Can Listeners Hear How Many Singers are Singing? The Effect of Listener's Experience, Vibrato, Onset, and Formant Frequency on the Perception of Number of Simultaneous Singers

***Molly L. Erickson and †Christopher S. Gaskill,** **Knoxville, Tennessee, †Tuscaloosa, Alabama*

Summary: Objective/Hypothesis. This study investigated whether listener's experience, presence/absence of vibrato, formant frequency difference, or onset delay affect the ability of experienced and inexperienced listeners to segregate complex vocal stimuli.

Study Design. Repeated measures factorial design.

Methods. Two sets of stimuli were constructed: one with no vibrato and another with vibrato. For each set, each stimulus was synthesized at four pitches: A3, E4, B4, and F5. Stimuli were synthesized using formant patterns appropriate for the vowel |ɑ|. Frequencies for formants one through four were systematically varied from lower to higher in an attempt to simulate the acoustic results of corresponding changes in vocal tract length. Four formant patterns were synthesized (patterns A–D). Three pairs were created at each pitch, pairing the formants AB (mezzo-soprano/mezzo-soprano), CD (soprano/soprano), and AD (mezzo-soprano/soprano). Each of these three pairs was constructed in three separate conditions: simultaneous onset; the first voice in the pair with an onset delay of 100 milliseconds; and the second voice in the pair with an onset delay of 100 milliseconds. Using a scroll bar, listeners rated how difficult it was for them to hear each stimulus pair as two separate voices.

Results. The most difficult combinations to segregate were produced with no vibrato and used simultaneous onset. The easiest conditions to segregate were combinations including a "soprano-like" formant pattern (D) in the vibrato condition. Overall, listener's experience did not affect the perceived difficulty of segregation; however, in the presence of vibrato cues, inexperienced listeners did not use delay cues as an aid in segregation in the same manner as did experienced listeners. Once vibrato was removed from the experimental context, inexperienced listeners were able to use delay to aid in segregation in a similar manner to experienced listeners.

Conclusion. Presence/absence of vibrato, formant pattern difference, and onset delay interact in a complex manner to affect the perceived difficulty of voice segregation.

Key Words: Onset–Voice category–Spectral difference–Vibrato–Listener's experience–Perception–Simultaneous singers.

INTRODUCTION

Auditory segregation is a psychoacoustic phenomenon, whereby a listener is able to distinguish individual sound sources within a signal composed of more than one complex sound.[1] The process by which this is accomplished has been termed by Bregman[1] as "auditory scene analysis." During auditory scene analysis, the human auditory system must determine which auditory events should be integrated into a sequential stream produced by one sound source, while also separating those events from others produced by competing sound sources.[2] Researchers have investigated auditory segregation using pure tones,[2] musical instruments,[3] and synthetic complex sounds.[4,5] Several factors have been identified that appear to be relevant in making an auditory segregation task more or less difficult to perform. These include temporal cues, such as relative onset times,[6–8] spatial location of the individual sound sources,[1,9] differences in pitch,[1,10,11] differences in timbre,[12–14] and the presence or absence of frequency modulation (FM).[5,15,16]

Traditionally, these auditory phenomena have been studied in isolation using synthetic sounds using a bottom-up approach in which not only highly controlled but also highly unnatural synthetic stimuli are used. In contrast, singing voice researchers often use a top-down approach in which natural stimuli, with no parameter controls, are used. At present, there remains a huge gap between these two approaches. However, by using a singing voice synthesizer, such as Aladdin Interactive DSP Workbench (Hitech Development, Stockholm, Sweden), it is possible to study these phenomena simultaneously using stimuli that, while highly controlled, sound extremely natural, thereby closing the gap. Only in this way can we understand how the factors that influence auditory segregation interact in real-world phenomena. This understanding is particularly relevant to speech and hearing scientists who are interested in listeners' ability to attend to one auditory stimulus in the presence of other competing auditory stimuli. However, understanding how the listener segregates one competing stimulus from many may also be of interest to choir directors, composers, and music arrangers, as these individuals may wish to either enhance the listener's

Accepted for publication April 26, 2012.

This paper was presented at the 32nd Annual Symposium: Care of the Professional Voice; June 5, 2003; Philadelphia, PA.

From the *Department of Audiology and Speech Pathology, University of Tennessee, Health Sciences Center, Knoxville, Tennessee; and the †Department of Communicative Disorders, University of Alabama, Tuscaloosa, Alabama.

Address correspondence and reprint requests to Molly L. Erickson, Department of Audiology and Speech Pathology, 578 South Stadium Hall, University of Tennessee, Knoxville, TN 37996. E-mail: merickso@utk.edu

Journal of Voice, Vol. 26, No. 6, pp. 817.e1-817.e13

0892-1997/$36.00

doi:10.1016/j.jvoice.2012.04.011

ability to hear multiple stimuli or to reduce the listener's ability to hear multiple stimuli and therefore, they might assign voices to parts or write arrangements accordingly.

PURPOSE

The purpose of this experiment was to investigate possible factors affecting the ability of experienced and inexperienced listeners to perform an auditory segregation task using synthesized female voices as stimuli. Paired synthesized vocal stimuli were constructed that manipulated combinations of the presence of vibrato, relative onset, and formant patterns. Stimuli were synthesized at four separate pitches. It was hypothesized that: (1) experienced listeners would be able to perform the segregation task better than would inexperienced listeners; (2) stimuli with vibrato would be easier to segregate than stimuli without vibrato; (3) the more different the paired formant patterns were from one another, the easier it would be to segregate the two voices; (4) presenting the stimuli with a delayed onset would allow listeners to segregate the two voices more easily than would a simultaneous presentation; and (5) higher frequency stimulus pairs would be harder to segregate than would lower frequency stimulus pairs.

METHOD

Stimuli

The stimuli were synthesized with first and second formant patterns appropriate for the vowel |ɑ| using the Aladdin Interactive DSP Workbench. The frequencies of the first four formants were varied to simulate varying vocal tract lengths to create four distinct formant patterns (labeled as patterns A–D) at pitches ranging from A3 to A5. Formant pattern A had the lowest formant frequencies and was modeled from a professional mezzo-soprano who had been unambiguously categorized as such for more than 8 years. Formant pattern D had the highest formant frequencies and was modeled after a professional soprano who also had been unambiguously categorized as such for more than 8 years. Formant patterns B and C were linearly interpolated to fall between patterns A and D such that the formants in the pattern B were closer to those of A than those of D and the formants in pattern C were closer to those of D than to those of A. Table 1 displays the synthesized formant frequencies for all four formant patterns. Two sets of stimuli were created, one without vibrato and one with a constant vibrato rate of 5.9 Hz and constant vibrato extent of 50 cents. All stimuli were synthesized with a constant source slope of −12 dB per octave, and the formant bandwidths for formants one through four were held constant at 125, 150, 150, and 150 Hz, respectively.

TABLE 1.
Formant Frequencies in Hertz for Formant Patterns A–D

Pattern	F1	F2	F3	F4
A	625	1074	3027	3600
B	680	1141	3098	3674
C	806	1287	3244	3827
D	878	1367	3320	3906

The stimuli were synthesized at four pitches: A3 ($F_0 = 220$), E4 ($F_0 = 330$), B4 ($F_0 = 494$), and F5 ($F_0 = 699$). All stimuli were amplitude normalized using *Cool Edit Pro* 2.0 (Syntrillium Software Corporation, Phoenix, AZ) and trimmed to 1 second in duration. Spline curves were applied to eliminate any onset or offset clicks.

Two equal sets of stimulus pairs were constructed, one set with vibrato and one set without. For the vibrato stimuli, the stimulus pairs were constructed with their vibrato cycles 180° out of phase. This was done to simulate a real-world situation, where the vibrato rates would likely create paired signals that were incoherent with each other.

Three pairs of stimuli were created at each pitch, pairing the formants AB (similar to pairing two mezzo-sopranos), AD (similar to pairing a mezzo-soprano with a soprano), and CD (similar to pairing two sopranos) for a total of 12 pairs. Each of these 12 pairs was constructed in three separate conditions: simultaneous onset; the first voice in the pair with an onset delay of 100 milliseconds, and the second voice in the pair with an onset delay of 100 milliseconds (for a total of 36 stimulus pairs). Tokens were constructed, one for each stimulus pair, consisting of four repetitions of the stimulus pair with a 500 milliseconds delay between each repetition. Listeners made one judgment per token.

Listeners

Inexperienced listeners were recruited from introductory general education undergraduate psychology courses at the University of Tennessee, Knoxville, TN, USA. They had no formal vocal or instrumental training or experience or interest in classical singing. One hour of extra credit toward their course grade was offered in exchange for their participation. Experienced listeners were recruited from the University of Tennessee School of Music and the Knoxville Opera Company and had a minimum of 5 years formal training and experience in vocal music. Fifteen experienced listeners and 33 inexperienced listeners were recruited for the study. All participants demonstrated hearing ability within normal limits bilaterally as evidenced by a 25-dB hearing screening at 1000, 2000, and 4000 Hz. All participants signed statements of informed consent as approved by the Institutional Review Board at the University of Tennessee.

Procedure

The listeners were instructed that they would hear pairs of simultaneous female voices producing the sustained vowel |ɑ|, and that each stimulus would be presented four times with a short pause between each presentation. They were told that they were to rate how difficult it was for them to hear each stimulus as two separate voices. The participants were seated in an IAC (Winchester, Hampshire, UK) sound booth (model 101302) for both the hearing screening and the experimental task. During the task, participants were seated in front of a computer monitor wearing Sennheiser H265 (Old Lyme, CT) linear headphones. All stimuli were presented in random order for each participant. They were instructed to choose a difficulty rating after the presentation of each stimulus by moving a scrollbar on a 0–50 visual analog scale presented on the monitor with the descriptors "Very Easy" and "Very Hard" at the opposite poles. Each