An Automatic Vowel Space Generator for Language Learners' Pronunciation Acquisition and Correction

Xinyuan Chao[1], **Charbel El-Khaissi**[2], **Nicholas Kuo**[1], **Priscilla Kan John**[1], and **Hanna Suominen**[1, 3, 4]

[1]Research School of Computer Science, The Australian National University, Australia
[2]College of Arts and Social Sciences, The Australian National University, Australia
[3]Data61, Commonwealth Scientific and Industrial Research Organisation, Australia
[4]Department of Future Technologies, University of Turku, Finland

{*u6456596, charbel.el-khaissi, nicholas.kuo, priscilla.kanjohn, hanna.suominen*}*@anu.edu.au*

Abstract

Speech visualisations are known to help language learners to acquire correct pronunciation and promote a better study experience. We present a two-step approach based on two established techniques to display tongue tip movements of an acoustic speech signal on a vowel space plot. First, we use *Energy Entropy Ratio* to extract vowels; and then, we apply the *Linear Predictive Coding* root method to estimate Formant 1 and Formant 2. We invited and collected acoustic data from one *Modern Standard Arabic* (MSA) lecturer and four MSA students. Our proof of concept was able to reflect differences between the tongue tip movements in a native MSA speaker to those of a MSA language learner at a vocabulary level. This paper addresses principle methods for generating features that reflect bio-physiological features of speech and thus, facilitates an approach that can be generally adapted to languages other than MSA.

1 Introduction

Second language (L2) learners have difficulties in pronouncing words as well as native speakers (Burgess and Spencer, 2000) which can create inconveniences in social interactions (Derwing and Munro, 2005). Difficulty in providing pronunciation instructions by language teachers add extra challenges on L2 pronunciation training and corrections (Breitkreutz et al., 2001).

One solution to assist pronunciation acquisition is through the adoption of educational software applications (Levis, 2007). A well-designed language educational software can provide straightforward guidance to correct L2 pronunciation through multiple information sources. One instance of auxiliary systems is *Pronunciation Learning Aid* (PLA), which supports language students towards native-like pronunciation in a target language (Fudholi and Suominen, 2018). PLA achieves this via evaluating students' produced speech to reflect their pronunciation status. Another instance of auxiliary systems is visual cues, which serves as a friendly and accessible form of feedback to language students (Yoshida, 2018).

Through combining language lecturers' teaching with auxiliary systems, our aim is to assist students in both a classroom setting and in their individual practices. We present a prototype system that displays visual feedback on tongue movements to assist language learners to acquire correct pronunciation in the process of L2 studying. We have adopted a human-centred approach for the development of the system using a design-oriented perspective through applying a methodology that draws from *Design Science Research* (DSR) (Hevner et al., 2004) and *Design Thinking* (DT) (Plattner et al., 2009). Unlike machine learning methods, which train deep neural networks to predict articulatory movements (Yu et al., 2018), our proposed system uses vowel space plots based on bio-physiological features to help visualise tongue movements.

In this present work, we introduce a versatile prototype of our vowel space plot generator to address these challenges for students primarily learning MSA. Our design aims to allow L2 beginner learners to quickly visualise their status of pronunciation compared to those by their language teachers. We provide a reference vowel space plot adjacent to the students' own plots to reflect clear differences to support self-corrections. The envisioned applicability ranges from in-class activities to provide immediate and personalised suggestion to remote learning where in both cases glossary files are pre-uploaded by teachers or textbook publishers.

2 Related Work

Traditional acoustic plots, such as waveforms, spectrograms, and other feature plots are applied to vi-

sualise speech signals and can provide sufficient information to phoneticians, expert scientists, and engineers (Fouz-González, 2015). However, these methods fall short in providing straightforward suggestions for improving language students' pronunciation or otherwise lack an intuitive and user-friendly graphic user interface (Neri et al., 2002). A study proposed by Dibra et al. (2014) adopted the combination of waveform and highlighting syllables to visualise pronunciation in ESL studying shows using acoustic plots to support pronunciation acquisition is an implementable method.

Different from acoustic plots, another thinking of pronunciation visualisation was considered based on people's bio-physiological features. A pioneer study with this idea was introduced by Tye-Murray et al. (1993), in which they discussed the effect of increasing the amount of visible articulatory information, such as non-visible articulatory gestures, on speech comprehension. With the improvement of equipment, Ultrasound imaging, *Magnetic Resonance Imaging* (MRI), and *ElectroMagnetic Articulography* (EMA) can be alternative approaches to visualise the movement of articulators, and several study cases on pronunciation visualisation were implemented by Stone (2005), Narayanan et al. (2004), and Katz and Mehta (2015). However, these approaches are still difficult to be implemented in daily language studying since relevant equipment are often not available for in-class activities and self-learning, and generated images and videos are hard to be understood by ordinary learners.

Enlightened by imaging the movement of articulators, the idea of talking head, which is using 3D mesh model to display of both the appearance articulators and internal articulators, was introduced. Some of the fundamental works of talking head were completed by Odisio et al. (2004), and Serrurier and Badin (2008). With the techniques of articulatory movement prediction, such as *Gaussian Mixture Model* (GMM) (Toda et al., 2008), *Hidden Markov model* (HMM) (Ling et al., 2010), and popular deep learning approach (Yu et al., 2019). Although talking head is developing swiftly, the research about performance of talking head for pronunciation training is still insufficient.

The place and manner of articulation are well established variables in the study of speech production and perception (e.g. Badin et al., 2010). Early research has already realised the potential of using vowel space plots to achieve pronunciation visualisation, such as the studies by Paganus et al. (2006) and Iribe et al. (2012). These studies indicate that for language learners, vowel space plots are easy-to-understand, straightforward, and provide the necessary information for understanding their own tongue placement and movement. Therefore, vowel space plots are considered a useful tool for language learners to practice and correct their pronunciation relative to other pronunciation correction tools, such as ultrasound visual feedback or more traditional pedadogical methods like explicit correction and repetition.

3 The Proposed Approach

To visualise the tongue movement based on students' pronunciation practice, our proposed system needs to receive students' pronunciation audio signal as its input. After the process of vowel detection, vowel extraction, and formant estimation, the system can automatically generate the corresponding vowel space plot as its output. In this section, we will introduce how engineering and linguistics insights inspired our proposed method, and the details of audio signal processing procedures.

3.1 Design Methodology

To find a reliable solution for language students on the challenges about pronunciation acquisition, we adopted a design-based approach and implemented a human-centred approach by using the Design Thinking framework (Plattner et al., 2009) to find the students' needs in terms of pronunciation practice and transform these into requirements. In the Empathy and Define phases of DT, we defined our research question as "Finding an implementable and friendly approach for language learners to help them practice their pronunciation". After this, we participated in an MSA tutorial and observed students' behaviours during the process of pronunciation acquisition. Finally, we generated an online questionnaire for students which asks their in-class pronunciation training experience and their study preferences. The details of this survey were introduced in the thesis by Chao (2019).

Based on the observation of MSA tutorial, we found that students feel comfortable to interact with other people (lecturer or classmates) during pronunciation process. One advantage for interaction is other people can provide feedback on students' pronunciation. Another finding from observation

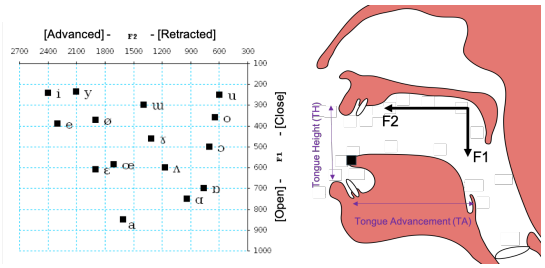
is the process of pronunciation acquisition can be seen as a process of imitation. Students need a gold-standard, such as teachers' pronunciation, as a reference to acquire new pronunciation and correct mispronunciation. The survey gives us some insights into students preferences about pronunciation study pattern. One of the most important insight is that students are interested in multi-source feedback of pronunciation training. For ordinary pronunciation, training students can only receive auditory information of pronunciation. Therefore, if a straightforward and easy-understanding visual feedback can be adopted in our proposed method, students will have a better experience and higher efficiency on pronunciation training.

The DT Empathy and Define phases gave us the insight that an ideal auxiliary pronunciation system should interact with learners, provide gold-standard pronunciation reference, and display reliable visual feedback to learners. The insight gained led to ideation discussions leading to the selection of vowel space plots as visualisation tool. We augmented the use of DT with the DSR approach, in the manner of John et al. (2020)'s study, to guide the development of our the artefact generated from our insights. Using the DSR method introduced by Peffers et al. (2007), we (1) identified our research question based on a research project which is about assisting new language learner on pronunciation acquisition with potential educational softwares, (2) defined our solution according to our observation and survey, (3) designed and developed our prototype of vowel space plot generator, (4) demonstrated our prototype to MSA lecturers and students, (5) and evaluated the prototype's performance. The DT and DSR process underpin all our methods.

3.2 Vowel Space Plot

Our proposed prototype uses vowel space plots as a tool to visualise the acoustic input. This visualisation then forms the basis for subsequent feedback on pronunciation features.

A vowel space plot is generated by plotting vowel formant values on a graph that approximates the human vocal tract (Figures 1(a) and 1(b)). F1 and F2 vowel formant values correlate with the position of the tongue during articulation (Lieberman and Blumstein, 1988). Specifically, F1 is associated with the height of the tongue body (tongue height) and plotted along the vertical axis, while its



(a) An example of vowel space (b) Vowel space plot and plot which shows the location oral cavity – the Formant- of different vowels in the vowel Articulation Correlation space

Figure 1: Vowel space plot and oral cavity

F2 counterpart is associated with tongue placement in the oral cavity (tongue advancement) and plotted along the horizontal axis.

The correlation between formant values and the tongue's height and placement is referred to as the formant-articulation relationship (Lee et al., 2015). These F1-F2 formant values can be rendered as x-y coordinates on a 2D plot to visualise the relative height and placement of the tongue in the oral cavity during articulation. When visualised alongside the tongue position of a native speaker's pronunciation, users can then see the position of their tongue relative to a standard reference or benchmark of their choice, such as an L2 teacher or native speaker. This visualisation supports pronunciation feedback and correction as users could then rectify the placement and/or height of their tongue during articulation to more closely align with its position in an equivalent native-like pronunciation.

3.3 Vowel Detection and Perception

To extract vowels from input speech signal, first, we calculate relevant energy criteria and find speech segments. Once speech segments were confirmed, we then use defined thresholds and detect vowels from these speech segments. This section will introduce the energy criteria and the thresholds we adopted in our practice.

Before detecting vowels in a speech signal, detrending and speech-background discrimination are two necessary steps of pre-processing. These steps ensure that only the correct speech information from the original signal is extracted, while other possible noise is ignored. In this way, the prototype minimises the possibility of including irrelevant signals during the feature extraction process.

Our prototype adopted the spectral subtraction

algorithm to achieve speech-background discrimination, as first introduced by Boll (1979). And the detrending can be achieved by the classic least squares method.

Our approach used *Energy Entropy Ratio* (EER), which is a calculated feature from input signal, as the criteria to find vowels from input speech signal. The EER can be calculated as following steps.

The *spectral entropy* (SE) of a signal describes its spectral power distribution (Shen et al., 1998). SE treats the signal's normalised power distribution within the frequency domain as a probability distribution and calculates its Shannon entropy. To demonstrate the probability distribution of a signal, let a sampled time-domain speech signal be $x(n)$, where the ith frame of $x(n)$ is $x_i(k)$ and the mth of the power spectrum $Y_i(m)$ is the *Discrete Fourier Transformation* (DFT) of $x_i(k)$. If N is the length of *Fast Fourier Transformation* (FFT), the probability distribution $P_i(m)$ of the signal can be then expressed as

$$p_i(m) = \frac{Y_i(m)}{\sum_{l=0}^{N/2} Y_i(l)}. \quad (1)$$

The definition of short-time spectral entropy for each frame of the signal can be further shown as

$$H_i = -\sum_{k=0}^{N/2} p_i(k) \log p_i(k). \quad (2)$$

The spectral entropy reflects the disorder or randomness of a signal. The distribution of normalised spectral probability for noise is even, which makes the spectral entropy value of noise great. Due to the presence of formants in the spectrum of signals in human speech, the distribution of normalised spectral probability is uneven, which makes the spectral entropy value small. This phenomenon can be used with speech-background discrimination to find out endpoints of speech segments.

In its practical application, SE is robust under the influence of noise. But spectral entropy cannot be applied for signals with a low *signal-to-noise ratio* (SNR) because when SNR decreases, the time-domain plot of spectral entropy will keep the original shape, but with a smaller amplitude. This makes SE insensitive to distinguishing speech segments from background noise. To provide a more reliable method of detecting the beginning and end of speech intervals, we introduce

$$EER_i = \sqrt{1 + |E_i / H_i|}, \quad (3)$$

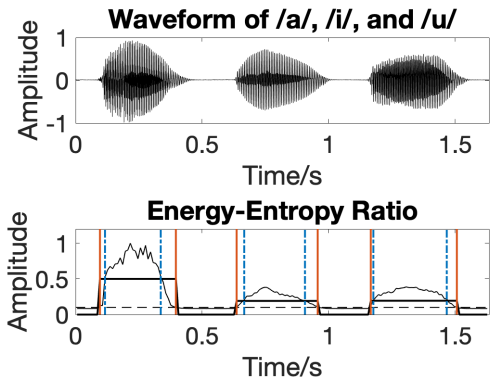


Figure 2: Vowel detection and segmentation

where E_i is the energy of the i^{th} frame of a speech signal, and H_i is the corresponding SE. Speech segments will have larger energy and smaller SE than silent segments. A division of these two short-term factors makes the difference between speech segments and silent segments more obvious.

The first threshold T_1 was implemented as the criterion to judge if the segment contains speech or not. The value of T_1 can be adjusted, and in our case we chose $T_1 = 0.1$ which performs well. Thus, segments with an energy entropy ratio larger than T_1 were classified as speech segments.

In each speech segment that is extracted, the maximum energy entropy ratio, E_{max}, and scale factor r_2, were used to set another threshold T_2 for detecting vowel segments:

$$T_2 = r_2 E_{max}. \quad (4)$$

Since different speech segments may have a different threshold T_2, segments with an energy entropy ratio larger than T_2 were used to detect vowels.

In an example visualisation of vowel detection and segmentation (Figure 2), three vowel phonemes — /a/, /i/, and /u/ — are contained in the speech signal. The black dashed horizontal lines show the threshold value $T_1 = 0.1$ for speech segment detection, while the solid orange lines show the detected speech segments within the speech signal. Similarly, the black vertical lines in bold indicate a dynamic threshold value T_2 for vowel detection across different speech segments, while the blue dashed lines display the vowel segments.

3.4 Formant Estimation

Formant value estimation is the next task after the detection of vowel segments from input speech signals. Our prototype adopted the *Linear Predictive*

Coding (LPC) root method to estimate the F1 and F2 formant values for vowels.

A common pre-processing step for linear predictive coding is pre-emphasis (highpass) filtering. We apply a straightforward first-order highpass filter to complete this task.

A simplified speech production model, which we adopted in our work is represented in Figure 3 following Rabiner and Schafer (2010). As shown in Figure 3, $s[n]$ is the output of the speech production system, $u[n]$ is the excitation from the throat, G is a gain parameter and $H(z)$ is a vocal tract system function. Let us consider the transfer function of $H(z)$ as an *Auto-Regression* (AR) model

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \qquad (5)$$

where $A(z)$ is the prediction error filter, which is used in the LPC root method below.

The polynomial coefficient decomposition of prediction error filter $A(z)$ can be used to estimate the centre of formants and their bandwidth. This method is known as the LPC root method, which was first introduced by Snell and Milinazzo (1993). Notably, the roots of $A(z)$ are mostly complex conjugate paired roots.

Let $z_i = r_i e^{j\theta_i}$ be any value of a complex root of $A(z)$, where its conjugate $z_i^* = r_i e^{-j\theta_i}$ is one of the roots of $A(z)$. Further, if F_i is the formant frequency corresponding to z_i, and B_i is the bandwidth at 3dB, then we have the relationships $2\pi T F_i = \theta_i$ and $e^{-B_i \pi T} = r_i$, where T is sampling period. Their solutions are $F_i = \theta_i/(2\pi T)$ and $B_i = -\ln r_i/\pi T$.

Since the order p of prediction error filter is set in advance, the pair number of complex conjugate roots will be up to $p/2$. This makes it straightforward to find which pole belongs to which formant, since extra poles with a bandwidth larger than a formant's bandwidth may be conveniently excluded.

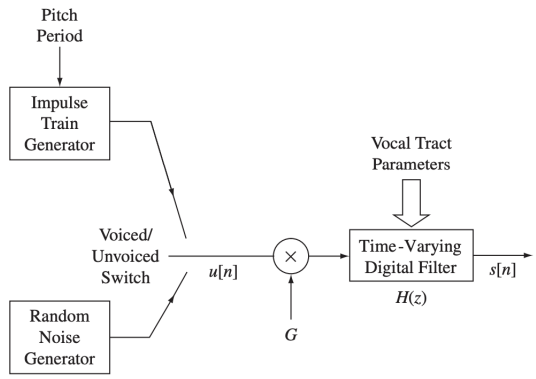


Figure 3: A simplified model of speech production

4 Preliminary Evaluation Experiment

We conducted two experiments to evaluate the performance of our prototype. First, we invited a native Arabic speaker who is a *Modern Standard Arabic* (MSA) lecturer at The Australian National University (ANU) to provide a glossary of MSA lexicon and their corresponding utterances. These utterances constituted the gold-standard or target pronunciation for users. Then, we invited four MSA language students to use our prototype by pronouncing four MSA words. For each lexical item pronounced, the articulation was visualised on a vowel space plot so users can compare their pronunciation alongside the native-like, target pronunciation of their lecturer. Following this visual comparison, users were prompted to pronounce the same word again.

In the experiments, we want to verify the feasibility and accessibility of our prototype. The feasibility of our prototype was determined by whether the interpretation of the comparison plots in the first instance supported improved pronunciation of the same word in subsequent iterations. And the accessibility refers to whether our prototype can provide implementable and correct feedback for learners to visualise their pronunciation.

Ethical Approval (2018/520) was obtained from the Human Research Ethics Committee of The Australian National University. Each study participant provided written informed consent.

4.1 Feasibility Test

The functionality of the prototype, including speech detection, vowel segmentation and plot generation, was first verified by using a series of acoustic signals as input to observe the accuracy of the output vowel space plot. The MSA lecturer's pronunciation of MSA lexicon was used here to test the veracity of the prototype output. The MSA dataset comprised of ten lexical items[1] and their corresponding pronunciation, henceforth referred to as the "standard reference" (see Table 1).

For each vocabulary item and corresponding audio input, we observed the vowel space plot gen-

[1]Refer to MSA Vocabulary Selection (Section 8) on our selection criteria of this list.

Vocabulary	MSA	Transliteration	Vowels
clock	ساعة	/sāʿa/	2
eggs	بيض	/bayḍ/	1
mosque	جامِع	/jāmiʿ/	2
phone	هاتِف	/hātif/	2
shark	قِرْش	/qirš/	1
soap	صابون	/ṣābūn/	2
spring	ربيع	/rabīʿ/	2
street	شارِع	/šāriʿ/	2
student(male)	طالِب	/ṭālib/	2
student(female)	طالبة	/ṭāliba/	3
watermelon	بَطّيخ	/bāṭṭīḵ/	2

Table 1: Ten reference vocabularies

Vocabulary	MSA	Transliteration	Vowels
shark	قِرْش	/qirš/	1
soap	صابون	/ṣābūn/	2
student(male)	طالِب	/ṭālib/	2
student(female)	طالبة	/ṭāliba/	3

Table 2: The student test data of four MSA words

erated by our prototype. The accuracy and accessibility of our prototype's speech and vowel detection functionality was determined by its ability to correctly visualise tongue positioning for each vowel in a word. This was determined based on a comparison with statistical averages of formant values for the same vowel. We use a Sony Xperia Z5 mobile phone to collect the utterance of glossary from the MSA lecturer. The utterances were recorded as individual mp3 files which can be used as input of our prototype. Each mp3 file contains one MSA vocabulary in the glossary. These mp3 files were recorded in the lecturer's office to reduce background noise.

4.2 Accessibility Test

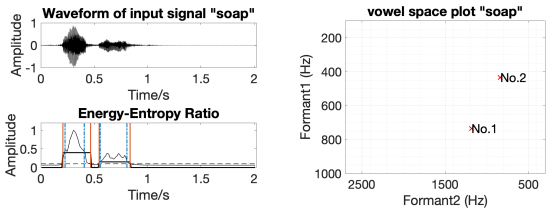
The verification of our prototype's functionality alone is insufficient to prove that the prototype can assist in providing valuable corrective feedback to users. Therefore, we invited two male students and two female students who were enrolled in a beginner MSA course (ARAB1003) at ANU to voluntarily participate in our accessibility test The success of our prototype's feedback function was determined by whether the language learners can interpret their pronunciation on a vowel space plot against the standard reference in order to produce a more native-like pronunciation for the same word.

Volunteers were aged between 19 and 22 and had completed an introductory MSA course (ARAB1002), which meant they had basic knowledge of MSA and were familiar with its alphabet and phonetic inventory. Four lexical items from the glossary in the standard reference were selected as test items which shown in Table 2 for the volunteers to pronounce. Volunteers pronounced each of the four vocabulary items independently, which were recorded respectively as audio files. These files were processed by our prototype and the corresponding vowel space plots were generated to visualise their pronunciation for each word. Then, their vowel space plots were compared to the corresponding vowel space plot of the standard reference. Participants were advised to use this comparison plot as the basis for their pronunciation feedback prior to repeating the pronunciation of the word. Then, participants pronounced the word a second time and the generated plot was once again compared to the standard reference. This time, the comparison assessed whether the participant's articulation of the vowel was more closely aligned to the standard reference compared to the first pronunciation. In other words, the second iteration of pronunciation allowed for an assessment of whether our prototype provided valuable visualisation information to participants, and whether it helped them immediately correct and improve their pronunciation relative to the standard reference.

We participated in one of the MSA course tutorials and were keen to see the quality of acoustic data, which were collected from a noisy circumstance, like a classroom. The collecting device was a MacBook Pro 2017. We wrote a Matlab recorder function with GUI to collect the utterance provided by volunteers who were from this tutorial. The utterance were collected as individual wav files and each file contained one word from volunteers.

5 Results and Discussion

We used collected speech signals to test the feasibility and accessibility of our prototype. To test the feasibility, we fed the standard references to our prototype and verify whether the output vowel space plot can reflect the correct tongue motion of the corresponding word. As for accessibility, we used the student test data and generated the vowel space plot, and then found corresponding words from a standard reference and compare these two vowel space plots. An ideal result is the student test



(a) Vowel segmentation of standard reference "soap"

(b) Vowel space plot of standard reference "soap" with /ā/ and /ū/ two vowels

Figure 4: The waveform, energy-entropy ratio, and vowel space plot for standard reference word "soap" (provided by a MSA teacher)

data can reflect the student's tongue motion, and the student can find how to improve the pronunciation by compare these two vowel space plots. With the vowel space plots of the same words from student test data and standard reference, we compared the corresponding plots to see if the corresponding plots and if the vowel space plots can provide useful feedback on pronunciation correction. In this paper, we display the MSA word "soap" (صابون , /ṣābūn/) as an example of our results.

5.1 Feasibility

To test the feasibility of our prototype, we picked one vocabulary item (the word "soap") from standard reference and verify whether the output vowel space plot can reflect the tongue motion. The waveform, energy-entropy ratio, and vowel space plot for standard reference word "soap" (Figure 4).

From Figure 4(a), we found two voice segments between solid orange lines that were recognised from the input speech signal, and the two voice segments, which contained one vowel between dash blue lines for each. In Figure 4(b), the two vowels of /ā/ and /ū/ were mapped in the vowel space. This vowel space plot was made available to the users so they can get familiar with their tongue position in the oral cavity and use this visual feedback towards pronouncing the word "soap" correctly (Figure 5).

5.2 Accessibility

To test the accessibility of our prototype, we compared the vowel space plot of standard reference and the vowel space plot of student test data. We continue to use the word "soap" here as an example. Figures below show the results of MSA vocabulary "soap" pronounced by the four anonymous students. Students will see two vowel space plot from the

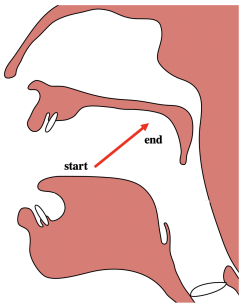


Figure 5: The tongue motion for the MSA word "soap"

prototype: one shows the standard reference, and another reflects their own pronunciation.

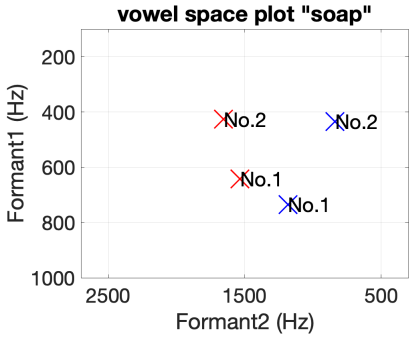
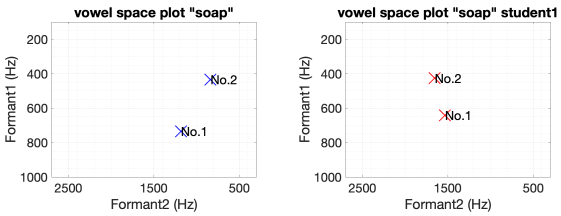


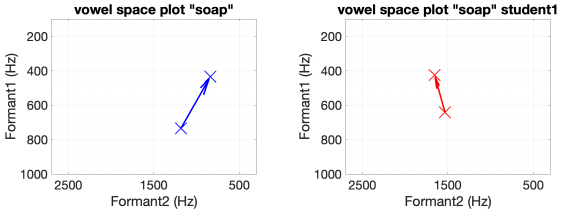
Figure 6: The tongue movement (reference and student1's practice) for the MSA word "soap"



(a) Standard reference "soap"

(b) Vowel space plot of user input-1 "soap" with /ā/, /ū/

Figure 7: The vowel space plot from standard reference and student1

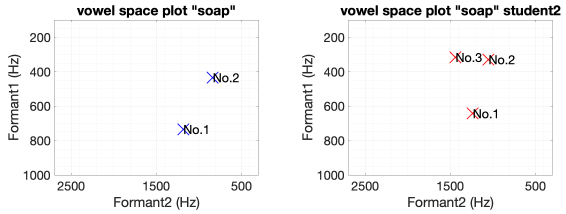


(a) Standard reference "soap" with arrow

(b) Vowel space plot of user input-1 "soap" with arrow

Figure 8: The vowel space plot from standard reference and student1 with arrow

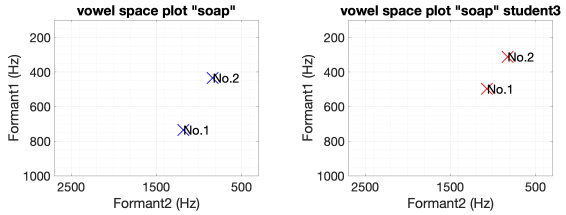
Figure 6 shows the overlay vowel space plot of standard reference (blue crosses) and student1's pronunciation practice (red crosses). Since the key information from vowel space plot is the trend of tongue movement, it is not necessary to compare the standard reference and students' pronunciation on the same vowel space plot. From Figure 7, student1's tongue should be drawn back instead of moving it to the front of the oral cavity. The vertical down-up movement of the tongue was correct. Figure 8 shows the tongue movement with an arrow. This is more readable and friendly for students to help them perceive their tongue movement.



(a) Standard reference "soap" (b) Vowel space plot of user input-2 "soap" with /ā/, /ū/ following wrong trajectory

Figure 9: The vowel space plot of standard reference and student2

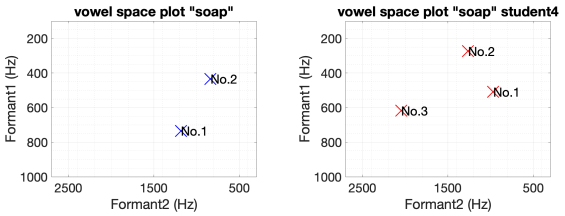
Student2, on the other hand, should focus on the pronunciation of the second vowel /ū/. According to Figure 9, we can see that the pronunciation of "soap" pronounced by student2 had the correct tongue motion trajectory when compared with the standard reference of Figure 1. This student's vertical down-up movement of the tongue was correct. A small defect for this practice was that there existed an unexpected vowel for the end of this pronunciation practice. For further practice, the advice for student1 targeted pronouncing a clean and neat end of the word "soap".



(a) Standard reference "soap" (b) Vowel space plot of user input-3 "soap" with /ā/, /ū/

Figure 10: The vowel space plots of standard reference and Student3

Student3, in turn, had the correct tongue motion, and the pronunciation was good as well. However, the starting point of the first vowel /ā/ was somewhat higher than its standard reference. Hence, our suggestion for Student3 was to lower the starting position of the word "soap".



(a) Standard reference "soap" (b) Vowel space plot of user input-4 "soap"

Figure 11: The waveform, energy-entropy ratio, and vowel space plot of Student4

Finally, student4 and student1 made similar mispronunciation: student4 should draw the tongue back instead of moving it forward while pronouncing the second vowel /ū/. Besides this mistake, another interesting point worthy of notice was that another unexpected vowel occurred by the end of this speech signal. According to waveform analysis, this vowel was not pronounced by student4 but originated from the background noise due to the data collection during an in-class activity. This meant that the sudden noise from background can still influence the analysis result although our prototype already applied its denoising algorithm to this speech signal. Hence, we made a suggestion to try to adopt a more effective denoising function as the future development of the system to satisfy the requirements from students to practice their pronunciation anywhere, including noisy settings.

6 Conclusion

This paper presented the initial proof of concept that used vowel space plots to enhance language learning in second languages. The idea of our prototype was based on our early stage DSR process and MSA language student survey (Chao, 2019). Our prototype was designed to generate clear visual feedback from speech input, and it was tested to assist the pronunciation of L2 MSA beginners.

Our main contribution is the vowel space plot generator prototype which produces easily understandable visual cues from analysing the biophysiological features of user speech. Our prototype is hence user-friendly for improving language learner pronunciation.

To gain evidence of our prototype being effective on assisting language learners' pronunciation training, we designed an experiment to test at the vocabulary level the feasibility and accessibility of the prototype and invited language students to provide their audio data for experimental use. Also, according to students' feedback, we proposed a series of future developments that are described in the next section. One limitation of our presented work is that there was no re-testing of pronunciation after the students received feedback from the system to check that their pronunciation improved. We plan to deploy re-tests as mentioned in our next stage experiments

7 Future Work

In the future, we aim to build on this current work to verify and quantify the pronunciation improvements gained from each user. This will help us to understand the effectiveness of this current design of the prototype and enable us to select appropriate extensions to enhance L2 learning experiences.

We are currently considering to build a correction subsystem for pronunciation practice. In addition to the existing vowel space plots, we theorise that it would be helpful to construct a system that could directly compare our users' speech to a set of externally stored standard references. This should enable the users to correct their pronunciation with higher precision and efficiency. Such a design could also potentially provide personalised pronunciation assistance via analysing user-specific pronunciation patterns.

Future iterations also intend to test a much more varied selection of MSA words that capture both short and long vowels in word initial, medial and final positions, as well as the two MSA dipthongs /aw/ (e.g. ضوء /ḍaw/ 'light') and /aj/ (e.g. بيت /bajt/ 'house') and MSA consonant.

Another potential future direction is to animate the tongue motion. Iribe et al. (2012) showed that such animations could achieve better results than their static counterparts. We expect the animated version of the vowel space plot to display tongue motions while people speak to help users to better conceptualise pronunciation in real-time.

8 Clarification: MSA Vocabulary Selection

The justification for the selection of the above ten words was based on a variety of factors. First, the selected vocabulary items were basic MSA words chosen in consultation with an MSA teacher to ensure students had been explicitly taught or otherwise been exposed to them during the course of their language learning.

Second, the selected words were restricted to one-to-three syllabic words only. This restriction ensured that sentence-level factors affecting the articulation of vowels were excluded (e.g. /t/-insertion rule in *Iḍāfah* structures; ساعة /sāʿa/ "clock" vs. ساعة يوسف /sāʿa**t** jusif/ "Joseph's clock"), thus allowing for a straightforward assessment of how the prototype detected speech boundaries and extracted the relevant features from vowel segments.

Finally, the ten words selected captured the three, cardinal MSA vowels: /a/ i/ and /u/. Although these vowels exist in the English phonemic inventory and do not theoretically pose a challenge for English-speaking L2 learners of MSA, when they are considered alongside surrounding MSA consonants then their articulation becomes more difficult, such as in the well-known case of emphatic spreading caused by the presence of pharyngeal or pharyngealised consonants ('emphatics') (e.g. Shosted et al., 2018).

Acknowledgement

The authors express their gratitude to participants and other contributors of this study. Furthermore, we would like to thank our three anonymous ALTA reviewers for their careful comments, which helped us to improve this present work.

We would also like to thank Ms Leila Kouatly, a MSA lecturer who works at the Australian National University (ANU) for helping us on the selection of the MSA glossary. She also provided us a series of opportunities to join her classes and tutorials. We acquired many valuable observations on her pedagogical methods and skills. Her activity in promoting our study ensured that students actively participated in our student experience survey and preliminary evaluation experiments.

Moreover, we thank Dr Emmaline Louise Lear and Mr Frederick Chow. Dr Lear helped us to acquire ethic approval for our study and provided us inspirations from an educator's perspective. Mr Chow helped us on communication with ANU Centre for Arab and Islamic Studies which is crucial for our study and commented on engineering details of our project. They also provided insightful suggestions for an early presentation for this study as examiners. We would like to express our sincere appreciation for their help and remarkable work.

Finally, we acknowledge the funding and support by Australian Government Research Training Program Scholarships and ANU for the first three authors' higher degree research studies.

References

Pierre Badin, Yuliya Tarabalka, Frédéric Elisei, and Gérard Bailly. 2010. Can you 'read' tongue movements? evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, 52:493–503.

Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120.

Judy Breitkreutz, Tracey M Derwing, and Marian J Rossiter. 2001. Pronunciation teaching practices in canada. *TESL Canada journal*, pages 51–61.

John Burgess and Sheila Spencer. 2000. Phonology and pronunciation in integrated language teaching and teacher education. *System*, 28(2):191–215.

Xinyuan Chao. 2019. *Supporting students' ability to speak a foreign language intelligibly using educational technologies:The case of learning Arabic in the Australian National University*. College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia.

Tracey M. Derwing and Murray J. Munro. 2005. Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3):379–397.

Dorina Dibra, Nuno Otero, and Oskar Pettersson. 2014. Real-time interactive visualization aiding pronunciation of english as a second language. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 436–440.

Jonás Fouz-González. 2015. Trends and directions in computer-assisted pronunciation training. *Investigating English Pronunciation Trends and Directions*, pages 314–342.

Dzikri Fudholi and Hanna Suominen. 2018. The importance of recommender and feedback features in a pronunciation learning aid. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 83–87, Melbourne, Australia. Association for Computational Linguistics.

Alan R. Hevner, Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. Design science in information systems research. *MIS Quarterly*, 28(1):75–105.

Yurie Iribe, Takurou Mori, Kouichi Katsurada, Goh Kawai, and Tsuneo Nitta. 2012. Real-time visualization of english pronunciation on an ipa chart based on articulatory feature extraction. *Interspeech 2012*, 2:1270–1273.

Priscilla Kan John, Emmaline Lear, Patrick L'Espoir Decosta, Shirley Gregor, Stephen Dann, and Ruonan Sun. 2020. Designing a visual tool for teaching and learning front-end innovation. *Technology Innovation Management Review*, 10.

William F. Katz and Sonya Mehta. 2015. Visual feedback of tongue movement for novel speech sound learning. *Frontiers in Human Neuroscience*, 9:612.

Shao-Hsuan Lee, Jen-Fang Yu, Yu-Hsiang Hsieh, and Guo-She Lee. 2015. Relationships between formant frequencies of sustained vowels and tongue contours measured by ultrasonography. *American journal of speech-language pathology / American Speech-Language-Hearing Association*, 24:739–749.

John Levis. 2007. Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, 27:184.

Philip Lieberman and Sheila E. Blumstein. 1988. *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science and Communication. Cambridge University Press.

Zhen-Hua Ling, Korin Richmond, and Junichi Yamagishi. 2010. An analysis of hmm-based prediction of articulatory movements. *Speech Communication*, 52(10):834–846.

Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd. 2004. An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America*, 115(4):1771–1776.

Ambra Neri, Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning*, 15(5):441–467.

Matthias Odisio, Gérard Bailly, and Frédéric Elisei. 2004. Tracking talking faces with shape and appearance models. *Speech Communication*, 44:63–82.

Annu Paganus, Vesa-Petteri Mikkonen, Tomi Mäntylä, Sami Nuuttila, Jouni Isoaho, Olli Aaltonen, and Tapio Salakoski. 2006. The vowel game: Continuous real-time visualization for pronunciation learning with vowel charts. In *Advances in Natural Language Processing*, pages 696–703, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. 2007. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77.

Hasso Plattner, Christoph Meinel, and Ulrich Weinberg. 2009. *Design-thinking*. Springer.

Lawrence Rabiner and Ronald Schafer. 2010. *Theory and Applications of Digital Speech Processing*, 1st edition. Prentice Hall Press, Upper Saddle River, NJ, USA.

Antoine Serrurier and Pierre Badin. 2008. A three-dimensional articulatory model of the velum and nasopharyngeal wall based on mri and ct data. *The Journal of the Acoustical Society of America*, 123:2335–55.

Jia-lin Shen, Jeih-weih Hung, and Lin-shan Lee. 1998. Robust entropy-based endpoint detection for speech recognition in noisy environments. In *Fifth international conference on spoken language processing*.

Ryan K Shosted, Maojing Fu, and Zainab Hermes. 2018. *Arabic pharyngeal and emphatic consonants*, chapter chapter3. Routledge.

R. C. Snell and F. Milinazzo. 1993. Formant location from lpc analysis data. *IEEE Transactions on Speech and Audio Processing*, 1(2):129–134.

Maureen Stone. 2005. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, 19(6-7):455–501. PMID: 16206478.

Tomoki Toda, Alan W Black, and Keiichi Tokuda. 2008. Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model. *Speech Communication*, 50(3):215–227.

Nancy Tye-Murray, Karen Iler Kirk, and Lorianne Schum. 1993. Making typically obscured articulatory activity available to speech readers by means of videofluoroscopy. In *NCVS Status and Progress Report*, volume 4, pages 41–63.

Marla Tritch Yoshida. 2018. Choosing technology tools to meet pronunciation teaching and learning goals. *The CATESOL Journal*, 30(1):195–212.

Lingyun Yu, Jun Yu, and Qiang Ling. 2018. Synthesizing 3d acoustic-articulatory mapping trajectories: Predicting articulatory movements by long-term recurrent convolutional neural network. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4.

Lingyun Yu, Jun Yu, and Qiang Ling. 2019. Bltrcnn-based 3-d articulatory movement prediction: Learning articulatory synchronicity from both text and audio inputs. *IEEE Transactions on Multimedia*, 21(7):1621–1632.