Say ‘Aaaaa’
Interactive Vowel Practice for Second Language Learning

Preben Wik, David Lucas Escribano

Department of Speech Music and Hearing
KTH, Stockholm, Sweden
preben@speech.kth.se, davidle@kth.se

Abstract

This paper reports on a system created to help language students learn the vowel inventory of Swedish. Formants are tracked, and a 3D ball moves over a vowel-chart canvas in real time. Target spheres are placed at the target values of vowels, and the students’ task is to get the target spheres. A calibration process of capturing data from three cardinal vowels is used to normalize the effects of different size vocal tract, thus making it possible for people to use the program, regardless of age, size, or gender. A third formant is used in addition to the first and second formant, to distinguish the difference between two Swedish vowels.

1. Introduction

Computer assisted pronunciation training (CAPT) could ideally incorporate a wide variety of exercise on different levels. CAPT can fill an important gap in language learning on articulatory exercises on phoneme level, segmental difficulties on syllable and word level, prosodic features on sentence level and conversational skills on discourse level. CAPT is a highly language specific discipline, because the difficulties a language student is likely to encounter, are precisely those aspects of the target language (L2) that differs from the native language (L1). There is no one-size-fits-all in CAPT, where the same sets of exercises apply to all languages. One such language specific exercise for language students to master is the vowel inventory in a new language.

Vowels are used in all languages in the world, but there is a wide variation on how many, and which vowels are used. The range goes (according to Indopedia [1]) from as few as two, (Abkhaz, Xoo) to the world record, held by the Sedang language (a relative to Vietnamese) where they distinguish between 55 different vowels. Someone with Abkhaz as L1 trying to learn Sedang, will have quite a challenge trying to acquire the new vowel system, whereas the other way around is likely to be trivial.

The most common vowel system among languages contains no more than five vowels, although some of the most widely spoken languages have larger vowel inventories, like for example English with 14-16. For many language learners it is thus a difficult and important aspect of language learning to gain insights into a larger or different set of vowels than ones L1. This paper presents an approach for language students to learn the vowel inventory of Swedish.

1.1. The Swedish vowel system

Swedish is notable for having a large vowel inventory, with 17-22 different monophthongs, depending on how one count. The orthographic base is three back vowels (/O/, /Å/ /A/), three front vowels (/I/, /E/, /Ä/), and three rounded front vowels (/Y/, /U/, /Ö/). These occur in pairs of long and short, with a quality difference apart from the length, thus 18 vowels in all. Because of the small difference in vowel quality between short /Ä/ and /E/ in standard Swedish, it is sometimes counted as the same, thus 17 vowels. Changes in vowel quality in many dialects (including standard Swedish) due to the vowels position in a word, makes the count come up to 22 [2]. Many L2 learners of Swedish find the vowel system very complex and difficult to master. A CAPT system allowing language learners to practice this in a self paced manner, on their own computer at home, is therefore an attractive and potentially valuable asset.

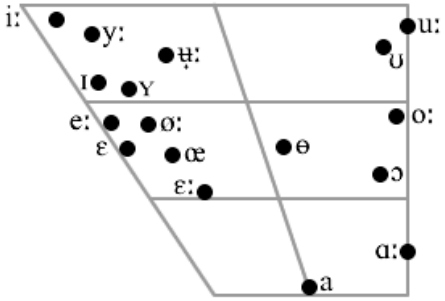

*Figure 1:*A vowel chart (with IPA notation) of the Swedish vowel system, with 17 monophthongs.

2. Method

2.1. Formants and vowel charts

Formants are concentrations of acoustic energy around particular frequencies in the speech wave, and are an effect of resonance in the vocal tract. These characteristic harmonics can be used to identify vowels. The first formant (F1) corresponds to the front-back dimension and the second formant (F2) to the open-closed dimension of a vowel. They map nicely on the traditional vowel chart (as in figure 1), when F2 is plotted in negative direction.

Most vowels can be separated by the F1-F2 plane alone, but there are exceptions. Most notably for this paper, the distinction between Swedish /I/ and /Y/ lies in changing the lips from a wide spread position to a pouted. This change will acoustically be noted by a shift of the third formant (F3). To cover the Swedish vowel inventory, tracking F1 and F2 is thus not enough, but also F3 must in some cases be taken into account.

The size of the vocal tract affects the formant values so that a man, woman, or child saying the same vowel will get different formant values. Fant [3] drew attention to the fact that the relationship between male and female formant frequencies cannot be described by uniform scaling. This non-uniform scaling of the vocal tract means that if vocalizations

of people with different height, gender, or age are to be compared using the formant frequencies, a normalization method must precede the comparison. Our solution to this is to make use of the cardinal vowels.

2.2. Cardinal vowels as calibration points

A cardinal vowel is a vowel sound produced when the tongue is in an extreme position, either front or back, high or low. Since the cardinal vowels are extreme points of articulation, they mark the outer rim of an individual's vowel space and all other vowels are lying within this space. If we are able to elicit some of these cardinal vowels from the users, they can be used as reference points, by scaling the canvas to fit these points. All target vowels can then be measured in relative distances from them.

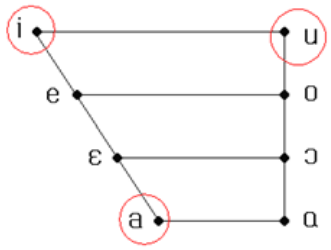

Figure 2: The cardinal vowels, with the three corner vowels used for calibration marked with a circle.

Three cardinal vowels, the corner vowels (see figure 2) are elicited from the user by an initial interactive calibration phase using an embodied conversational agent (ECA) (see figure 3). The ECA starts by giving a short explanation to why this is necessary in order to get accurate measurements. The ECA then proceeds to elicit each individual corner vowel.

The corner vowels are given articulatory definitions. [i] is produced with spread lips, and the tongue as far forward and as high in the mouth as is possible. [u] is produced with pursed lips, as in a whistle, and the tongue as far back and as high in the mouth as possible. [a] is produced with an open mouth, and with the tongue as low as possible, as when going to the dentist, saying Aaaaaa.

There is a bootstrapping problem involved in the calibration phase. A human being can *hear* if a vowel is mispronounced. Our software will measure the formant frequencies, and normalize them relative to a person's corner cardinal vowels. If the cardinal vowels are off, (or from a different person) the analysis of the software will also be off. Since the formant values are based on the size and shape of every individual's vocal tract, we cannot know what the expected values should be. If a user for some reason fails to do the correct articulatory movements, as instructed by the ECA, we could end up with a canvas that is too small, or skewed, and that would affect the quality of the analysis.

We have made some efforts to eliminate this potential problem. First of all by making the ECA's explanations as clear as possible, coaching the student into stretching his/ her personal vowel canvas as much as possible. After the initial elicitation of the corner cardinal vowels, the ECA asks the student to say three easy words, /BI:/ /BO:/ /BA:/, containing the easiest, most common vowels. These words are then run through a forced alignment [4], the center piece of the vowel is cut out, and the formant extraction method is applied on each of the vowels respectively. If the F1,F2 coordinates fit in the expected areas with a reasonable accuracy, the calibration phase is finished. If not the whole calibration phase is repeated. Although this method worked successfully on all test-subjects in the experiment described in section 3, (with some of them doing a second calibration), we will not know if this method is adequate until the system has been tried on a larger set of students.

2.3. Software

The main part of the software is a 3D canvas with a vowel chart, and a ball. When a language learner speaks into a microphone the ball moves around on the canvas, and will in real time move to the place on the vowel-chart canvas that corresponds to the vowel uttered by the student, thus giving immediate feedback on the consequences of his/her articulatory movements. The movements of the ball are accomplished by extracting the formants of the acoustic signal, and using the values of the first and second formant as coordinates on the canvas. To make the movements of the ball smooth, we extract the formants and calculate the median over a sliding time window. With a longer window we get smoother movements, but with the downside of a latency in the movement in relation to the spoken utterance. With a lower latency, and more immediate response, the movement of the ball becomes jerky. We have found that a time window of 50 ms, i.e. a refresh rate of 20 frames per second, makes the movements of the ball smooth, without a disturbing latency.

The direct, immediate feedback the moving ball gives the language learner is a great facilitator for discovering relationships between configurations of the mouth and tongue, and positions on the vowel chart. By moving the tongue forward and backward in the mouth, the ball moves from right to left on the canvas, and by opening and closing the mouth, the ball moves up and down on the canvas.

Anyone playing around with the software for a few minutes will be able to establish a relationship between articulatory movements and positions on the canvas.

2.4. Target spheres

In addition to the vowel-chart canvas and the moving ball, stationary target spheres can be placed at specific pre-determined positions on the canvas. These positions correspond to the locations where the vowels of the target L2 language are found (Swedish in our case).

These positions are determined by having native speakers say the desired vowels and storing the coordinates. The target spheres are a little larger than the moving ball and are as opposed to the moving ball not solid, but made of a wire-frame mesh, thus making it possible to see the moving ball when it enters the target sphere. A slider is available, allowing the students to change the size of the target spheres, as a way to adjust the difficulty level of the task of getting the moving ball inside the target sphere.

2.5. Practice mode and game mode

Two modes are available for the student to choose between. In practice mode the student is free to choose a vowel to practice on, and no time restrictions are given. By clicking on a button with a vowel, the corresponding target sphere will appear on the canvas. When there is no sound input, the moving ball will return to its starting point, which is in the center of the canvas (see figure 3)

Game mode is a ‘catch-the-target-spheres’ race against time. Target spheres are placed on the vowel chart, one at the time, and stays until the student has managed to keep the moving ball steadily inside the target sphere for 500 ms. The target sphere then turns green, and is replaced by a new one at another position, corresponding to another vowel. Two versions of the game have been tried: See how many targets one can get in one minute, alternatively, -how long time does it take to get all the targets. For the experiment reported in section 3, the latter was chosen, to facilitate comparison across subjects and vowels.

2.6. /Y/ and the third formant

The main difference between the Swedish /I/ and /Y/ sound lies in a shift in the third formant (F3), we experimented with different ways of visualizing this in an intuitive way that students would be able to understand. Since the vowel chart canvas, the moving ball, and the target spheres are all modeled in 3D, our first attempt was to use the z-axis to represent F3. Since the standard way of representing the vowel chart is in a plane, where F1 and F2 occupy the x-axis and y-axis respectively. If any movement in the z-axis should be visualized, the vowel chart, now a 3-D cage, must be viewed from an angle. After some initial attempts by students, this idea was abandoned, because it weakened some of the beneficial, intuitive aspects of moving the ball in the traditional x-y plane. Attempts were also made to change the color and size of the moving ball as a representation of shifts in the z-plane. In the end we settled for a solution where a binary red/green icon was made visible, close to the location of /Y/ in the chart.

3. Experiment

10 subjects were enrolled for a user study, to investigate the usefulness of the software as a vowel-learning tool. Five subjects were international language students, and five were native Swedish speakers used as a reference. Among the international students, two were Spanish, two were Italian, and one was from Syria. Both groups had three males and two females.

From the 18 vowels of Swedish, 10 were selected as part of the experiment. The nine long variants (see section 1.1) and the open fronted short /A/, which was selected because it in vowel quality has a close resemblance to the /A/ sound used in many languages. Since the task in the experiment was to keep the moving ball steadily inside each target sphere for at least 500 ms, it was decided that the long vowels were the most appropriate to try.

The experiments were conducted on a laptop computer with a microphone headset in a quiet private room. Each student performed the experiment on two separate occasions with a few days in between. Each session consisted of a calibration phase, and an initial training period of five minutes, getting acquainted with the program, before the tests started. On each occasion every student did three consecutive tests, and the times for reaching each target sphere were logged.

4. Results

To analyze the results, the data was split into four groups: Swedish subjects session one and two, and international subjects session one and two. The distinction between the data from the Swedish subjects and the international subjects is motivated to isolate the effect of getting acquainted with the use of the program under the assumption that all the Swedes already master the Swedish vowels. Comparing first and second session for the Swedish subjects will show the effect of that. Comparing the differences between first and second session for the international subjects and the Swedish subjects is thought to show some learning effects beyond learning to use the program. Inside each of these groups a different mean value was calculated for the different Swedish vowels in all the tests.

Learners of Swedish usually exhibit varying degrees of difficulties mastering different vowels. A reasonable assumption would be that they are difficult because they are unfamiliar, and therefore harder to reach. Our hypothesis is that the immediate feedback provided by the program would enable students to explore the unfamiliar regions, and that they initially would take a longer time to reach, but that after some training with the program, these areas would not pose a bigger problem than other areas.

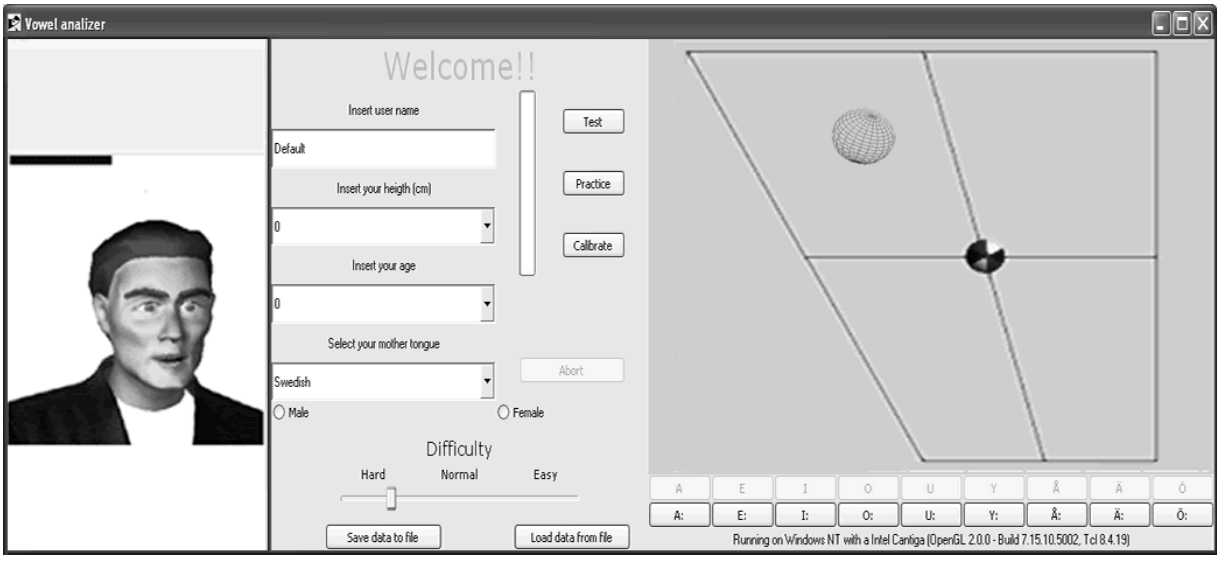

Figure 3 Screenshot of the software, with the moving ball in its resting position, and with one target sphere visible.

In figure 4 we see that the 'exotic' /Ä:/, /Ö:/, /Y:/ and /A:/ along with /E:/, which is more fronted than in many languages, are the vowels the international subjects spent most time on in the first session.

The top plot of Figure 5 shows that the biggest *gain* the international subjects made in time between session one and session two for the different vowels are the same. The gain for the Swedish subjects in the bottom plot of Figure 5, show a very different distribution. Although a t-test and ANOVA was calculated to see whether the differences between session one and session two were significant (which they were), we feel that the sample size is too small to draw any general conclusions in that direction yet.

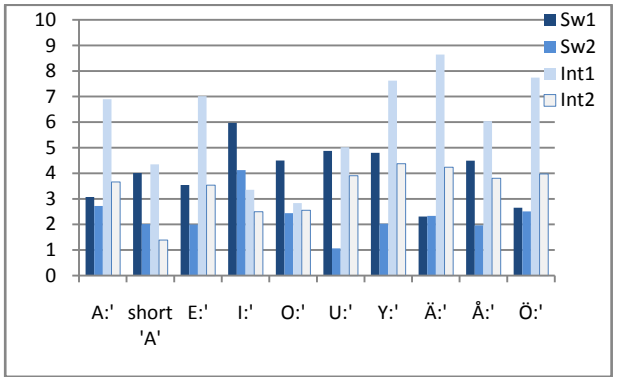

Figure 4: Mean times in seconds for the different vowels divided into four groups: Swedish subjects and international subjects session one and two.

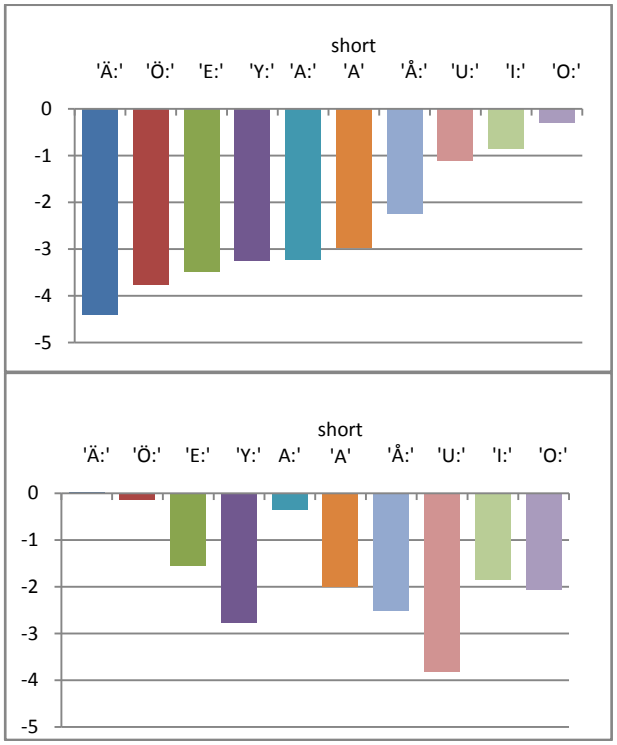

Figure 5: Difference in time between session two and one. Top plot: International subjects (sorted), Bottom plot: Swedish subjects

5. Discussion & future work

This paper has presented a new system to practice vowels in Swedish with real-time interactive feedback. Making CAPT systems to practice vowel production has been done before (see for example [5,6,7]) . The main contribution of this paper is thus a calibration technique based on an ECA that elicits cardinal vowels from the user, and uses those to normalize the vowel-space canvas, thus allowing all users, regardless of vocal tract size to use the system. We also extract the third formant, F3 in order to distinguish between certain vowels in Swedish. The system is not limited to Swedish, as it is fast and easy to make another set of targets, based on the vowel inventory of another language, as long as it is based on formant extraction.

The system was made as a standalone application. The intention is however to make it an integrated part of Ville - the virtual language teacher, [8] a language learning system developed at the Centre for speech technology, KTH. Work in this respect is underway, which will make it possible to try the system on a larger audience, and make some longitudinal studies of its effects and usefulness.

6. Acknowledgements

This work was partly financed by the Swedish Graduate School of Language Technology (GSLT). Many thanks also to Samer Al Moubayed for discussions, and help with Matlab.

7. References

[1] Indopedia.org [Online][Cited: 0515, 2009.] http://www.indopedia.org/Vowels.html

[2] Elert, C. C. (1966). *Allmän och svensk fonetik.* Almqvist \& Wiksell.

[3] Fant, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *STL-QPSR, 7*(4), 022-030.

[4] Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proc of Fonetik 2003, Umeå University, Dept of Philosophy and Linguistics PHONUM 9* (pp. 93-96).

[5] Auberg, S., Correa, N., Rothenberg, M., & Shanahan, M. (1998). Vowel and intonation training in an English pronunciation tutor. In *STiLL-Speech Technology in Language Learning*.

[6] Paganus, A., Mikkonen, V. P., Mantyla, T., Nuuttila, S., Isoaho, J., Aaltonen, O., & Salakoski, T. (2006). The Vowel Game: Continuous Real-Time Visualization for Pronunciation Learning with Vowel Charts. *Lecture Notes in Computer Science, 4139*, 696.

[7] Zahorian, S. A., & Correal, N. S. (1994). Vowel training experiments with a computer-based vowel training system. *The Journal of the Acoustical Society of America, 95*, 3014.

[8] Wik, P., & Hjalmarsson, A. (in press). Embodied conversational agents in computer assisted language learning. *Speech communication*.