

# The Vowel Game: Continuous Real-Time Visualization for Pronunciation Learning with Vowel Charts

Annu Paganus<sup>1</sup>, Vesa-Petteri Mikkonen<sup>2</sup>, Tomi Mäntylä<sup>1</sup>, Sami Nuuttila<sup>1</sup>,  
Jouni Isoaho<sup>1</sup>, Olli Aaltonen<sup>2</sup>, and Tapio Salakoski<sup>1</sup>

<sup>1</sup> University of Turku, Department of Information Technology, FI-20014 Turku  
Finland  
{Annu.Paganus, Tomi.Heikkimikael.Mantyla, Sami.Nuuttila,  
Jouni.Isoaho, Tapio.Salakoski}@utu.fi  
<sup>2</sup> University of Turku, Department of Phonetics, FI-20014 Turku  
Finland  
{Petteri.Mikkonen, Olli.Aaltonen}@utu.fi

**Abstract.** Learning to pronounce new speech sounds is difficult. Visual feedback helps in identifying the errors and indicating the achieved progress. The Vowel Game uses a visualization method that symbolizes the vocal tract. This instructs the user on how to adjust e.g. the tongue position during pronunciation. It gives information about the correctness and goodness of the uttered vowel. Preliminary evaluation suggests that continuous real-time feedback can be obtained, but the effect on learning remains to be tested.

## 1 Introduction

Visual feedback has been proven to be beneficial in language learning applications. When using Sona-Speech 3600-ESL [2] from KAY Elemetrics the user pronounces a vowel according to an example sound. Then the application draws a dot in real-time into a vowel chart where correct places of the vowels are presented with IPA-symbols. The application also opens a new window where it presents the “closest vowel” to the users production. There is also authentic video material of native speakers pronouncing the sample vowels. Baldi [12] is an animated 3D talking head that have been used for example for training the perception and production of speech for people with hearing loss. The head provides examples and feedback (smiling etc.) to the learner if one is making the right interpretation of words which the head says. In their investigation Massaro and Light [12] have used also the voice recognition system in the CSLU toolkit to evaluate the validity of the learners ability to produce certain words. In the Video Voice system [6] the learner gets information about how near his/her pronunciation is to the right phoneme. Those phonemes are shown at the coordinate system in which the axes are F1 and F2 values. Dowd et al [5] have visualized vowels with separate oval areas for each vowel. They don't use formants but resonances of the vocal tract using an acoustic impedance spectrometer. They claim this technique is more precise than using formants. They have got encouraging

results in learning: results with visual feedback and training were 25 percent units better than with only auditory feedback. The Optical Logo Therapy (OLT) [7] provides same kind of real-time feedback. It shows e.g. phonemes /s/, /z/, /sh/, /i/ and /u/ in the same picture. There is undefined space between the presented phonemes which they consider to be a problem. The tool tells when the learner has said the right phoneme but doesn't give information about how to improve.

In this paper we introduce the Vowel Game, a pronunciation learning tool based on formants. The phonetic background underlying the game is based on the Turku Vowel Test that is a research project build up to study the perception of vowels by speakers of different languages [15]. It is a perception test where the listener is asked to judge first what category the stimulus belongs to and second how good of an example of the given category the stimulus is. Application then produces a vowel chart according to listeners choices. While the study had been going on for several years, an idea emerged of how the produced vowel charts could be used in pronunciation training. If we can produce a vowel plotter that shows the exact relevant acoustic values of the utterance, we could use it as an instructional tool on a vowel chart of any given language.

The phonetic background underlying the game is discussed first in section 2. Third section describes the workings of the game. The learning with the system and real-time issues are discussed in section 4. Finally the paper is concluded with a discussion on future work.

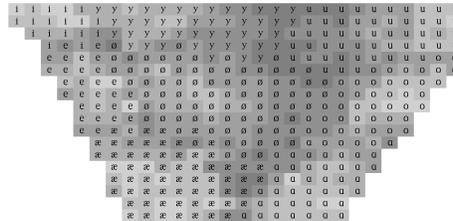
## 2 Phonetic Background

A vowel chart is a simple tool for visualizing speech. Place in the chart is determined by first and second formant frequency. Formants are energy peaks at a certain frequency resulted from vocal tract resonances. A vowel chart is a diagram where the first formant is presented as values of hertz or mels growing from top to bottom. The second formant is presented similarly from right to left. Vowel chart can be viewed as a simplification of a persons individual vocal tract. Width of the vowel chart corresponds to the length of the vocal tract, and height of the vowel chart corresponds to vocal tract height. Figure 1 shows two correspondences.

The perception of speech sounds is categorical. Lieberman et al. [11] found out that people tend to have little difficulties in discriminating sounds near phoneme boundaries, even though the acoustic qualities of phonemes are continuous. In cooing a prelingual infant produces all vowels of the vowel space. She can also discriminate sounds nonexistent in her native language. The infant listens and mimics adult speech. That makes her brains to start constructing permanent memory traces about the nature of her native speech sounds. At the age of six months, all humans have usually learned the phonetic categories and prototypes of their native language, making it extremely difficult to distinguish foreign speech sounds. The prototype is the best example of a phoneme category. This best example acts like a magnet drawing the other phonemes of the category towards it perceptually. This results in better discrimination and



A prototype chart demonstrates the structures inside categories. The chart is based on perceived goodness at a scale of 1 to 7. Goodness is demonstrated by grey scale colors. Lighter grey areas represent the more prototypical vowels. A prototype chart of Finnish is presented in Figure 3.



**Fig. 3.** Prototype chart of the eight Finnish vowels. Lighter areas represent the more prototypical vowels, whereas darker areas are rarely used in Finnish. [15]

### 3. The Vowel Game

The Vowel Game is an application that uses vowel charts in order to let the user train to pronounce vowels. The software is built with Java™ using version 1.5.0.

#### 3.1 The Idea of the Vowel Game

When the application is started, the user sees the Finnish vowel chart and the target phonemes circled at the chart as illustrated in Figure 4. The idea of the game is to learn to pronounce all the target vowels. The *Play* button starts the game. The user is expected to pronounce a vowel, which is continuously traced on the chart. When the users utterance is located on the chart, it is shown by switching the vowel yellow. Last five locations are shown at a time. When a target vowel has been hit it turns red. The *Pause* button pauses the speech data recording

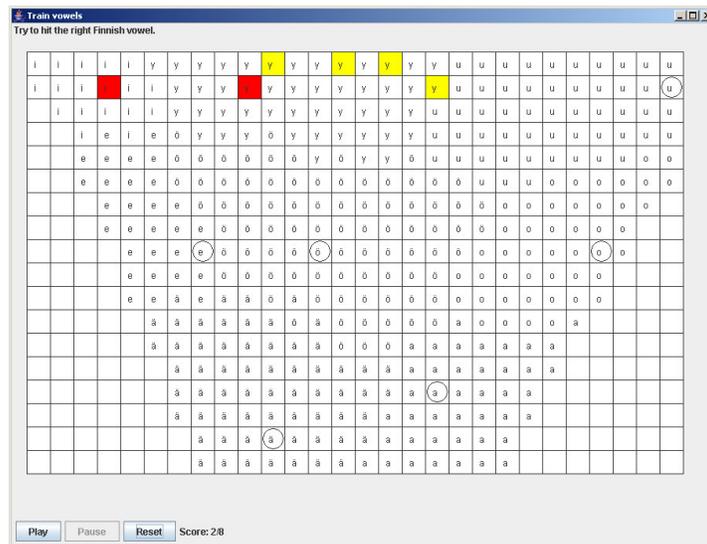
Figure 4 illustrates how the game looks like when the user is uttering the vowel /y/ and has already hit the vowels /i/ and /y/. Notice that the IPA-symbols are not used in the game, instead the written characters of standard Finnish of the corresponding phonemes are used.

We are currently working to offer auditory feedback with formant speech synthesis. When it is ready the user can get a sample from each vowel by clicking on the chart.

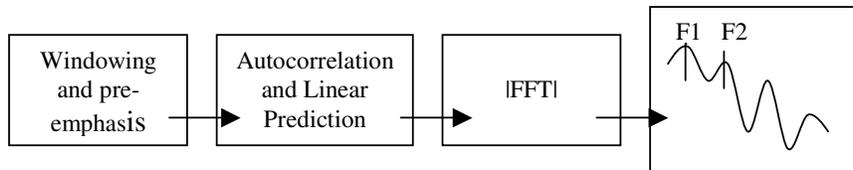
#### 3.2 Implementation Aspects

Figure 5 shows the steps that are taken while the formant locations are determined. The sampling frequency we use is 8 kHz. This enables us to review formants at

frequencies below 4 kHz according to Nyquist theorem [4]. The voice signal is windowed using a Hann window (aka Hanning) of length 256. One window takes then 32 ms which should include at least one glottal pulse. The signal is then pre-emphasized by a whitening filter that increases the spectral slope by 6 dB per each octave. The pre-emphasis stage thus increases the relative energy of the high-frequency spectrum [4] so that the higher frequencies with naturally lower relative energy get the same weight as the lower ones.



**Fig. 4.** A snap shot of the Vowel Game while user is saying vowel /y/ and has been hit the target vowels /i/ and /y/.



**Fig. 5.** Steps taken while the formant locations are determined.

Next, the autocorrelation coefficients are calculated for the 10th order Linear Prediction (LP) analysis that is used. The actual LP coefficients are then found by using a decomposition method to solve the normal equations [4]. Finally, the impulse response of the resulting analysis filter is Fourier transformed to find the spectral envelope of the speech signal. In this the Fast Fourier Transform (FFT) algorithm is used. Formants F1 and F2 are then found by locating the first two maxima in the spectral envelope.

Generally it is difficult to determine the formants precisely, rapidly and automatically [5]. In the current state of development of the Vowel Game the values are extracted rapidly and automatically but not always precisely. Problems occur when F1 and F2 are merged to one peak. That happens e.g. sometimes when the user utters the Finnish vowel /u/.

## 4 Discussion

Visual feedback has positive influences in learning foreign languages [5], and overcoming problems with speech production [6, 7]. The challenge in providing visual feedback is to make it easy to understand [6]. The Vowel Game shows how close the pronunciation is to the prototype, as the Sona-Speech [2] does, and also, if it falls within the correct category. This is a guide to shift the pronunciation for example towards a familiar phoneme in the same direction as the target prototype. Unlike in OLT [7] there is no undefined area between the prototypes, so the presentation is continuous.

We believe it would serve a purpose to have the feedback in real-time instead of after each attempt. The speaker shouldn't have to wait to see, how close the attempt came, before trying again. Another important factor is continuity of the training session. Real-time and continuous feedback can be used to search the correct pronunciation, or play around and see how the outcome is affected.

Systems delivering non-judgmental, immediate feedback during the pronunciation, such as our application and many parts of the SPECO system at KTH [17], are also seen beneficial by Zhang [18]. There was also a weakness noted in the Baldi system, where it occasionally gave false negative feedback [12], which is an argument on behalf of using directing, non-judgmental feedback.

The probable future of the system is to be a part of a toolbox of several applications. This game is for a teacher to apply, when the student's mastery of the language is at a point, where focusing on the correct pronunciation is useful.

The Vowel Game helps visual learners, as they can see what the vowel "looks like". For kinesthetic learners, real-time feedback would seem to us as equally helpful, as the learner given the visual guidance can feel the vowel around the mouth, and work with their own vocal tract and see what is happening. For people with hearing loss, visual feedback has also been found successful [12]. People with no hearing at all might find our application helpful e.g. on a mobile device as a tool for pronunciation confirmation when they talk. A third group of people who should be interested in the game are language professionals. They could use the application to train the awareness of their own vocal tract and the nature of speech production.

Providing the system for a PDA-platform poses a real challenge. We use Java to achieve code mobility at the cost of efficiency. It remains to be seen how well we can comply with the real-time requirements in the PDA environments.

Cost efficiency is one practical aspect. As noted by Zhang [18], feedback of this kind has been impractically expensive in the past. Today, a typical PC with enough

processing power is affordable. Where a small elementary school can not afford a real language studio, a desktop computer might be a low cost equivalent.

A research by Alais and Carlile [1] supports, that the human perception system is capable of adapting to a time difference of at least 68 ms, which is consistent with other researches mentioning the requirement of maximum video delay of 16-42 ms [14] and even 150 ms [16]. The risk in failing to achieve real-time is that the effects of notable delay “include overcompensation, lack of trust in the feedback and confusion and disorientation” [3].

Because the vowel chart is a simplification of the vocal tract, and there are as many vocal tract sizes and shapes as there are speakers, our tool needs to be calibrated. This can be done for example by the use of the so called “point vowels”, /i/, /a/ and /u/. Because these vowels are the articulatory and acoustic extremes of the vowel space [9], we can ask the user to articulate them and then set the vowel chart size accordingly. There are also more sophisticated methods for calculating vocal tract length and shape. These methods use formant data and pitch period estimations [13].

Another problem for accurate analysis of formants comes from differences in fundamental frequency. Because there is acoustic energy present only at the multiples of the fundamental frequency, there is more “empty space” between the multiples when the fundamental frequency is higher. This kind of “empty space” can in some cases be at a crucial point in the spectra. There are also ways to normalize the effect of fundamental frequency, in which we will look into in the future.

Preliminary tests suggest that the game is currently more suitable for men than for women. Women’s F1 values are sometimes higher than the values in the used chart, which is natural because the used vowel chart is based on synthesized male voice samples. However, once the calibration is finalized the problem should be solved.

The vowel chart is a simplification also in that there are other ways to inflict formants than tongue position. One of these ways is lip rounding. In the two dimensional model it is impossible to analyze or visualize whether a certain change in the second formant frequency is a cause of lip rounding or movement of the tongue. An application of the third formant could prove to be beneficial in determining lip rounding, but would also cause the system to become seriously more complex. At the current time we feel that the information provided by the two formants gives us satisfactory outcome and the application of higher formants seems unnecessary.

## **5 Conclusion and Future Work**

In this paper we introduced the Vowel Game - a tool to help people to learn to pronounce vowels. The study shows that real-time continuous visual feedback about correctness and goodness of the pronunciation is viable through formant charts. The game will be a part of a larger system including consonants and prosody training. Future work will also include e.g. fundamental frequency normalization and research on the applicability of the third formant. Also the effect on learning has to be studied.

## References

1. Alais, D., Carlile, S.: Synchronizing to real events: Subjective audiovisual alignment scales with perceived auditory depth and speed of sound. *Proc Natl Acad Sci USA*. 2005 Feb 8; 102(6), (2005) 2244-7
2. Carey, M.: CALL visual feedback for pronunciation of vowels: KAY Sona Speech. *CALICO Journal* 21, (3) (2004)
3. Day, P.N. Holt, P.O.B. Russell, G.T.: Modelling the Effects of Delayed Visual Feedback in Real-Time Operator Control Loops: A Cognitive Perspective. *Proceedings of XVII European Annual Conference on Human Decision Making and Manual Control*, Loughborough October, Group D Publications Ltd (1999) 70-79
4. Deller, J.R., Proakis, J.G., Hansen, H.L.: *Discrete-Time Processing of Speech Signals*. Macmillan, New York (1993)
5. Dowd, A., Smith, J.R., Wolfe, J.: 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 (1998) 1-20
6. Fitzgerald M., Gruenwald A., Stoker R.: Software review – Video Voice Speech Training System, *Volta Review*, vol. 89 (1989) 171-173
7. Hatzis, A.: *Optical Logo-Therapy (OLT): Computer-Based Audio-Visual Feedback Using Interactive Visual Displays for Speech Training*. PhD thesis, Department of Computer Science, University of Sheffield (1999)
8. Iverson P, Kuhl P.K, Akahane-Yamada R, Diesch E, Tohkura Y, Kettermann A, Siebert C.: A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87 (2003) B47-B57
9. Jakobson, R., Fant, G., Halle, M.: *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, Massachusetts. MIT Press. (1969)
10. Kuhl P.K.: Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* 50(2) (1991) 93-107
11. Liberman. A.M., Harris, K.S. Hoffman, H.S., Griffith, B.C.: The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54, (1957) 358-368
12. Massaro, D.W., Light, J.: Using Visible Speech to Train Perception and Production of Speech for Individuals With Hearing Loss, *Journal of Speech, Language and Hearing Research*, Vol. 47, April (2004) 304-320
13. Paige A., Zue V.: Calculation of Vocal Tract Length, *IEEE Transactions on Audio and Electroacoustics* 18, no. 3 (1970) 268-70
14. Regan, M. Pose, R.: Priority rendering with a virtual reality address recalculation pipeline. In *Proceedings of the 21st Annual Conference on Computer Graphics and interactive Techniques SIGGRAPH '94*. ACM Press, New York, NY (1994)
15. The Turku Vowel Test. Dept. of Phonetics, University of Turku. <http://fon.utu.fi/> 3.4.2006
16. Vaghi, I., Greenhalgh, C., Benford, S.: Coping with inconsistency due to network delays in collaborative virtual environments. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (London, United Kingdom, December 20 - 22, 1999)*. VRST '99. ACM Press, New York, NY (1999) 42-49
17. Vicsi, K. Roach, P. Öster, A-M. Kacic, Z. Csatári, F. Sfakianaki, A., Veronik, R.: A multilingual, Multimodal, Speech Training System SPECO, *Proceedings of Eurospeech 2001* (2001) 2807-2810
18. Zhang, F.: Using interactive feedback tool to enhance pronunciation in language learning. S. Mishra & R.C. Sharma (Eds.) *Interactive Multimedia in Education and Training*, Idea Group Publishing (2004) 377-399