Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Spatial separation and timbral difference affect identification of pitch modulation in one of two sources

Hong Jun Song and William L. Martens

Faculty of Architecture, Design and Planning
University of Sydney
Australia

PACS: 43.66.Lj

ABSTRACT

The ability of listeners to identify which of two simultaneously presented stimuli exhibited a sudden increase in periodic pitch modulation (vibrato) was examined in a divided attention task. While it is well established that stream segregation is influenced by both timbral differences and spatial separation between simultaneously presented stimuli, the interaction between these two salient factors had not been systematically investigated for such a task. The results of the study reported here showed that the facilitation of identification performance due to spatial separation of sources differing in timbre was greatest when those stimuli differed least in timbre. In particular, when two simultaneously presented sounds had very distinct vowel coloration, spatial separation of the two did not improve the ease of identifying which of the two sources exhibited a sudden increase in periodic pitch modulation. By quantifying vowel coloration differences for pairs of stimuli in terms of the Euclidean distance between their first two formant frequencies, it was shown that identification performance for stimuli exhibiting the most similar vowel coloration was most affected by spatial separation. In all cases tested, however, identification performance for spatially separated stimuli was superior to that for co-located stimuli. The results reveal the relative salience of these two prominent factors, spatial separation and timbral difference, in determining the effectiveness of concurrent auditory information displays.

INTRODUCTION

When two or more sound sources are presented simultaneously, a number of factors influence the listener's ability to extract target information from each, and to identify which of the targets contains some pattern of interest. For example, if one of two simultaneously presented sources exhibits a sudden increase in pitch modulation, as was the case for the stimuli presented in the current study, a listener's ability to identify which of the two sources was so modulated may be influenced by factors such as the timbral difference between the sources as well as the spatial separation between the two sources. Within such divided attention tasks, there has been a growing interest in determining the relative salience of spatial separation versus other factors that might improve performance on identification tasks, such as those used in speech intelligibility studies [1, 2, 3]. The task emplyed in the current study required listeners to perform an identification in terms of a more elemntary attribute than that required in speech intelligibility studies, testing the relative influence of speech-like tonal coloration differences on pitch modulation identification given sources that were either co-located or spatially separated.

The context in which this work takes place is that of empirical evaluation of deign features for auditory display applications. Often in such applications, there is a need for information in a target stimulus to be extracted within the presence of other competing sources. The process in which listeners assign multiple sound elements to independent sources is called auditory streaming [4]. This process can be influenced in many ways, including physical cues or experiential and cognitive factors. The ability of the listener to separate individual meaningful signals from an acoustic mixture depends on how effectively the relationships of these attributes are expressed. Bregman [4] described the perception of sound variables that affect streaming. *Frequency separation*: when the frequency alters between high and low tones, the perception of input stimuli changes into two separate tone streams. If lower tones are below the "fission boundary", the sound sequence is always perceived as a single stream, whereas higher tones beyond the "temporal coherence boundary" can form the segregation of two streams. *Rate of alternation*: faster presentation rates of alternating high and low frequency tones can cause the two tones to be segregated into two streams. *Duration*: sufficient intervals are able to split a single sound input into separate streams. Other factors, such as rhythm influencing the perception of a phrase boundary, synchronicity of tone pairs, and harmonicity, have all been considered efficient cues for streaming.

With the aim of relieving interference and organizing differentiation between simultaneous signals, there has been much interest in using spatial location of sound sources as a cue to direct the listener's attention and to unmask simultaneously presented sound signals. However, a comprehensive framework to explain our ability to understand non-speech sound signals in complex auditory scenes is lacking. Although spatial separation of simultaneous auditory streams seems to be strong, it was not treated as a dominant cue. The effect of spatial separation on streaming seems ambiguous and there was no clear evidence to verify that spatial location was the single cue assisting in separating and categorizing the mixture of sound signals. Bregman [4] argued that spatial location is just one of a number of cues contributing to stream segregation – "location differences alone will not be powerful influences on grouping". Culling and Summerfield [5] believed that lateralization cues can be used to identify simultaneously presented sound sources only after other separation processes have occurred.

Arguments also exist related to attention that listeners distribute to filtered-out inputs. The term *selective attention* describes the process of focusing on a specific subset of all inputs and only target information needs to be extracted within the presence of other competing sources. In selective attention tasks, spatial separation of multiple stimuli has proven ability to improve performance nearly as much as other segregation cues. This has been studied extensively with the "Cocktail Party" phenomenon, and the intelligibility of spatial separation in unmasking competing sources has been verified either in distance or in direction (discussed in the following sections). Another class of auditory display study emerged from *divided attention*, in which listeners are typically asked to perform more than one auditory information processing task simultaneously. It is believed that in a listening task involving divided attention, listeners have to split attentional focus and shift between different subsets of inputs. The complexity of concurrent auditory display has limits owing to the limitation of our information processing ability. Hawkins and Presson [6] stated that knowing where the sound is coming from seems to be of little help in detection because the listener's task is to report the contents. Shinn-Cunningham and Ihlefeld [7] made a similar claim that spatial separation could not dramatically improve identification performance in a divided attention task due to the cost of distributing auditory attention across locations.

THE PRESENT EXPERIMENT

The weight of evidence against spatial separation is mainly on its reliability in contributing to perceptual segregation during auditory scene analysis. Timbral dissimilarity between pairs of displayed streams can be understood in terms of the differences in formant frequencies that were used to modify the tone color of the stimuli in each of the streams. The aim of this study was to explore the effects of spatial separation and tonal color differences on perceptual segregation in a situation where listeners were required to identify the occurrence of periodic modulation within one vowel sound when two vowel sounds were simultaneously presented. This study also examined whether differentiating two simultaneous vowel sounds by spatial presentation (with interaural level difference) and tonal coloration would increase the accuracy of discriminating pitch modulations.

Two competing vowel sounds lasting approximately 4 seconds, were presented simultaneously via headphones for modulation discrimination. Fundamental frequencies of the stimuli glided either upwards or downwards, converging onto a single mean value, after which one of them was periodically modulated in pitch vibrato. A divided attention to both vowel sounds was needed since the listeners had no prior information about when the vowel sounds, and which stream of vowel sounds, would have pitch modulation because of being periodically modulated with pitch vibrato.

Two simultaneous vowels were either spatially co-located or separated, by altering the interaural level difference from $0dB$ (both centered) to $-10dB$ (one close to the left and the other close to the right channel). For the spatially separated presentation, listeners had to distribute their attention across spatial locations within the head (lateralization) to monitor the paired vowel sounds. This study employed Interaural Level Difference (ILD) as a lateralization cue for perceptual segregation.

METHOD

Participants

All subjects reported that their hearing was considered normal and they had normal or corrected vision. Written consent was obtained prior to the experiment from all participants. There were 21 subjects between the ages of 18 and 20 years ($mean = 18.5, SD = 0.68$), 13 females and 8 males.

Stimuli

The simultaneous headphone presentation of two vowel sounds was either co-located in the center or separated with left/right interaural level difference. Single vowels were three-formant synthetic English vowels, with individual formant bandwidths and fundamental frequencies (F_0) either gliding from $100Hz$ upwards or from $300Hz$ downwards, and merging at a constant F_0 of $200Hz$. Single vowels were of approximately 4 seconds duration, including 2.5 seconds with F_0 gliding and 1.5 seconds at the constant F_0.

Sound stimuli resembling three English vowels /ɪ/, /æ/, and /ɒ/ (as in "hit", "hat" and "hot") were used in the current experiment. In the follow sections, these three vowels are represented using their IPA (International Phonetic Alphabet) symbols. Those three vowels were synthesized by using MATLAB with the first formant (F_1), second formant (F_2), and third formants (F_3). Formant frequencies did not vary over the duration of the stimuli and were set to the values reported in [8]. The formant frequencies and bandwidths are shown in Table 1.

Table 1: Formant frequencies and bandwidths used to synthesize the three vowels. The formant frequencies (in Hz) were suggested by [8].

		/ɪ/		/æ/		/ɒ/	
	Formant	Frequency (Hz)	Bandwidth (Hz)	Frequency (Hz)	Bandwidth (Hz)	Frequency (Hz)	Bandwidth (Hz)
Formant	F1	390	65.2	660	86.8	730	92.4
	F2	1990	193.2	1720	171.6	1090	121.2
	F3	2550	238	2410	226.8	2440	229.2

The fundamental frequency F_0 had a rising glide from $100Hz$ to $200Hz$ for one vowel within the pair and a falling glide from $300Hz$ to $200Hz$ for the other vowel. A vocal jitter creating a natural sound character was used for all vowels - at a maximum jitter 2.2% of F_0. This value was selected by the experimenter through comparison with a few higher and lower rates. The duration of the F_0 glides was around 2.5 seconds and for the remaining 1.5 seconds, the pitch contour for both vowels remained at a constant (F_0) frequency of $200Hz$ (shown in Figure 1). Within the 1.5-second duration, one of the paired vowels was modulated with a pitch vibrato function and the other was given no additional modulation (beyond the jitter). A periodical time delay of the vocal signals produced a periodical variation in pitch that introduced a vibrato effect in the vowel stimulus. The implementation of vibrato effect employed the method described in [9]. The vibrato in this study had a modulation frequency rate of $5Hz$ and with pitch variation of around 6%.

For spatially co-located presentation, the vowel sounds had $0dB$ of gain for both channels; for separated presentation, the vowel sound power at one channel was attenuated by $10dB$, leaving the other channel at $0dB$. By applying $10dB$ level attenuation of vowel sound level at left or right channel, the stereo display generated a left-and-right sound source lateralization along the interaural axis. That is, spatial separation of two simultaneous vowels by ILD could contribute to perceptual segregation of the paired vowels.

The stimulus set had 24 stimuli ([3 pairs of vowels] ×[2 gliding directions] ×[2 vibrato choices] ×[2 spatial presentations]) and was played twice. Vowels were presented in one of three pairs:

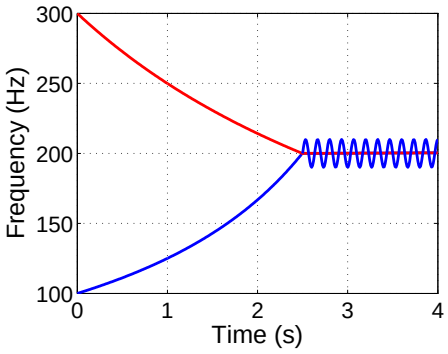


Figure 1: Vibrato implementation. One vowel had falling F_0s gliding from $300Hz$ to $200Hz$ and the other vowel had rising F_0s gliding from $100Hz$ to $200Hz$. After both vowels reached the same F_0 frequency of $200Hz$ at 2.5 seconds, one was modulated with delay-based vibrato till the end. In this figure, the vowel signal with rising F_0 was augmented with vibrato, which is exaggerated here for clarity.

/ɪ/|/æ/, /ɪ/|/ɒ/, and /æ/|/ɒ/; the pitch contour of each was termed up or down, the two gliding directions before the pairs merged at the identical F_0; the vibrato was shown after they merged, either in the vowel gliding up or in that gliding down; the spatial sound source of the paired vowels was either co-located in the center or separated into the left/right channel.

Procedure

All subjects took the test with Sennheiser HD 415 headphones. The training and test modules were designed with Max/MSP [10]. Pairs of vowel sounds were presented in random order and in different random order when replaying.

In training, subjects first listened to the normal individual vowels and modulated ones (a brief presentation of each was about 1 second), and then experienced with 4 trials. At the end of training, four stimuli that the listeners were practising on and the modulated vowels within the pairs were provided. Participants were able to check their responses and replay the stimuli as many times as they needed until they were ready to start the test. Through the training, all participants were expected to recognize the difference between normal vowel stimuli and modulated vowel stimuli and be familiar with the modulation discrimination and keypunching tasks.

In the test, subjects were required to listen to the pairs of vowels and to identify the vowels with the vibrato as quickly and accurately as possible. During the test, subjects performed a two-alternative forced-choice judgment (2AFC). The simultaneous vowels were preassigned keys of "∼" or "-" on the keyboard. The paired vowels to be presented and their corresponding preassigned keys were prompted on screen when the stimuli were loaded before starting play. Thus participants were reminded through visual presentation of the two button symbols and the corresponding vowel IPA symbols. Once the modulation was found, subjects responded on a conventional terminal keyboard with either of the two keys.

The duration of presentation for each pair of vowels was approximately 4 seconds. There was an 8-second interval that allowed participants to change their responses, before the next stimulus was automatically loaded. The time left for loading the next stimulus was also shown in seconds on the screen. Participants were able to press the same key twice but were aware that the 8-second interval was compulsory and that they had no control over the pace. At the end of each trial, no feedback was given.

Data Analysis

In the following discussion, a set of symbols is introduced to represent the variables:

D - Discrimination response;
S - Spatial presentation;
R - Round;
P - Vowel pairs;
V - Vowel.

The Pearson Chi-square test of association was employed to examine the significance between two variables. Simple two-way contingency tables containing counts (i.e. response frequencies) were constructed. The significance was calculated based on the analysis of deviation of observed frequencies from expected frequencies.

RESULTS

The Max/MSP experiment interface recorded response keys and response time (RT). When participants pressed buttons on the keyboard more than once for one pair of vowels, e.g. to change their responses, those keys pressed and the keypunching time in milliseconds (counting from the start of the stimulus) were recorded. For the sake of analysis, only the last key pressed and its time were marked as their decision; if the time of the last key press was less than 2.5 seconds (the button pressed prior to the beginning of vibrato), it indicated that the participant had failed to make a proper response, and the response was marked as incorrect regardless of which vowel they had responded to. This involved 39 (out of 1008) responses.

Spatial presentation and paired vowels on discrimination performance

First, a Chi-square test confirmed the improvement of paired-vowel segregation in a spatially separated presentation compared with a co-located presentation. The correct responses had changed from 255 of a total of 504 in the spatially co-located presentation to 300 of a total of 504 in the separated presentation. This increase was significant, $\chi^2 = 9.75$, $p < 0.005$, which suggested spatial separation was a critical factor for enhancing the discrimination of modulation occurring in paired vowels. Second, the effect of tonal separation of paired vowels on modulation discrimination was measured on the basis of the degree of performance difference and the degree of dissimilarity of tonal coloration.

The tonal coloration difference between vowel stimuli was quantified as a summed Euclidean distance between the formant frequencies of each in (F_1, F_2) space. The pair /ɪ/|/ɒ/ had the highest accuracy level, and this could be explained by /ɪ/ and /ɒ/ are far apart in (F_1, F_2) space with an Euclidean distance on the log-transformed formant frequencies of 1.25 (log) (see Figure 2). The Euclidean distance between /ɪ/ and /æ/ is 0.78 (log), which is greater than that of 0.67 (log) between /æ/ and /ɒ/. The two most dissimilar tonal colorations resulted when the two stimuli differed in terms of both F_1 and F_2, as was the case for the stimuli that sounded as /ɪ/ vs. /ɒ/. For the other two pairs, /ɪ/ vs. /æ/ differed primarily in terms of F_1 and /æ/ vs. /ɒ/ differed primarily in terms of F_2. These differences were quantified by taking the Euclidean distance of the formant frequencies after these frequencies had been expressed in terms of octave scaling.

The differences in performance were measured by a Z-score

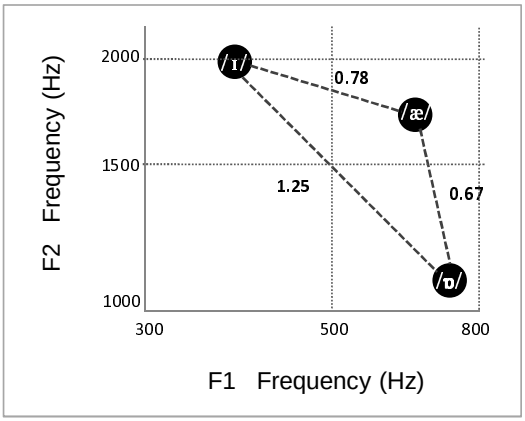


Figure 2: Vowel (F_1, F_2) space with log scaling. The formant frequencies were represented where the axes were scaled logarithmically. /ɪ/ and /ɒ/ differed in both F_1 and F_2, resulting in the most dissimilar tonal coloration; /ɪ/ and /æ/ differed primarily in terms of F_1; /ɒ/ and /æ/ differed primarily in terms of F_2. An Euclidean distance on the log-transformed formant frequencies was calculated between two vowels, marked along the corresponding dashed connection line.

Table 2: Summary of performance and Euclidean distance on the log-transformed formant frequencies. The difference of performances were measured by a Z-score test.

	Accuracy (%)		Z-score Difference	Euclidean Distance (log)
	co-located	separated		
/ɪ/|/æ/	52.38%	63.10%	0.27	0.78
/ɪ/|/ɒ/	65.48%	71.43%	0.16	1.25
/æ/|/ɒ/	55.36%	67.26%	0.31	0.67

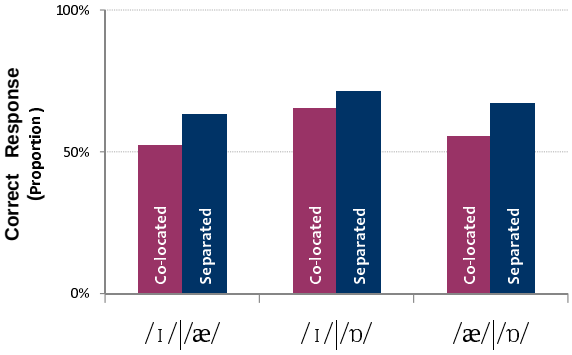


Figure 3: Proportion of correct responses of vowel pairs. More distinct tonal coloration separation of /ɪ/ and /ɒ/ produced a higher proportion of correct responses than for the other two pairs.

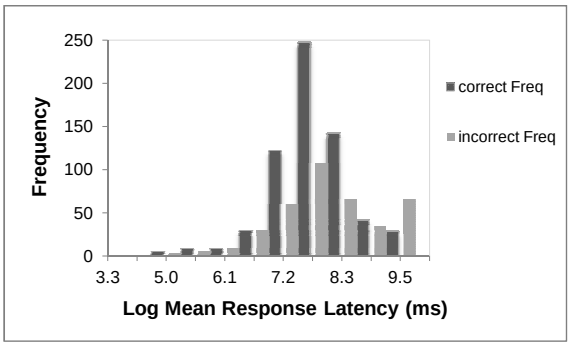


Figure 4: Histograms of log-latencies on correct/incorrect discrimination. The vertical axis represents the corresponding frequency of (in)correct responses. The log-latency on correct responses was mainly located between 7 and 8.3 ms; the accuracy decreased dramatically when latencies were out of this range.

test listed in Table 2. The Z-score differences in performance between the co-located and separated displays related to the Euclidean distance, which was calculated on the differences between the log-transformed formant frequencies. The greater the Euclidean distance was, the smaller was the performance difference between the two displays. Since a greater (F_1, F_2) separation led to a better performance, the result confirmed that the separation in tonal coloration was effective in producing perceptual segregation, and that the greatest effect of the spatial separation on modulation discrimination was found in the stimuli that least differed in tonal coloration.

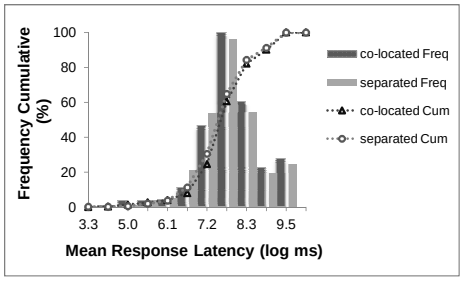
Finally, if both tonal coloration and spatial separation were applied, the segregation of one vowel from the other could be more effective than using only the spatial cue. The most obvious improvement in discrimination performance, due to spatial separation, was found when the vowel sounds differed least in tonal coloration. The bar chart shown in Figure 3 demonstrated that when the vowels were separated in tonal (F_1, F_2) space, the additional cue of spatial separation was able to enhance the perceptual segregation of paired vowels.

As shown above, /ɪ/ and ɒ/ had the most distinct tonal coloration difference due to the greatest Euclidean distance on the log-transformed formant frequencies in the (F_1, F_2) space. The proportion of correct responses for this pair was higher than for the other two pairs (/ɪ/|/æ/ and /æ/|/ɒ/). With both spatial co-located and separated displays, /ɪ/|/ɒ/ produced the highest discrimination accuracy. This reflected that with the spatially co-located or separated display, the further tonal coloration separation of this pair resulted in higher accuracy.

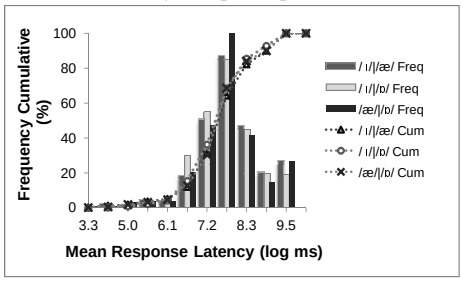
With the comparison of the correct responses in spatially co-located and separated displays, the discrimination performance of /ɪ/|/ɒ/ was poorer for the further separated vowels in the (F_1, F_2) space than for the other two pairs that were closer. When the tonal coloration difference was relatively smaller, the effect of spatial separation presented more obviously.

Measuring response latency

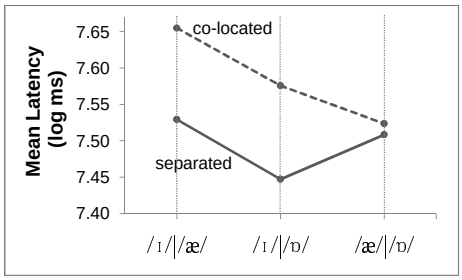
A relationship between log latency and response likelihood was found. The listeners who responded faster gave responses with greater correct-response likelihood. Significant rapid-responders could certainly mask this relationship because rapid-responders gave nearly random responses which, on average, have very low correct-response likelihood (for example, below 7.2 ms in Figure 4). Since the listeners were encouraged to respond within a limited time, it was possible that listeners simply hazarded a guess before time-out. Future study on latency could be conducted to encourage listeners to think about the answer at their own pace and evaluate whether it is better to guess or to omit the item.



(a) Latency on spatial presentation



(b) Latency on vowel pairs



(c) [S×P×D×latency] (correct responses only)

Figure 5: (a)(b)Histograms of latencies: the curves represent the corresponding cumulative frequency distributions; (c) Mean log response latencies.

The mean value of log response latency on individual spatial presentation or on vowel pairs had a relatively similar density distribution (Figure 5 (a) and (b)). The frequency cumulative of the correct response showed that the increase in correct responses was sharp from 1.3 seconds (7.2 ms) to 4 seconds (8.3 ms) with the average latency located in the middle of bins. The frequency of correct responses being longer than 4 seconds after the end of the sounds did not contribute much to the overall accuracy of the score, only an additional 5%.

Figure 5-(c) illustrated the average latency of correct responses on individual pairs of vowels with the two spatial displays. The average latency time on the spatially separated display was less than that on the co-located display for all pairs (Figure 5-(c)). The mean latency on the accurate judgment of the pair /æ/|/ɒ/ seemed not to be influenced by spatial presentation, since the average latency on co-located displays was almost equal to that on the separated display. For all three pairs of vowels, the separated presentation produced a slightly quicker correct judgment, but when an incorrect judgment was made, the separated display took a longer time than the co-located. It seemed that listeners spent a longer time thinking when they felt it was difficult to make a clear discrimination.

It was not surprising that with spatial separation subjects responded slightly more quickly to the pair of /ɪ/|/ɒ/ due to their relatively larger tonal coloration separation (Euclidean distance shown in Figure 2). But it is hard to explain that with the co-located display, the average latency time on /æ/|/ɒ/ was less than that on /ɪ/|/ɒ/, which would have had the smaller latency time if accuracy and latency were closely associated (perhaps because of the larger formant separation). With the separated display, the mean latency on correct judgment was not dropped explicitly from the co-located display. These results are complicated and may be explained by the fact that for the purpose of perceptual segregation, tonal coloration separation made prominent independent contributions to response time; thus, large tonal coloration separation was more likely to lead to quick responses than spatial separation.

CONCLUSION

This study employed a task that required monitoring two simultaneously presented vowel sounds. The effect of spatial separation was measured by comparing modulation discrimination performance in a binaural display when pairs of vowels were from a single location or from different locations (within the head). It found that the ability of listeners to detect a vowel having been modulated was improved when paired vowels were spatially separated (see more details in [11]). The results indicated that spatial separation, produced by sound lateralization associate with interaural level difference, and by the tonal coloration difference, were able to influence the modulation discrimination of two simultaneously presented vowels, in a manner similar to those other central attributes of sound such as frequency difference, intensity and tempo, which were already well known in eliciting stream segregation.

The statistical significance of spatial separation in the divided-attention task supported the conclusion that the spatially separated display of two concurrent vowels can provide a significant benefit in modulation discrimination within concurrently presented vowel stimuli, when a divided attention task is required.

When the performance in identifying pitch modulation in one of two simultaneously presented vowel sounds that were spatially co-located was compared with the performance of those that were separated, a higher accuracy level was found with the spatially separated display. The spatial separation was constructed on an interaural level difference; and timbral difference was constructed on vowel tonal coloration. Before two vowels fused at an identical fundamental frequency, the difference of fundamental frequency and frequency gliding direction could also contribute to the perceptual segregation of two streams. The better discrimination performance with separated display than with co-located display showed that the cost of distributing attention across lateral locations did not outweight its benefit in aiding in segregation and pitch modulation discrimination.

The results also provided evidence that tonal coloration differences of paired vowel sounds reinforced the improvement of discrimination, where a greater Euclidean distance on the log-transformed formant frequencies in the (F_1, F_2) space between paired vowels was able to obtain a higher accuracy of discrimination. The findings indicated that the benefit of spatial separation was greatest when the timbres of paired streams were least distinct (shortest Euclidean distance on the log-transformed formant frequencies).

Response latency did not demonstrate an association either with spatial presentation nor with tonal coloration difference, but showed a common correlation with accuracy - longer latencies associated with lower accuracy score.

REFERENCES

[1] Douglas S. Brungart, "Evaluation of speech intelligibility with the coordinate response measure," *The Acoustic Society of America*, vol. 109, no. 5, pp. 2276–2279, 2001.

[2] Douglas S. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," *The Acoustic Society of America*, vol. 109, no. 3, pp. 1101–1109, 2001.

[3] Douglas S. Brungart, Mark A. Ericson, and Brian D. Simpson, "Design considerations for improving the effectiveness of multitalker speech displays," in *the 8th International Conference on Auditory Display*, Kyoto, Japan, 2002.

[4] Albert S. Bregman, *Auditory scene analysis*, The MIT Press, Cambridge, MA, USA, 1990.

[5] John F. Culling and Quentin Summerfield, "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," *Journal of Acoustical society of America*, vol. 98, no. 2, pp. 785–797, 1995.

[6] Harold L. Hawkins and Joelle Presson, "Auditory information processing," in *Handbook of perception and human performance*, KR Bott, L Kaufman, and JP Thomas, Eds., vol. II of *Cognitive processes and performance*, pp. 26–64. Wiley, New York, 1986.

[7] Barbara G. Shinn-Cunningham and Antje Ihlefeld, "Selective and divided attention: Extracting information from simultaneous sound sources," in *10th International Conference on Auditory Display (ICAD'04)*, Sydney, Australia, 2004.

[8] Gordon E. Peterson and Harold L. Barney, "Control methods used in a study of the vowels," *Journal of Acoustical Society of American*, vol. 24, no. 2, pp. 175–184, 1952.

[9] Udo Zölzer, Ed., *DAFX - Digital Audio Effects*, John Wiley & Sons, 2002.

[10] Miller Puckette and David Zicarelli, "Max/MSP," 1990-2010.

[11] Hong Jun Song, *Evaluation of the effects of spatial separation and timbre on the identifiability of concurrent auditory streams*, Ph.D. thesis, University of Sydney, submitted on March, 2010.