Hearing Research 344 (2017) 235–243

Contents lists available at ScienceDirect

Hearing Research

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

Research Paper

Sequential stream segregation of voiced and unvoiced speech sounds based on fundamental frequency

Marion David [a, *], Mathieu Lavandier [b], Nicolas Grimault [c], Andrew J. Oxenham [a]

[a] Department of Psychology, University of Minnesota, Minneapolis, MN, 55455, USA
[b] Univ. Lyon, ENTPE, Laboratoire Génie Civil et Bâtiment, Rue M. Audin, F-69518, Vaulx-en-Velin Cedex, France
[c] Cognition Auditive et Psychoacoustique, Centre de Recherche en Neurosciences de Lyon, Université Lyon 1, UMR CRNS 5292, Avenue Tony Garnier, 69366, Lyon Cedex 07, France

ARTICLE INFO

Article history:
Received 18 July 2016
Received in revised form
22 November 2016
Accepted 29 November 2016
Available online 5 December 2016

Keywords:
Stream segregation
Fundamental frequency
Speech sounds

ABSTRACT

Differences in fundamental frequency (F0) between voiced sounds are known to be a strong cue for stream segregation. However, speech consists of both voiced and unvoiced sounds, and less is known about whether and how the unvoiced portions are segregated. This study measured listeners' ability to integrate or segregate sequences of consonant-vowel tokens, comprising a voiceless fricative and a vowel, as a function of the F0 difference between interleaved sequences of tokens. A performance-based measure was used, in which listeners detected the presence of a repeated token either within one sequence or between the two sequences (measures of voluntary and obligatory streaming, respectively). The results showed a systematic increase of voluntary stream segregation as the F0 difference between the two interleaved sequences increased from 0 to 13 semitones, suggesting that F0 differences allowed listeners to segregate speech sounds, including the unvoiced portions. In contrast to the consistent effects of voluntary streaming, the trend towards obligatory stream segregation at large F0 differences failed to reach significance. Listeners were no longer able to perform the voluntary-streaming task reliably when the unvoiced portions were removed from the stimuli, suggesting that the unvoiced portions were used and correctly segregated in the original task. The results demonstrate that streaming based on F0 differences occurs for natural speech sounds, and that the unvoiced portions are correctly assigned to the corresponding voiced portions.

1. Introduction

Speech intelligibility in complex auditory environments, such as a cocktail party (Cherry, 1953), relies on our natural ability to perceptually segregate competing voices. To be intelligible, the sequence of sounds spoken by each person must be integrated into a single perceptual stream, and must be segregated from the speech sounds produced by other people. Auditory stream segregation and integration have been studied using both speech and non-speech sounds.

A large body of literature has documented the cues by which simple (non-speech) sounds are perceptually integrated and segregated (e.g., Bregman, 1990; Moore and Gockel, 2002, 2012). One important segregation cue involves differences in frequency or fundamental frequency (F0) between pure tones (Miller, 1957; van Noorden, 1975) and complex tones (Vliegen and Oxenham, 1999), respectively. One difficulty with generalizing the results from studies of streaming to real-world listening is that streaming studies often use sequences of sounds that are exact repetitions of each other, without the variations that are common in everyday situations. Some exceptions include studies of melody discrimination (e.g., Hartmann and Johnson, 1991), and a study involving two interleaved sequences of vowels that differed in F0 (Gaudrain et al., 2007). Listeners in that study were asked to report the order of presentation of the vowels either between or within the two interleaved sequences. Performance in the between-sequence task decreased significantly, while performance in the within-sequence task improved significantly, as the difference in F0 (ΔF0) between the two streams increased. Although this result shows that sequential voiced speech sounds can be segregated

* Corresponding author.
E-mail address: david602@umn.edu (M. David).

http://dx.doi.org/10.1016/j.heares.2016.11.016
0378-5955/

based on F0 differences, real speech also includes many unvoiced sounds, such as fricatives, which must be assigned to the correct speaker and segregated from other competing sounds.

Numerous studies of speech perception in the presence of competing speech have shown that F0 and intonation differences between a target and an interfering speaker can indeed improve the intelligibility of a target (Brokx and Nooteboom, 1982; Assmann and Summerfield, 1990; Bird and Darwin, 1998; Darwin et al., 2003), along with other cues, such as differences in vocal tract length (Darwin and Hukin, 2000; Darwin et al., 2003; Gaudrain and Başkent, 2015) or intensity differences (Brungart, 2001). However, these measures were based on sentence intelligibility. Because of the numerous linguistic and other context effects present in speech, such stimuli do not provide a strong test of whether all voiced and unvoiced segments are correctly assigned to the correct speaker, as some degree of reconstruction could occur based on linguistic or lexical context and constraints.

A stronger test of the binding between consonants and vowels was provided by Cole and Scott (1973), who studied the perceptual organization of repeating syllables consisting of an unvoiced fricative consonant and a voiced vowel (CV), all with the same vowel (/a/) but with different consonants. They found that listeners' ability to judge the order of the sounds was best when the natural sounds were presented, and worsened if the formant transitions between the consonant and its vowel were removed from the vowels. They argued that these vowel transitions play an important role in binding adjacent segments of speech. A more recent study (Stachurski et al., 2015) used the verbal transformation effect (Warren, 1961) to determine the extent to which formant transitions bind vowels to their preceding consonant. Stachurski et al. (2015) found that the number of verbal transformations reported decreased when the formant transitions were left intact, suggesting that the transitions provided additional binding between the consonant and its following vowel, particularly when the formant transition itself was more pronounced.

Although these studies suggest that formant transitions assist in binding successive consonant and vowel pairs, none of them has studied the extent to which this binding is maintained in the presence of competing streams, as would be encountered in a multi-talker environment. The purpose of the present study was to test whether successful streaming of interleaved sequences of speech sounds can be achieved based solely on differences in F0 between the voiced portions of speech, and thus whether the unvoiced segments can be segregated into the correct streams by virtue of their companion voiced segments. On the one hand, the temporal proximity of the unvoiced and voiced portions of a CV pair, along with the formant transitions, might assist in the perceptual fusion of the unvoiced and voiced portions (Cole and Scott, 1973; Stachurski et al., 2015). On the other hand, repeating sequences of spectrally dissimilar sounds (such as the fricative consonant and vowel) can lead to perceptual segregation and, in some cases, spurious perceptual organization (Harris, 1958), even when formant transitions are maintained (Stachurski et al., 2015). Here, naturally spoken CV pairs were generated to produce speech sounds that contained both unvoiced and voiced segments. The speech sounds were then concatenated in random order into sequences. Two such sequences were temporally interleaved, and a difference in F0 was introduced between the interleaved sequences to produce a pattern of speech tokens with alternating F0, and thus induce stream segregation. Performance was measured in tasks that either favored perceptual integration of all the sounds into a single stream or favored perceptual segregation of the alternating sounds into two separate streams.

2. Experiment 1: within- and across-sequence repetition detection with consonant-vowel pairs

2.1. Rationale

The aim of this experiment was to test whether sequential stream segregation of CV tokens can be elicited by differences in F0 between the voiced portions of the tokens. Voiceless fricatives were used as consonants to provide noise-like aperiodic stimuli that did not carry F0 information. Therefore, successful streaming based solely on F0 differences would require additional binding of the voiced and voiceless segments of each CV token. Such binding can occur in naturally uttered speech signals due to spectral transitions between the consonant and vowel (Cole and Scott, 1973; Stachurski et al., 2015). The present experiment tests whether such binding is sufficient to allow segregation of competing streams.

2.2. Methods

2.2.1. Stimuli

The speech sounds were naturally uttered pairs of voiceless fricative consonants and voiced vowels. Because the consonant-vowel stimuli were recorded as a whole, they included a fricative part (the consonant), a transition part (the vocalic part still containing some consonant information) and a voiced part (the vowel). A set of 45 such sounds were recorded by two speakers, one male and one female, both of whom were native speakers of American English. The recordings were made with a microphone (Sennheiser E914) and portable digital recorder (Marantz PMD670) in a sound attenuating booth. The stimulus set was composed of five voiceless fricative consonants ([f], [s], [ɵ], [ʃ] and [h]) combined with nine vowels ([æ], [e], [iː], [I], [ɵ], [ɛ], [ʌ], [ɑ] and [uː]). The [h] is not often considered in studies investigating fricative consonants (Jongman et al., 2000); however, [h] is defined as a glottal fricative consonant in the International Phonetic Alphabet (IPA), and so was included here.

The stimuli had to be short enough to produce automatic or obligatory stream segregation (van Noorden, 1975), but long enough to contain information from both the consonant and vowel. The duration of each token was therefore limited to 160 ms, with 40-ms inter-token intervals, leading to an onset-to-onset time of 200 ms, which is close to the upper limit for observing obligatory stream segregation (van Noorden, 1975; Micheyl and Oxenham, 2010; David et al., 2015). The beginning and end of the recorded speech sounds were truncated and gated on and off with 10-ms raised-cosine ramps. The truncation points were chosen manually to ensure that the consonant and vowel parts of the stimulus had approximately the same length. The spectral shapes of the different vowels were, of course, different, but the spectral shape of the steady-state portion of each vowel did not differ much in the context of different consonants, as expected. The pitch contours of the tokens were flattened using Praat software (Boersma and Weenink, 2001). The stimuli were then resynthesized using a pitch synchronous overlap-add technique (PSOLA), widely used for F0 manipulations of speech sounds, which has minimal effect on the spectral shape of the CV tokens, including the vocalic portions.

Listeners were presented with interleaved sequences in an ABAB… format, with the A and B sequences presented at different F0s. There were 14 speech tokens in each of the A and B sequences, for a total of 28 speech tokens in each presentation, with the speech tokens selected randomly (without replacement) from the total set of 45 tokens for each presentation. The F0 of the A tokens was constant at 110 Hz and 220 Hz for the male and female voice, respectively, while the F0 of the B tokens was set to be ΔF0 semitones above the F0 of A (0, 1, 3, 5, 7 and 9 semitones, i.e.,