Vowel Shapes: An Open-Source, Interactive Tool to Assist Singers with Learning Vowels

Cynthia Ryan
University of Rochester
Rochester, NY, United States
cyndijo50@gmail.com

Katherine Ciesinski
Eastman School of Music
Rochester, NY, United States
KCiesinski@esm.rochester.edu

Mohammed (Ehsan) Hoque
University of Rochester
Rochester, NY, United States
mehoque@cs.rochester.edu

ABSTRACT
The mastery of vowel production is central to developing vocal technique and may be influenced by language, musical context or a coach's direction. Currently, students learn through verbal descriptions, demonstration of correct vowel sounds, and customized exercises. Vowel Shapes is an interactive practice tool that automatically captures and visualizes vowel sounds in real time to assist singers in correctly producing target vowels. The system may be used during a lesson or as a practice tool when an instructor is not present. Our system's design was informed by iterative evaluations with 14 students and their vocal professor from the Eastman School of Music, University of Rochester. Results from an exploratory evaluation of the system with 10 students indicated that 70% of the participants improved their time to reach an instructor-defined target. 90% of the students in the evaluation would use this system during practice sessions.

Author Keywords
Auditory I/O and Sound in the UI; Visualization; Interdisciplinary Design; Singing; vowels; interactive.

ACM Classification Keywords
H.5.m Information interfaces and presentation (HCI)

INTRODUCTION
Let's consider Maria, a student majoring in vocal performance at a music school. Maria takes individual lessons with an applied instructor once a week. The feedback Maria receives from her instructor is subjective and only available when the instructor is physically present. Otherwise, she spends most of her singing time practicing alone using her memory and notes scribed during her individual lesson. The practice methodology is basic and limited. Thus, when Maria practices alone, she cannot verify whether she is singing correctly. Could an interactive system with real-time visual feedback allow her to determine whether she is singing with the correct technique? Could this system supply information to the instructor that could then be shared with the student at the next lesson? Could the system be designed to help Maria practice anywhere and at any time?

UbiComp '15, September 07-11, 2015, Osaka, Japan
 ISBN 978-1-4503-3574-4/15/09 $15.00
DOI: http://dx.doi.org/10.1145/2750858.2805829

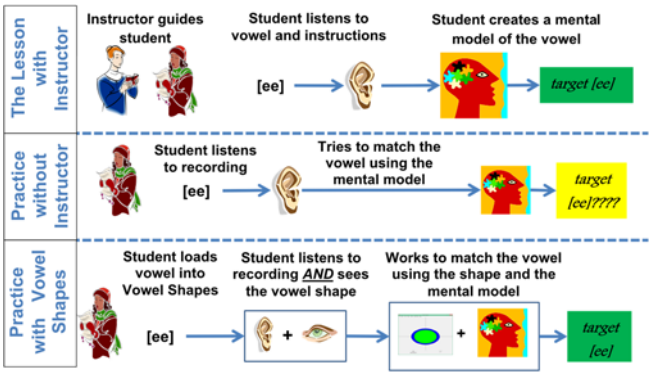

Figure 1: Learning a vowel – current and Vowel Shapes

In this paper we propose, develop and evaluate a dynamic interactive practice tool for vocal students. Vowel Shapes is a tool that automatically translates a vowel sound to a visual shape. When students practice singing a vowel, they continuously make minute physical adjustments to match an internalized vowel sound. They are currently limited to their memory of the target vowel and their aural, kinesthetic, and subjective abilities to decide when the vowel has been matched. In our work, we hypothesized that a real-time visual vowel shape could significantly improve the students' ability to master a vowel. The visual vowel shape would continually provide feedback to students as they make subtle vocal changes to the vowel (Figure 1). Ideally as students vocalize the vowel they should be able to understand how their physical adjustments map to vowel shape changes. The music student Maria should be able to say to herself, "*When I move my tongue upwards, the vowel sound is made brighter and the shape becomes flatter*."

Vowel Shapes captures a live audio feed of a sung vowel and represents the vowel visually based on analysis of the vowel audio. We informed the design consideration and evaluation of Vowel Shapes by collaborating with a vocal instructor and volunteer vocal students. We conducted a number of design iterations to improve ease of use and technical accuracy. Based on design findings, we proposed a visualization algorithm and a normalization scheme to allow for comparison of individuals vocalizations.

Our technical implementations had the following research challenges. Visualizations needed to be represented using information that a student was familiar with. A target visualization would need to inform the students on how to match the target vowel. *How could we develop a visualization that informed the singer about vowel production – tongue position and shape of mouth?* Each target vowel would need to be practiced singly and by a specific individual. *What methods would allow the analysis of a single vowel?* Each student would be at a different level of education, implying varying teacher expectations. *How would we allow for an instructor to assist the student to overcome previous erroneous training or refine a vowel?*

BACKGROUND RESEARCH

Teaching vocal students how to produce vowel sounds appropriate to various languages is an important aspect of applied study [4]. Current methods include instructor or vocal coach-based learning. For practice, students may access handwritten notes and/or recordings from their lessons. Other methods for practicing vowels may include watching laboratory videos of the vocal tract or reviewing written instructions available through scientific publications, pedagogy books or web sites [6, 9, 10, 19, 21]. Many commercial software tools exist that automatically provide spectrograms or other visualizations of the audio [6, 3]. While they reflect the student's vowel production in real time and allow the instructor to provide feedback, these tools lack a target visualization. During individual practice, the student must still rely on the memory of visual images and the instructor's advice.

Since the late 1980's there has been research in the area of singing and real-time visual feedback applications [22]. A review of such applications [11] provides some insight into why visualization improves the learning experience. Domains such as pronunciation training, second language acquisition, and speech training for those with speech disorders or other challenges have used interactive visual feedback successfully [1, 8, 14, 20].

One difference between the singing and the speaking domains is the specificity of the vowel required when singing. A spoken vowel may be understandable to a native speaker even when formed generically by a non-native speaker. A sung vowel, for aesthetic reasons, needs to precisely evoke the language of the written text. Further, it must permit proper emission of tone on the notated pitches and be replicated consistently in varying verbal and musical contexts. Thus correct vowel production is a complex coordination of mental intention and physical activity that, until mastered, is only partially verifiable by the student's auditory acuity.

CONTEXTUAL INQUIRY

To inform our design process, we interviewed a professor, with over 30 years of professional singing experience and 5 vocal students. We provided a survey (found here: http://tinyurl.com/kt4stpj) to which 9 students responded. The survey questions gathered information on three topics: *(i)* the student's understanding and opinion of their current procedure to learn vowel sounds; *(ii)* the student's familiarity with spectrograms and their potential usefulness in a lesson or practice environment; *(iii)* the type of visualization that would be understood by the student.

Our review of practice methodologies and interaction with vocal students and the professor revealed the following: *(i)* a student leaves a voice lesson with only written notes, possibly an audio recording of the session, and published reference material; *(ii)* mastery of a vowel consistently lies with instructor feedback and with the students' memory of the instructors' subjective recommendations; *(iii)* music students are not familiar with spectrograms and would require a steep learning curve to interpret them; *(iv)* to the best of our knowledge, no tool exists, other than spectrograms, for students to visually compare their vowel sound during practice with a target image.

Based on our contextual inquiry, we aimed to achieve the following objectives in our prototype: *(i)* recognize vowel sounds in real-time and provide visual feedback; *(ii)* produce a target visualization based on an instructor's vowel audio; *(iii)* produce visual feedback to indicate how well the target is being matched; *(iv)* the target vowel visualization must be easy to understand by the student during individual practice sessions.

In order to achieve our goals, we had to solve the following technical challenges and design considerations. First the visuals need to be based on a common definition of how a vowel is produced—a specific articulatory and vocal tract shape. Second, the instructor's sung vowel model had to be captured, analyzed, characterized and saved. Third, the characterization needed to accommodate differences in gender and pitch.

VOWEL CHARACTERIZATION, SHAPE AND AUDIO

Vowels may be characterized by using frequencies, or formants [14], filtered from audio input. Formants (F_n) are "the frequencies which are most successful in travelling through the vocal tract" [7] and can quantitatively distinguish a vowel sound. A singer tracks the physical changes that produce formant modification. A closed or open oral cavity affects the frequency of F_1. Likewise, the vowel is considered front or back depending on tongue position, which modifies F_2. Further differences made by rounding or lateral spreading of the lips influence F_3. Vowel formants will vary based on an individual's gender and vocal tract, and while they will be within a given range, formants may not be compared directly. Normalization of the formants makes it possible to compare vowel sounds across physiological differences [18] and gender [2]. To characterize a single vowel in real time requires an intrinsic normalization algorithm. Vowel Shapes uses Barks Formant Normalization (BFN) [16, 17] to transform formants F_1, F_2, F_3, into normalized Z values; Z_i can be expressed by:

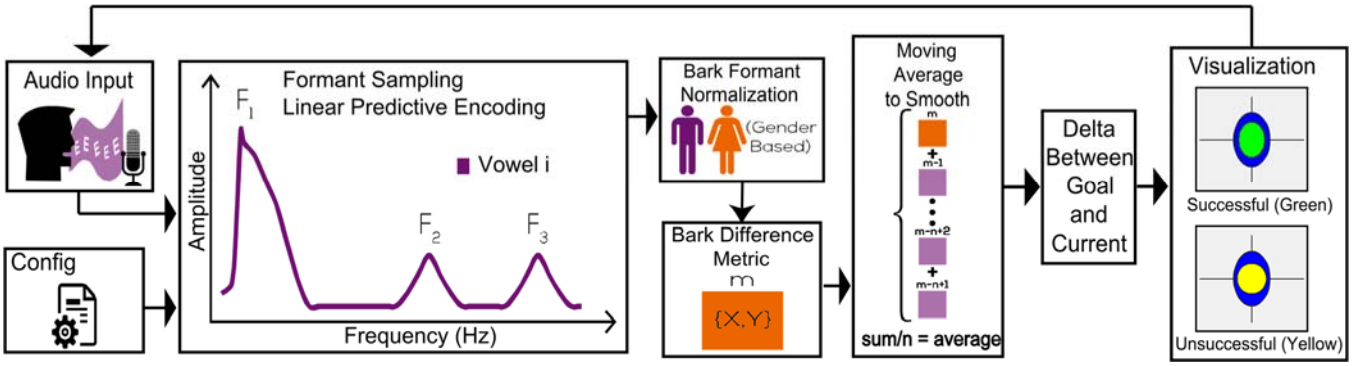

Figure 2: The design of Vowel Shapes

$$Z_i = \frac{26.81}{\left(1+\frac{1960}{F_i}\right)} - \Delta \quad \text{Male } \Delta = 0.53 \quad \text{Female } \Delta = 1.53$$

The BFN removes the gender disparity between singers. The Bark Difference Metric (BDM) removes pitch disparity and permits the comparison of vowels independent of pitch. A modified BDM metric [16, 17] transforms the three Z values into two dimensions, X and Y.

$$X = Z_3 - Z_2 \qquad Y = Z_3 - Z_1$$

The parameters are the width and height of the vowel. X may be understood by a singer as to how bright the vowel sound is and how high the tongue should be positioned to achieve the target width. A wider target shape would translate to a higher, forward tongue position producing a brighter vowel. Y may be understood as how rounded the mouth should be and how much the jaw should be opened. A taller shape would translate to a more open, sound and indicate the pharyngeal position for back vowels. X and Y thus define specific vowels and are translated by singers into various tongue, jaw and lip positions. Based on our contextual inquiry three initial vowels shapes were selected: an ellipse, a triangle and an {X, Y} grid.

A "matching" algorithm is implemented to indicate an acceptance or tolerance for when a vowel is matched – when met the student's vowel shape color is changed to green indicating an acceptable vowel. The tolerance for an acceptable vowel is based on an application configuration value that may be set by the instructor for each student based on their instructional level. The vowel shape-matching algorithm is dependent on the shape – area for a triangle, minor/major axis ratio for an ellipse, and Euclidean distance for a graph (see Figure 2).

Vowel Shapes requires real time audio analysis capability for recording, saving, playing and analysis of audio, specifically formants. We use the freely available sound toolkit called Snack [13]. A linear predictive coding (LPC) algorithm is used to find the formants. A sampling rate of 44.1 KHz is used [5]. The analysis uses a window of 0.049 seconds over a sample of 0.186 seconds and an LPC analysis order of 12. To smooth the shape transition, a moving average of the Z values is used to update the {X, Y} values and then the display.

IN-LESSON EVALUATION

The prototype was evaluated during individual lessons. Our evaluation process allowed the instructor and student to work together with the tool as they would in a typical lesson. The student was then assessed on their ability to repeat the lesson content – both with and without the tool. Our evaluation aimed at two areas of investigation: *(i)* would the tool help the students realize the target when instructor feedback was not available? *(ii)* would the students find value in the tool for individual practice?

Evaluation Process and Participants

The evaluation group consisted of 10 university vocal students and their professor. The session would start with the instructor singing a target vowel while the student listened, allowing the tool to record and complete an analysis on the target vowel audio. The student then completed two practice attempts to correctly match the target vowel. The practice attempts included both the visualization and instructor feedback. In some instances, the professor made multiple recordings before achieving an acceptable model vowel shape. The instructor was ok with this, as it is normal to require several takes in any recording situation to make an optimum recording.

Evaluation timings were also completed using a diphthong (combination of two vowels in a syllable, such as coin, loud, boy) of the instructor's choosing. The instructor would demonstrate the diphthong for the student, again allowing for recording and analysis by the tool. Two assessments were then completed – one timing with the tool where the professor needed to concur with the tool on a successful vowel, one timing without the tool where the professor indicated success. During both assessments the professor was only an observer until the student succeeded in producing a good vowel. The students were split into two groups. Group A used the tool for their first assessment. Group B was first assessed without the tool. This counterbalancing design methodology allowed us to control order effects. The evaluation time without the tool should approximate that of a practice session. A follow up survey (available at http://tinyurl.com/pen9hcm) was provided after each session.

Evaluation Results and Questionnaire

Seven out of the 10 students improved the time required to match the evaluation diphthong using Vowel Shapes. Figure 3 provides a summary of our results. Of the 3 students that did not improve 2 of the students were graduate students that the professor described as "impatient with themselves". The sudden change of modality for perfecting the vowel – from hearing themselves and watching the professor for visual cues to placing trust in the visualization – likely caused a longer time to match the vowel. For the third student, a freshman, we theorize that the student may require more information on the acoustics of vowel production to become adept at reacting to a visual representation of those parameters.

Only one student, a doctoral vocal major with significant background in business, preferred the graph representation, while the rest used the ellipse.

The geometric mean for with the tool is 13.78 seconds, without the tool it is 24.01 seconds. The difference between with-tool/without-tool is significant ($F[1,8]=6.278$, $p<0.05$). We may consider a moderate skill transference effect. ANOVA analysis indicates that counterbalancing within the evaluation worked ($F[1,8]=0.106$ not significant).

The follow-up questionnaire asking the students about their experience with the tool revealed encouraging results. 90% of the students found Vowel Shapes to be very useful when trying to find the target vowel and would use this tool in their studies and practice. 80% of the students said the visualization helped to show that they were consistently on pitch. To clarify this reference to pitch, the professor explained that the visualization helped them fine tune the formants through a change in resonance, not frequency.

While well received by the students, written comments from the survey provided more information on how to further improve the tool. Some comments were, "*This works on basic shapes, but might be frustrating for fine tuning because it is visual instead of aural* ", "*I just want to better understand how the target works*", "*It would be helpful if it were easier to see why the vowels didn't match. The program told me when my vowels weren't accurate, but I couldn't tell how to fix the vowel to make it match*".

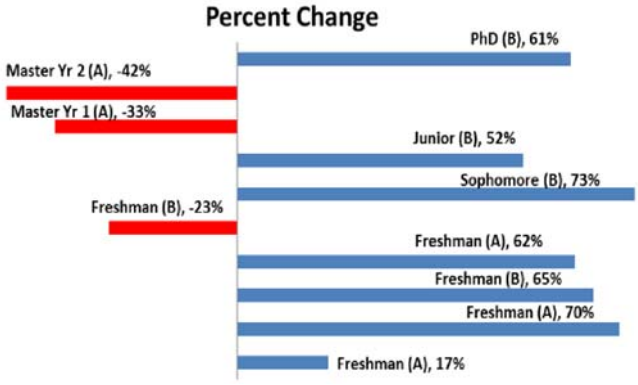

Figure 3: Student results – (Tool – w/o Tool)/Professor time

DISCUSSION

The initial evaluation of Vowels Shapes was positive and encouraging for future improvements, such as ensuring that formants are found accurately. The current LPC algorithm uses autocorrelation to determine the formants. While the length of the audio sample is large compared to the LPC order used, it may still allow for spectral line-splitting. Changing the algorithm to use modified covariance or Line Spectral Frequencies [12] would remove the potential of finding false formants.

It is possible that the vowel is not internalized by the student because it was visualized. As a consideration, a second study would retest the students on the vowel shape after an interval of rest time or after engaging in some other task. Most singers are adept and trained to recall extremely precise vowel shapes and many other physical aspects of tone production, working largely by ear and sensory memory. It is important to remember that the vocalists are attuning their articulators not only to external vibrations carried to the ear, but also to internal bone-conducted signals to the ear and other sensory information regarding positions of the larynx, pharynx, jaw and tongue. The addition of a visual cue provides a quicker means to get the desired result. The vocal professor commented *"... it is my experience (from working with a mirror, for example) that such aids do allow for retention and internalization. ..., repetition and practice are required to fully integrate and master a newly introduced vowel shape*."

An improved user interface is another area for improvement. The current vowel visualization permits the attempted vowel shape to completely overlay the target vowel shape potentially frustrating the student. Possible solutions may include using a transparency factor with the attempted vowel shape or introducing a third color when the attempted vowel overlaps the target vowel.

CONCLUSION AND ACKNOWLEDGEMENT

We demonstrate that when using Vowel Shapes 70% of our participants are now able to master more vowels in less time, increasing their overall productivity. The results also indicate that the students found the system engaging and would be eager to use the tool during their practice sessions. Vowel Shapes is at http://tinyurl.com/ky7zlqp.

Future instantiations could provide sharing of vowels among users and sharing of study results between instructors and diction coaches. Storage and sharing could be useful to expedite problem-solving and correct long-standing inaccurate concepts. Further, vowel analysis provides objective measures of accuracy that otherwise could slip off the radar of both teacher and student as study progresses. We believe our findings point to the usefulness of such tools for independent learning and fundamental skill building for singing in any style or language.

This paper evolved from a project collaborating with Nate Buckley, Josh Bronstein, Tait Madsen and Veronika Alex.

REFERENCES

1. Arends N. & Povel D.J. (1991). An evaluation of the visual speech apparatus. *Speech communication 10,* 405–414.
2. Bladon, R. A., Henton, C. G., &Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language & Communication, 4*(1), 1984, 59-69. doi: 10.1016/0271-5309(84)90019-3
3. CantOvation Ltd. (n.d.). Sing&See. Retrieved from http://www.singandsee.com
4. Coffin, B. (1982). *Phonetic readings of songs and arias* (2nd ed., with rev. German transcriptions.). Metuchen, N.J.: Scarecrow Press.
5. Colletti, J. (February 4, 2013). The Science of Sample Rates (When Higher Is Better—And When It Isn't) [Blog Post]. Retrieved from http://trustmeimascientist.com/2013/02/04/the-science-of-sample-rates-when-higher-is-better-and-when-it-isnt/
6. Diction for Singers (n.d.). Listening Labs, Retrieved from http://www.dictionforsingers.com/listening-labs.html
7. Doscher, B. M. (1994). *The Functional Unity of the Singing Voice: Second Edition*. Metuchen, N.J., & London: The Scarecrow Press, Inc.
8. Dowd A., Smith J. & Wolfe J. (1998). Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real-time. *Language and Speech 41,* 1–20.
9. Goldes, J. (n.d.). Vowel Sounds [Instructional]. Retrieved from http://www.thedialectcoach.com/index.php?option=com_content&view=article&id=380&Itemid=119
10. Hall, D. (n.d.). Interactive Sagittal Section [Interactive Demonstration]. Retrieved from http://smu-facweb.smu.ca/~s0949176/sammy/
11. Hoppe D., Sadakata M. & Desain P. (2006). Development of real-time visual feedback assistance in singing training: a review. *Journal of Computer Assisted Learning, 22*, 308–316
12. Kabal, Peter, and Ravi Prakash Ramachandran. "The computation of line spectral frequencies using Chebyshev polynomials." Acoustics, Speech and Signal Processing, IEEE Transactions on 34, no. 6 (1986): 1419-1426.
13. Kåre Sjölander (January 23, 2006). The Snack Sound Toolkit. Retrieved from http://www.speech.kth.se/snack/
14. Miyauchi, M., Kimura, T., & Nojima, T. (2013). A tongue training system for children with down syndrome. *In Proceedings of the 26th annual ACM symposium on User interface software and technology (UIST '13)*. ACM, New York, NY, USA, 373-376. DOI=10.1145/2501988.2502055 http://doi.acm.org/10.1145/2501988.2502055
15. O'Connor, K. (January1, 2011). Vowels, Vowel Formants and Vowel Modification [Instructional]. Retrieved from http://www.singwise.com/cgi-bin/main.pl?section=articles&doc=VowelsFormantsAndModifications&page=2
16. Syndal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of Acoustic Society of America*, 79, 1086. doi:10.1121/1.393381
17. Thomas, E. R., & Kendall, T. (April 11, 2010). NORM's Vowel Normalization Methods (v. 1.1) [Documentation]. Retrieved from http://ncslaap.lib.ncsu.edu/tools/norm/norm1_methods.php
18. Traunmüller, Hartmut (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88, 97-100 (1990), DOI:http://dx.doi.org/10.1121/1.399849
19. University of North Texas (July 3, 2014). For Singers [Reference]. Retrieved from http://music.unt.edu/voice/singers
20. Vicsi, K., P. Roach, A. Öster, Z. Kacic, P. Barczikay, A. Tantos, F. Csatári, Zs Bakcsi, and A. Sfakianaki. "A multimedia, multilingual teaching and training system for children with speech disorders." International Journal of speech technology 3, no. 3-4 (2000): 289-300.
21. Vocalist.org.uk (n.d.). Spectographs & Vocal Software [Reference]. Retrieved from http://www.vocalist.org.uk/vocal_software.html
22. Welch G.F. (1985). A schema theory of how children learn to sing in tune. *Psychology of Music, 13*(1), 3–18. doi: 10.1177/0305735685131001