Developing vowel mappings for an interactive voice synthesis system controlled by hand motions

Karl Nordstrom[1], Sidney Fels[1], Cameron Hassall[1] and Bob Pritchard[2]

[1]Media and Graphics Interdisciplinary Centre

[2]School of Music

University of British Columbia

Marguerite Witvoet (DIVA) in performance of *What Does A Body Know?*

Objective

Develop a map of vowel targets in space for gesture-based speech synthesis. The overarching goal is to make the synthesized speech (and singing) sound as natural as possible.

Speech synthesis system

A **DI**gital **V**entriloquized **A**ctor (DIVA) is a wearable gesture-controlled speech synthesis system. Hand gestures control a formant synthesizer for the creation of vowel and consonant sounds. The actor creates vowels by moving their right hand in space. The vowels are located on a horizontal plane with the closest vowel target determining the formants for synthesis.

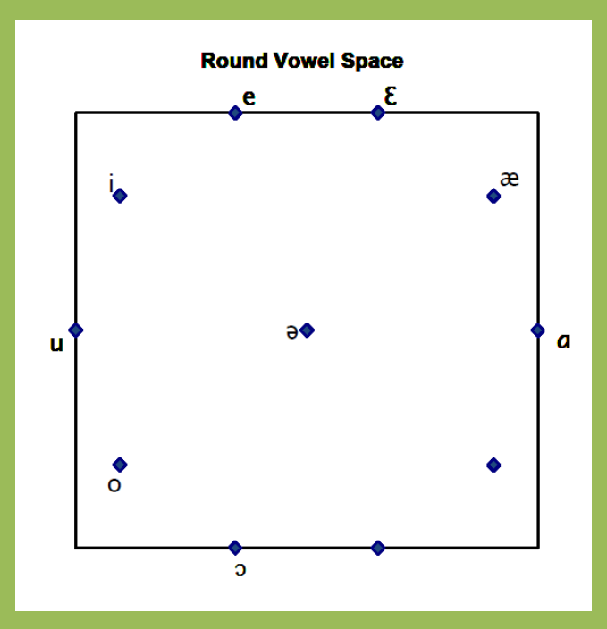
Previous method for defining vowel target locations

Previous iterations of the DIVA enabled the performer to choose the location of the vowels in the horizontal plane; however, it was difficult to locate the vowel targets in a uniform and orderly way. With little visual or tactile feedback, the vowel targets were poorly located. In some cases, the targets were almost on top of each other, making it difficult to find the desired vowel. Alternately, the targets were widely spread, making them difficult to reach. This limited the rate at which speech could be synthesized.

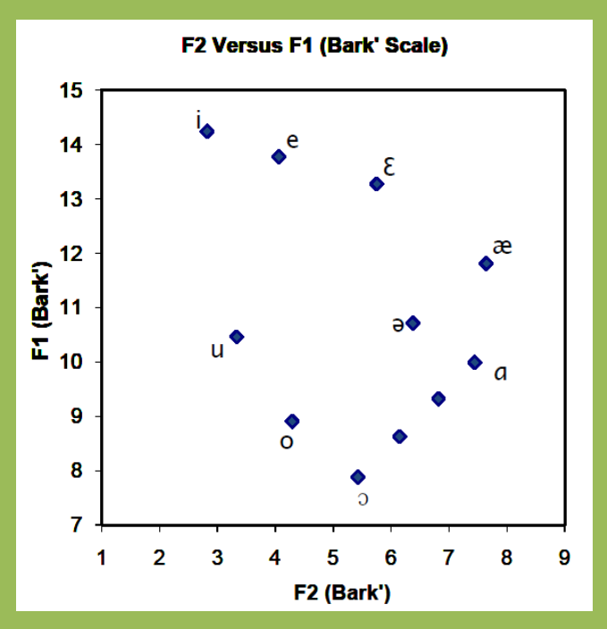
Two strategies for organizing the vowel layout

Two vowel layouts were explored.

1. Vowel targets evenly distributed in a circle to make the targets easier to find.

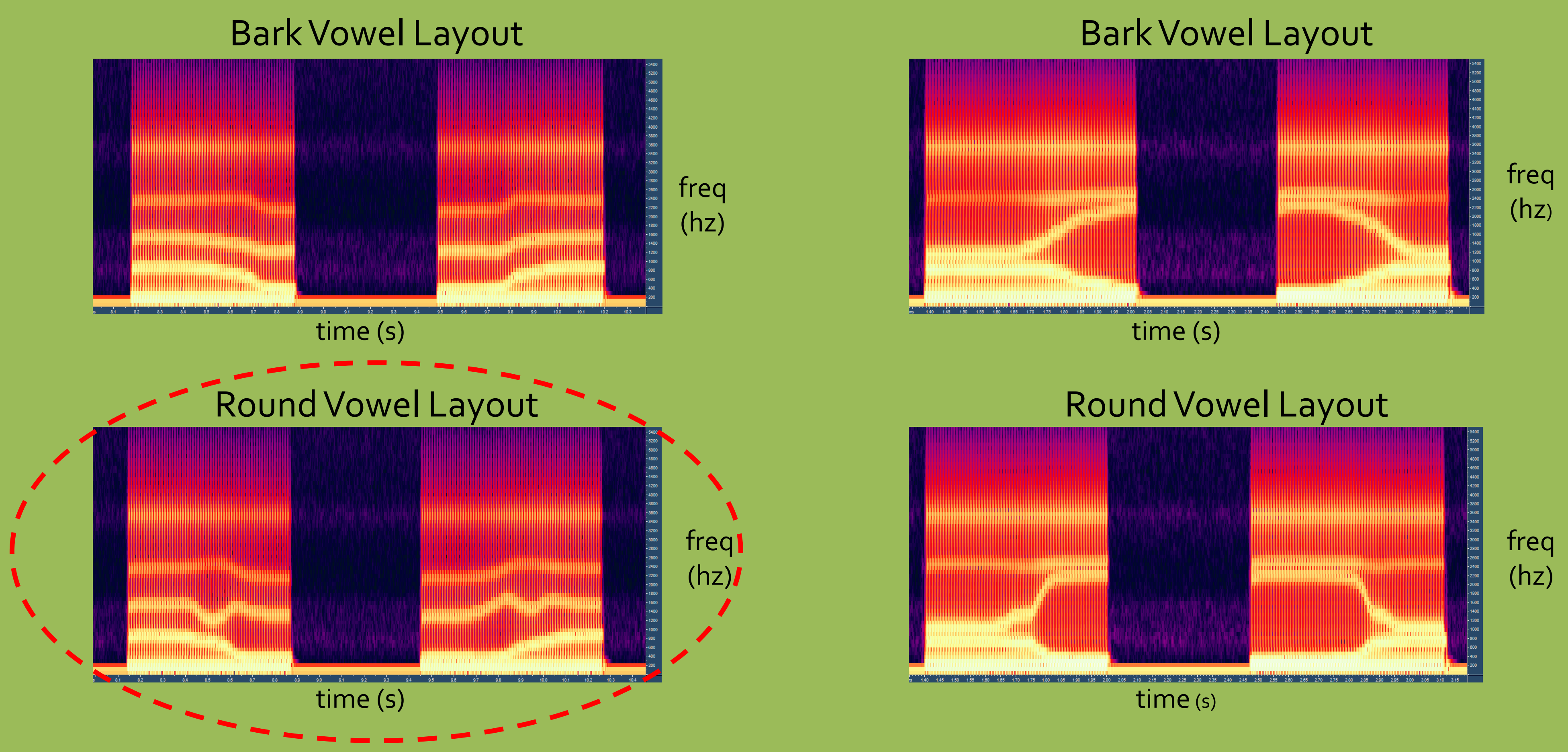


2. Vowel targets arranged according to Bark warped formant frequencies: F2 vs. F1.



Evaluation process

Straight-line hand motions were made through the two vowel layouts. The hand motions started at the same vowel target in both layouts and took a direct path to the same ending target. While carrying out these motions, we listened to the vowel transitions and recorded the results. See the spectral plots below.

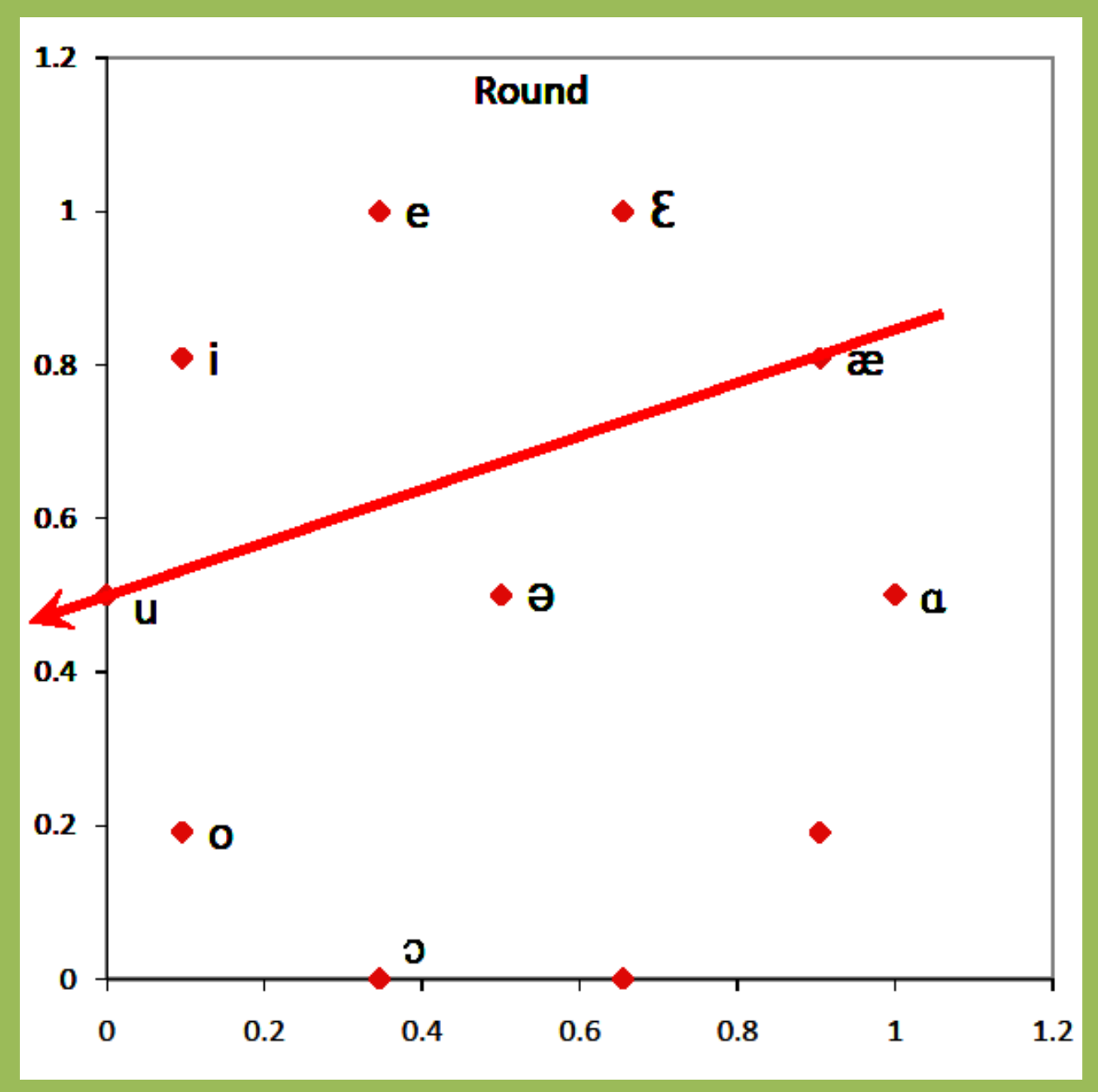


Results

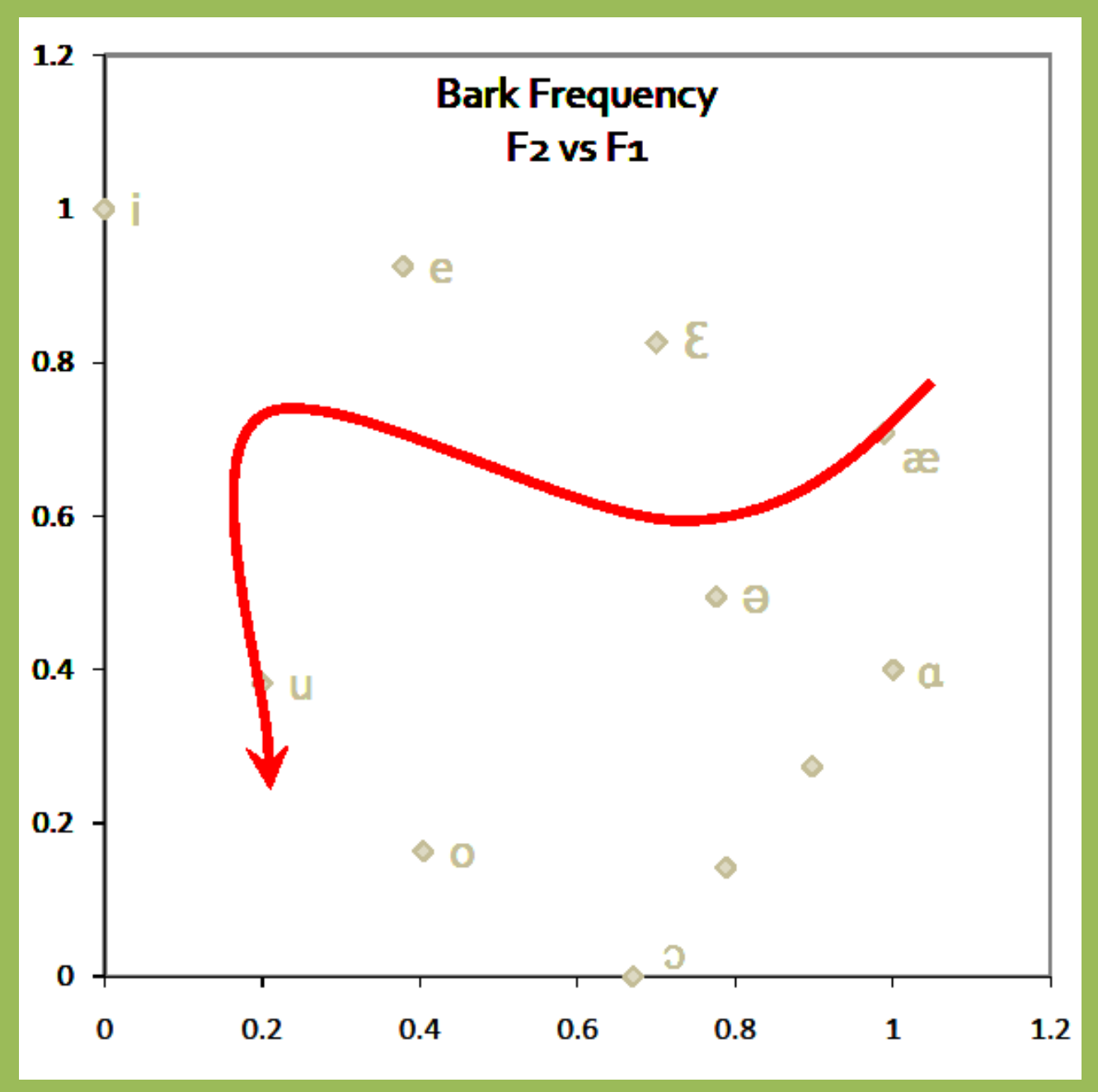
The Bark vowel layout resulted in formant transitions that were relatively smooth as seen in the plots above. In contrast, the round layout resulted in formant transitions that were not monotonic; the F1 and F2 pitch contours varied up and down as the hand carried out the straight-line motions. These variations were perceived as uneven formant transitions.

Explanation

As the hand moves through the vowel space, the formants for synthesis are determined by the closest vowel target. In the Bark vowel space, hand position corresponds to frequency, and a smooth motion through the vowel space results in a correspondingly smooth motion through the frequency space. In the round vowel space, the relationship between hand position and frequency is not as strong. As a result, some steady hand motions result in formant frequencies that vary up and down. One such example is shown in the figure to the right.



This straight line motion
Through the round vowel space…



… results in an F2 formant trajectory that
goes down, then up, then down again
(see circled spectral plot above)

Conclusion

Organizing the vowel space according to convenience can lead to uneven formant transitions. Instead, lay out the vowel space with a direct relationship between hand position and frequency. This will provide the performer with more direct and intuitive control over their instrument.

Social Sciences and Humanities Research Council of Canada · Conseil de recherches en sciences humaines du Canada · Canada · ICICS · MAGIC · EMILY CARR UNIVERSITY · UBC MUSIC · NSERC CRSNG · Canada Council for the Arts · Conseil des Arts du Canada