency. As expected, F0 differences was a strong cue for sequential segregation. Moreover, the number of 'two' stream judgments were greater at smaller IOIs, suggesting that factors that influence the obligatory streaming of tonal signals are also important in the segregation of speech signals.

Index Terms: streaming; sound segregation, speech perception

1. Introduction

The ability to communicate effectively in large, crowded spaces such as restaurants and classrooms depends largely on a listener's ability to extract a target speech signal from the background mixture of other interfering talkers and environmental noise. A listener's ability to do so is believed to be dependent of two mechanisms: first, the listeners must use frequency differences and other cues to separate the target voice from other masking voices that might be talking at the same time (simultaneous segregation), and second, the listeners must track intermittent segments of the target signal and separate them from segments of the masking voices that might alternate with the target over time (sequential segregation). For the simultaneous portion of this segregation problem, cues such as relative pitch differences, level differences, spatial separation etc., are known to be useful in disambiguating a target speech signal from concurrently presented masking voices [1, 2, 3, 4].

Sequential segregation has primarily been examined in studies using sequences of tones [5, 6]. Typically, a two-tone sequence (ABA) comprising of a high frequency and a low frequency tone, varying in frequency difference as well as alternation rate, is presented to listeners. The listeners are then asked to judge if the repeating sequence comprises of two streams or a single stream. The findings are as follows: 1) At relatively slow rates of high-low alternation and at small frequency separations, listeners perceive a single integrated stream, with a galloping rhythm, and 2) At relatively fast alternation rates and relatively large frequency separations, listeners tend to perceive the two streams separately (as a sequence of high frequency tones that is perceptually distinct from a sequence of low-frequency tones). One consequence of the failure to integrate the alternating tones in the latter case is that listeners are much less able to report information about the sequential ordering of the high and low frequency tones.

Following these early demonstrations of streaming using tonal signals, others have demonstrated auditory streaming using more complex sequences such as harmonic complexes (e.g. [7, 8, 9]). Fewer studies have used speech-like stimuli [10, 11]. For example, in an attempt to objectively measure sequential streaming for speech signals, Gaudrain and colleagues [11] asked listeners for temporal order judgments of five vowel sequences, both within a stream and across a stream, while the frequency difference between adjacent vowels was systematically shifted. Their results were consistent with results found for tonal studies, in that within stream correct vowel identification increased as the frequency difference between adjacent vowels increased. Conversely, as the frequency difference increased, listeners' accuracy to make across stream judgments reduced dramatically.

Even fewer studies have investigated sequential streaming of target phrases interleaved with masking signals. Broadbent [12] used two interleaved 3-word messages and showed that the F0 differences between a target and masker were critical in the comprehension of a target signal interleaved with a masker. Kidd and his colleagues [13] used a similar approach, but with 5-word nonsense sentences, and showed that word recall was improved by holding a parameter of the target sequence fixed (voice or ITD).

In a recent study [14], a speech analog of the tonal "ABA" streaming paradigm was used to study speech intelligibility, rather than temporal word order judgments, as a measure of speech stream segregation. The inherent assumption was that high intelligibility would be obtained only if the elements of the target speech were integrated into a separate and distinct stream from those of the interfering words. Results were inconsistent with previous tone/vowel stream segregation experiments, in that target intelligibility did not decline at slower presentation rates or improve at faster presentation rates for a given frequency separation. In fact, target intelligibility was best at slower rates of presentation, except at the two largest pitch difference conditions (8 and 12 semitones), where asymptotic performance was reached. This result suggests that the reductions in intelligibility that occurred for the most rapid presentation rates may have offset any corresponding benefits that might have been obtained from a greater propensity to stream the rapid speech sequences into two streams. Also, it is likely that listeners were able to exploit cognitive strategies to select the target signals at the slowest presentation rates (i.e., building temporal expectancies for the target by foot-tapping, rehearsing etc.).

Another possible interpretation of the results from [14] is that listeners simply are not good at separating two sequentially presented speech signals. In a study conducted by [15], listeners attained high intelligibility in conditions where a target signal that alternated on and off with a 50% duty cycle had to be segregated from an alternating noise that was interleaved with the speech signal, but not when it had to be segregated from an interleaved speech masker. In fact, target intelligibility was always at least as bad with an alternating target/masker stimulus as it was with a simultaneous, uninterrupted target and masker. This was true even when the pitch difference between the target and masker was 12 semitones and the listeners should have had relatively little difficulty segregating the two speech streams.

In this experiment, we used a speech-based paradigm analogous to the temporal order judgment technique that has often been used with tones to investigate sequential grouping for speech signals. This paradigm was based on the selection of four sets of words (ABCD) that were equally plausible to be grouped together as two spondees AB and CD or the alternating spondees AC and BD (see Table 1). For example, the ABCD sequence “BEACH CRAFT WOOD WORK” could either be perceived as the one stream saying “BEACHCRAFT WOODWORK” or two streams, one saying “BEACHWOOD” and one saying “CRAFTWORK”. These word patterns were presented in a continuous loop, and frequency manipulations similar to those used by [5] were used to modify the pitches of these four word patterns either in a sequential low-low-high-high pattern (thus increasing listeners’ propensity to report a single AB or CD stream of words) or a sequential low-high-low-high pattern (presumably increasing listeners’ propensity to report two streams; an AC stream or a BD stream). Differences in word presentation rate were also tested under the assumption that this rate would also influence the listener’s propensity to report the words a a single word stream or two simultaneous word streams. The underlying assumption was that the results for these speech signals would follow the general pattern of results found for tone by [5]. For faster rates of presentation and at relatively large frequency separations, we expected listeners to group a pairs of words that shared a common pitch (two streams) whereas at slower rates of presentation and relatively small frequency differences, we expected listeners to report a pair of words that were temporally adjacent to each other (one stream).

2. Methods

2.1. Listeners

Fourteen listeners (7 males, 7 females), with ages ranging from 21-56 participated in the study. All listeners were tested regularly for their hearing and had hearing thresholds within normal audiometric limits (<20 dB at octave frequencies between 250 - 8000 Hz). All subjects were well-practiced in speech perception tasks and were compensated for their participation.

2.2. Speech Materials

The spondees used in the experiment are shown in Table 1. The spondees in the same row were always presented together. Six talkers recorded six tokens of each word. The recordings were made at a 44.1 kHz sampling rate with a B&K 2131 microphone mounted on a stand positioned directly in front of the talker in an anechoic chamber. The words were hand-edited to remove preceding and trailing silent intervals. Then, fourteen listeners listened to the 6 tokens of each word and selected the two best tokens for each talker. Using the PSOLA algorithm [16] as implemented in the publically available PRAAT speech processing software [17], the duration as well as pitch of the two best tokens from all six talkers were resynthesized. Word durations were manipulated to be exactly 100 or 125 ms. The fundamental frequency (F0) was shifted to generate words with a base frequency of 80 Hz. Six additional frequency-shifted tokens were generated for each word, wherein the F0s were shifted by 3, 6, 9, 12, 15 or 18 semitones relative to the 80 Hz base signal. The actual frequencies corresponding to the semitone differences were 95, 113, 135, 160, 190 and 226 Hz respectively. The inter-onset interval (IOI, the time between the onsets of two adjacent stimuli) of the stimuli was varied by inserting a silent interval between adjacent words in the spondee, in order to create signals with IOIs of 100 (for the 100 ms only), 125, 150, 175, 200 and 200 ms.

Table 1: Spondee words used in the experiment

BEACH CRAFT	WOOD WORK
WORK DAY	WEEK END
BLACK BIRD	DOG HOUSE
FLASH LIGHT	FLOOD GATE
STOP SIGN	LIGHT POST

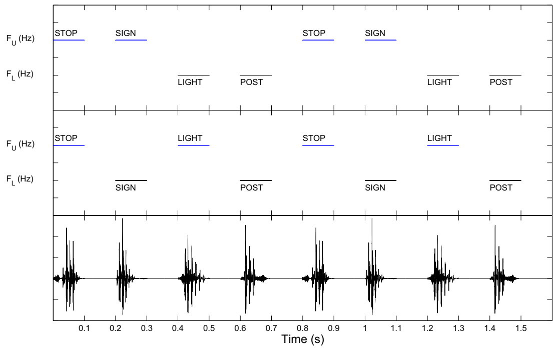

Figure 1: *Upper panel: A schematic representation of the F0 pattern in the ’Sequential Pairing’ pitch configuration. Middle panel: Schematic representation of the F0 pattern in the ’Alternated Pairing’ pitch configuration. Lower panel: Waveform of two repetitions of the two spondees presented at an IOI of 200 ms and word duration of 100 ms.*

2.3. Procedure

All stimuli were presented at a level of 65 dB SPL over headphones (Beyerdynamics DT 990 Pro) to listeners seated in quiet listening rooms. On each trial, listeners heard two spondees in one of two different pitch configurations (see Figure 1).

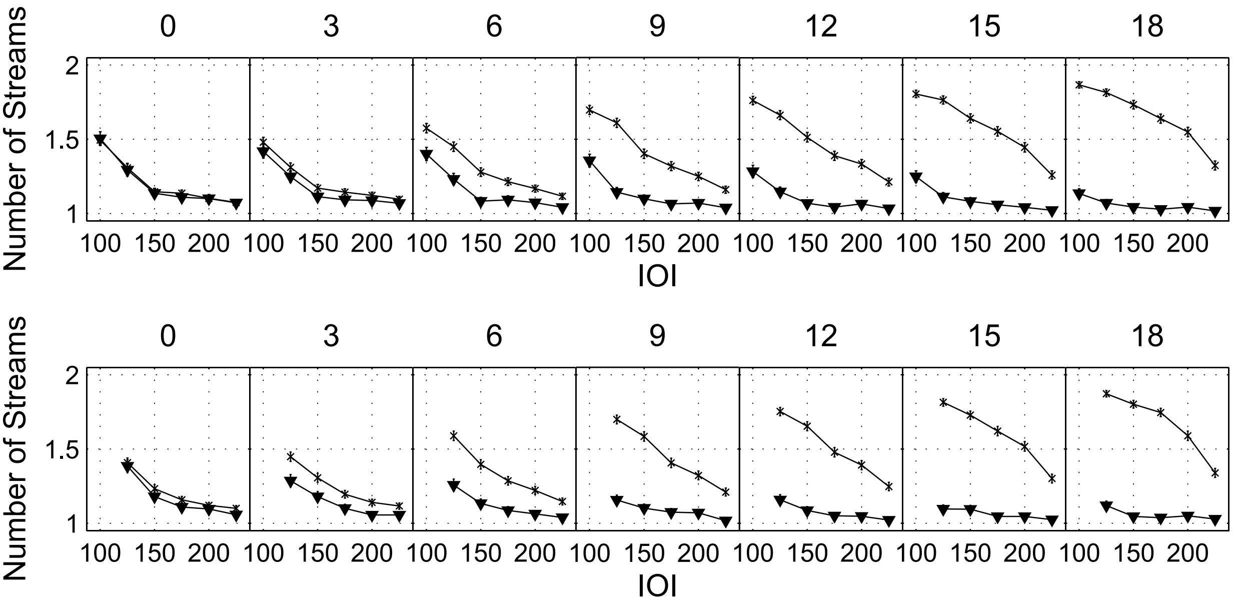

Figure 2: *Average number of 1 vs. 2 stream judgments as a function of IOI and the different F0 differences. The top row plots performance data for the 100 ms word duration, and the bottom row plots performance for the 125 ms word duration. Within each row, each column represents performance in one value of F0 difference condition. The triangles represent performance in the ’Sequential Pairing’ configuration whereas the ’x’ represent performance in the ’Alternated Pairing’ configuration. The error bars in both panels represent ± 1 standard error.*

- Sequential Pairing: In this configuration, both words in a single spondee were presented at the same frequency, with the F0 first spondee fixed at 80 Hz and the F0 of the second spondee systematically varying (i.e. a low-low-high-high pitch pattern).
- Alternated Pairing: The F0 of the first word in each spondee was 80 Hz, but the F0 of the second word of each spondee varied systematically (i.e. a low-high-low-high word pattern).

All possible two-word combinations arising from the spondee pair were displayed on a computer screen as possible responses. On each trial, sixty repetitions of the two spondee sequence (as depicted in Figure 1) were looped and presented to listeners. Listeners responded to the most “prominent” spondee at any time during the presentation of the looped stimuli by selecting a response using a mouse. As soon as they responded by clicking on a single word, the signal presentation was terminated. If the listeners failed to respond at the end of sixty repetitions, they were asked to guess before the next trial was initiated. Each block consisted of 36 trials. Within a block of trials, the relative F0 difference, word duration as well as the IOI were randomly selected from trial-to-trial. A 0 semitone condition was also included, where the F0 of all words was 80 Hz. No feedback was provided to the listeners. Each listener collected a total of 36 trials in each experimental condition.

2.4. Results

Listeners’ responses were operationally defined as “1” vs. “2” stream responses. In both configurations, if listeners responded by selecting either one of the original spondees presented (e.g., stopsign or lightpost; stream AB or CD), their response was assigned the number “1” to indicate that the listeners heard only one stream. If listeners responded to the first or the last words of the two spondees (e.g., stoplight or signpost; stream AC or BD), it was assigned the number “2” to designate an across stream response. In these cases, it was assumed that listeners can hear only the high or the low stream and thus can report only words grouped by common F0.

Figure 2 shows the average number of streams reported in each condition of the experiment. The top panel shows performance for the 100 ms word duration condition and the bottom panel shows performance in the 125 ms condition. Each panel depicts performance for a different ΔF0 condition. From the figure, it is evident that as the F0 difference increased, the subjects became more likely to group together words with the same F0 (leading to a mean response closer to 1 in the sequential pairing condition (triangles) and a mean response closer to 2 in the pairing conditions (x)). Moreover, the rate of presentation of the stimuli also had an effect on listeners’ responses; at a fixed F0 value, the propensity to group the stimuli into different streams by frequency increased as IOI decreased. The results follow similar trends for the 125 ms word durations. Both of these results are consistent with the well-known findings of van Noorden for the streaming of alternating tones [5].

One remarkable aspect of the results is that the listeners did not uniformly make the ’1’ stream response in the ’Sequential Pairing’ configuration, where the AB and CD syllable pairs were presented sequentially at the same frequency. In the condition with no F0 difference (leftmost panels of the figures), the mean number of reported streams approached 1.5 in cases where the IOI was equivalent to the word length. Thus, it seems that in fast presentation rates with silent intervals between the words, listeners were equally likely to respond to sequential or alternating words, especially when the words were presented at the same F0. Data from [11] also showed some evidence of the trend: in their study, across stream vowel order judgments were less accurate when the F0 of the two streams were the same. They hypothesized that listeners could utilize formant differences between the vowels to segregate the streams, even in the absence of F0 differences. It is possible that in the current study, timbral differences between the spondees could have resulted in increased streaming in conditions where other cues were either not present or not salient (i.e., F0 differences). It should be pointed out that studies on temporal order judgments using non-speech stimuli have shown that the stimuli need to be separated by at least 17-20 ms [18]; the time needed to report the order for speech stimuli might be in the order of 50 ms or more [19].

In cases where there were F0 differences, listeners still tended to exhibit a greater propensity to group syllables together by sequential order than by common F0. This is evident from the fact that the mean number of two-stream responses only approached '2' in the 'alternated pairing' conditions with the fastest alternation rates and the largest F0 values.

2.5. Summary and Conclusions

The current study was directed towards understanding if pitch differences and rate of presentation are important in the streaming of speech signals. To our knowledge, this is the first study that has demonstrated the role of these factors using word level stimuli. Our results showed that F0 differences have a powerful influence on how sequences of non-overlapping speech information are grouped into streams.

In practical terms, it is possible that advanced auditory displays might utilize F0 and rate of presentation changes in order to segregate a target talker from other interfering talkers, or to enable listeners to group the messages on one channel to the exclusion of others. Indeed, changes in rhythm and or pitch could be used to alert listeners to changes happening in a certain channel, especially when the change may not require any specific action. It is not yet clear what role top-down factors might play in the segregating speech signals, and a systematic study on these factors is warranted.

3. References

[1] A. S. Bregman, *Auditory Scene Analysis*. Cambridge: MIT Press, 1994.

[2] C. Darwin and R. Hukin, "Effectiveness of spatial cues, prosody, and talker characteristics in selective attention," *Journal of the Acoustical Society of America*, vol. 107, pp. 970–977, 2000.

[3] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acustica*, vol. 86, pp. 117–128, 2000.

[4] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *Journal of the Acoustical Society of America*, vol. 109, pp. 1101–1109, 2001.

[5] L. van Noorden, "Temporal coherence in the perception of tone sequences," *Eindhoven University of Technology*, 1975.

[6] M. Rose and B. Moore, "Perceptual grouping of tone sequences by normally hearing and hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 102(3), pp. 1768–1778, 1997.

[7] P. G. Singh, "Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre?," *The Journal of the Acoustical Society of America*, vol. 82, no. 3, pp. 886–899, 1987.

[8] N. Grimault, C. Micheyl, R. Carlyon, P. Arthaud, and L. Collet, "Perceptual auditory stream segregation of sequences of complex sounds in subjects with normal and impaired hearing," *British Journal of Audiology*, vol. 35(3), pp. 173–182, 2002.

[9] P. G. Singh and A. S. Bregman, "The influence of different timbre attributes on the perceptual segregation of complex-tone sequences," *The Journal of the Acoustical Society of America*, vol. 102, no. 4, pp. 1943–1952, 1997.

[10] M. F. Dorman, J. Cutting, and L. Raphael, "Perception of temporal order in vowel sequences with and without formant transitions," *Journal of Experimental Psychology : Human Percception and Performance*, vol. 1, pp. 121–129, 1975.

[11] E. Gaudrain, N. Grimault, E. W. Healy, and J.-C. Bera, "Effect of spectral smearing on the perceptual segregation of vowel sequences," *Hearing Research*, vol. 231, pp. 32–41, 2007.

[12] D. E. Broadbent, "Failures of attention in selective listening," *Journal of Experimental Psychology*, vol. 44, pp. 428–433, 1952.

[13] G. Kidd, V. Best, and C. Mason, "The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task," *The Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 938–945, 1999.

[14] N. Iyer, V. Best, D. Brungart, B. Simpson, C. Mason, and G. Kidd, "Factors affecting the sequential organization of speech," *Proceedings of the Midwinter Meeting of the Association for Research in Otolaryngology, St. Petersburg Beach, Florida*, 2009.

[15] N. Iyer, D. S. Brungart, and B. D. Simpson, "Effects of periodic masker interruption on the intelligibility of interrupted speech," *The Journal of the Acoustical Society of America*, vol. 122, no. 3, pp. 1693–1701, 2007.

[16] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[17] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer, version 3.4," *Institute of Phonetic Sciences, University of Amsterdam*, vol. 134, pp. 1–182, 1996.

[18] I. J. Hirsh, "Auditory perception of temporal order," *The Journal of the Acoustical Society of America*, vol. 31, no. 6, pp. 759–767, 1959.

[19] D. E. Broadbent and P. Ladefoged, "Auditory perception of temporal order," *The Journal of the Acoustical Society of America*, vol. 31, no. 11, pp. 1539–1539, 1959.