SLIONS: A Karaoke Application to Enhance Foreign Language Learning

Dania Murad[1] Riwu Wang[1] Douglas Turnbull[2*] Ye Wang[1]
[1]School of Computing, National University of Singapore, Singapore
[2]Department of Computer Science, Ithaca College, USA
{daniamurad,wangye}@comp.nus.edu.sg,riwu@u.nus.edu,dturnbull@ithaca.edu

ABSTRACT

Singing songs can be an engaging and effective activity when learning a foreign language. In this paper, we describe a multi-language karaoke application called *SLIONS: Singing and Listening to Improve Our Natural Speaking*. When developing this application, we followed a user-centered design process which was informed by conducting interviews with domain experts, extensive usability testing, and reviewing existing gamified karaoke and language learning applications. The key feature of SLIONS is that we used automatic speech recognition (ASR) to provide students with personalized, granular feedback based on their singing pronunciation. We also provided multi-modal instruction: audio of music and singing tracks, video of a professional singer and translated text of lyrics to help students learn and master each song in the foreign language. To test the efficacy of SLIONS, we conducted a one-week pilot study with English and Chinese language learning students (N=15). The initial quantitative results show that our application can improve pronunciation and may improve vocabulary. In addition, the qualitative feedback from the students suggests that SLIONS is both fun to use and motivates students to practice speaking and singing in a foreign language.

CCS CONCEPTS

• **Human Computer Interaction (HCI)**; • **Ubiquitous and Mobile Computing**;

KEYWORDS

Foreign Language Learning; Music Technology; Educational Technology; Intelligent Tutoring Systems; Mobile Application

ACM Reference Format:
Dania Murad[1] Riwu Wang[1] Douglas Turnbull[2] Ye Wang[1]. 2018. SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240691

*This work was conducted when the author was a senior visiting researcher at National University of Singapore.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240691

1 INTRODUCTION

Many people enjoy both playing games and singing songs. This may explain why karaoke applications like Smule's Sing![1] and Tencent's Quanmin K Ge,[2] have become popular with millions of users in recent years [23, 25]. Using these applications, an individual sings a song and receives a score based on some notion of singing quality [13]. The individual can then listen to, learn from, and re-record the song in an attempt to improve his or her performance.

Similarly, there is a large number of individuals who are interested in learning a foreign language. For example, in 2000 the British Council estimated that 1 billion individuals were learning English alone [2]. Moreover, individuals invest a significant amount of time and money by taking classes or using language learning software (e.g., Rosetta Stone[3], Pimsleur[4]) to learn a foreign language. In recent years, language learning applications like Duolingo[5] and Babbel[6] have used gamification to make language learning more enjoyable and effective [10, 30].

If we combine these two ideas together, we can create a new kind of application: a gamified karaoke application for foreign language learning. Research and pedagogical studies have shown that singing in a foreign language has clear benefits in language acquisition [6, 9, 11], particularly aiding in vocabulary [19], pronunciation [20, 27], memory recall [31] and cultural appreciation [8]. One of the ways of incorporating technology with language pronunciation is employing automatic speech recognition (ASR) for personalized feedback [3]. That is, it is often not feasible for traditional classroom teachers to listen and provide feedback to each student in a timely manner. This provides an opportunity to create a novel technological tool that automatically provides constructive language-based feedback to the students, thereby directly contributing to their language learning. However, despite the advancements in technological tools that aid in foreign language learning, we are unaware of any previous work that specifically uses ASR-based singing evaluation for language learning.

In this paper, we describe a multi-language karaoke application *SLIONS* (Singing and Listening to Improve Our Natural Speaking) that aims to provide a foreign language learning platform. We followed a user-centered design process when designing the karaoke application, and our initial prototype is aimed at young adult language learners. However, to target a broader range of individuals, we have provided a catalog which includes nursery rhymes, pop

[1]https://www.smule.com/listen/sing-karaoke/8
[2]http://kg.qq.com/
[3]https://www.rosettastone.com/
[4]http://www.pimsleur.com/
[5]https://www.duolingo.com/
[6]https://www.babbel.com/

songs, and classical music. For the current research, we are targeting Chinese and English language learners to study the effectiveness of SLIONS for language learning. Users can also earn points by completing singing exercises, thus adding elements of gamification. Our application provides users a platform where they can work at their own pace and in their own time, complete interactive exercises anywhere and receive immediate personalized feedback on their singing pronunciation. It also gives them the opportunity to practice and master different parts of the song so that they can improve their pronunciation.

We do not intend SLIONS to be a standalone language learning platform such as Rosetta Stone or Duolingo which have well-structured lesson plans and comprehensive syllabi. Rather, we have designed it as an entertaining tool that can be used to augment other classroom and software courses. That is, we include foreign language songs that are popular or have a familiar melody so that students find them easy to learn and enjoyable to sing. Through practice, we hypothesize that students who enjoy singing will not only improve their language skills but also be more motivated to learn the language in the classroom or other learning settings.

1.1 Contributions

The main contributions of the current research are:

(1) Provide design considerations while exploring recent research on singing, gamification, and foreign language learning.
(2) Introduce SLIONS as an application that integrates automatic speech recognition technology with language learning pedagogy.
(3) Conduct a pilot user study to test the efficacy of SLIONS for enhancing language learning through singing.

2 LITERATURE REVIEW

In this section, first we describe how music and speech are related to one another and contribute towards foreign language acquisition. Then we provide details on the intelligent tutoring software (ITS) for pronunciation training and how speech recognition tools and music technologies have been used for this purpose. Finally, we describe four commercial websites and applications that use music for language learning.

2.1 Language and Music

While language is an effective medium for verbal communication, music is a mean to perceive sound patterns. Mora [21] presented a close relationship between language and music, suggesting that music not only helps in improving pronunciation skills but also contributes towards the language acquisition process. McMullen *et al.* [18] provided some interesting insights regarding the similarities and parallels that have been found between language and music including a similar processing mechanism, especially in childhood. Their work leads to the claim that the capability of segregating music and language is initially not present in childhood, but is developed over time, thus adults have a separate language processing mechanism.

Schön *et al.* [28] hypothesized that pairing the musical sequence with the language sequence greatly aids in learning new words by leveraging the structural properties of the song to segment new words in a foreign language. Experimental results have shown that the word learning performance rate was highest for the group of users who were exposed to continuous singing with a constant mapping between syllable-pitch, thereby proving the claim that linguistic and music mapping enhance the performance and learning outcomes. Singing also has benefits over speaking and rhythmic speaking for language learning and is most evident on the verbal recall even after a certain delay [17]. Anvari *et al.* [1] described interrelation between music, phonological awareness and reading development by investigating a population of a hundred 4-5 years old children. They deduced that same auditory processing exists while perceiving music and developing reading skills.

2.2 Automatic Speech Recognition in Language Learning

Intelligent tutoring system (ITS) is a system that provides personalized intelligent feedback to the users with minimal involvement of teachers [24]. The use of ASR as an application of ITS has been effective for language learning, specifically pronunciation [29]. It is a pronunciation software that recognizes the words in a person's speech and provides recognition results based on speech features. ASR can be advantageous in a language learning classroom at different levels such as:

- Students can speak in their target language and receive feedback by comparing the pronounced words with the target words.
- Students can practice and learn the speaking tasks individually at their own pace.
- Students can be involved in an interactive dialogue with the ASR agent and learn through interactive sessions.

In the last decade, researchers have evaluated the use of speech recognition for interactive foreign language training [4, 5, 7, 15]. ASR has been utilized in various Computer Aided Language Learning (CALL) and Computer Aided Pronunciation Training (CAPT) systems for pronunciation training and effective speech interactions. An accurate ASR can be potentially utilized in pronunciation training systems. Chiu *et al.* [3] investigated the use of ASR in a web-based software *CandleTalk* to examine its effectiveness in speech and pronunciation in 49 Taiwanese students who were learning English. The pre-test and post-test results showed significant improvement in the oral performance. Hardison [14] conducted two studies with 26 English native speakers learning French. The first task consisted of pre-test and a post-test questionnaire where users provided feedback after exposure to a prosody training. The second task involved text recall. Improved text recall capability coupled with increased user's prosody and confidence level was observed by utilizing technology to increase practice and training activities and by increasing exposure to novel sentences. Zhao [32] tested the pronunciation of 20 English learning Chinese speakers and proved that the results returned by the ASR system were consistent with the teacher's scores at a higher level.

Even though the ASR technology is maturing at a rapid rate, the system design considerations are important when using it in a system. Some unreliability for CALL systems was indicated by Ambra [22] who reviewed various papers and systems on the use of the ASR technology for foreign language learning and concluded

that expected recognition accuracy is not achieved for non-native speakers. However, ASR systems that are deployed on a carefully designed system with intelligent feedback presentation and reliable score calculations can provide acceptable recognition accuracies. Eskenazi [7] provided some design principles regarding the development of human-computer interaction for effective pronunciation training systems and highlights the importance of providing corrective feedback at the appropriate time.

2.3 Commercial Applications using Music for Language Learning

As mentioned in the introduction, there are a large number of commercial language learning applications (e.g., Rosetta Stone, Duolingo). However, there is a small number of mobile and web applications that use music in the context of foreign language learning.

Lyrics training [7] is a website which uses song lyrics for vocabulary learning. The user is required to fill in words corresponding to the lyrics while listening to a song. When the user misses a word or enters a wrong word, the song is paused until the user fills in the correct word. This, in turn, prompts the user to listen to the song attentively. The software does not provide any definitions, meanings or translations of the words or lyrics which hinders the user from developing an understanding of the word or the context in which it is being used. *Lyrics Gaps* [8] is another music listening website to learn vocabulary and its translation. It contains multiple difficulty levels and the user is required to guess words by filling them in the blanks while listening to the song. However, if the user misses a word, the song still continues, making this software less interactive as compared to *Lyrics Training*. Both *Lyrics Training* and *Lyric Gaps* focus on listening to music and are not designed for singing.

My Lingo[9] lets the user watch music videos with subtitles in different languages. It primarily focuses on enabling people to watch a movie in a language different from their native languages. *Tubeoke*[10] is another karaoke website which requires the user to sing while the song is being played, thereby enabling the user to read out the song lyrics. However, both of these software applications are not interactive, do not provide any feedback on the performance of the user and are not aimed at providing any active language learning experience.

3 SLIONS DESIGN

Referring to the technologies discussed above, we are unaware of any existing applications or published user studies which integrates singing and language learning with ASR technology. To this end, we have designed and developed SLIONS to explore ASR-based computer-aided language learning (CALL). Our primary goals for SLIONS are:

- The user experience should be fun, engaging and easy to use so that it motivates foreign language learning.
- The application should help improve pronunciation through both instruction, feedback, practice and encouragement.
- The application should improve vocabulary by providing lyrics translations and highlighting words and phrases.

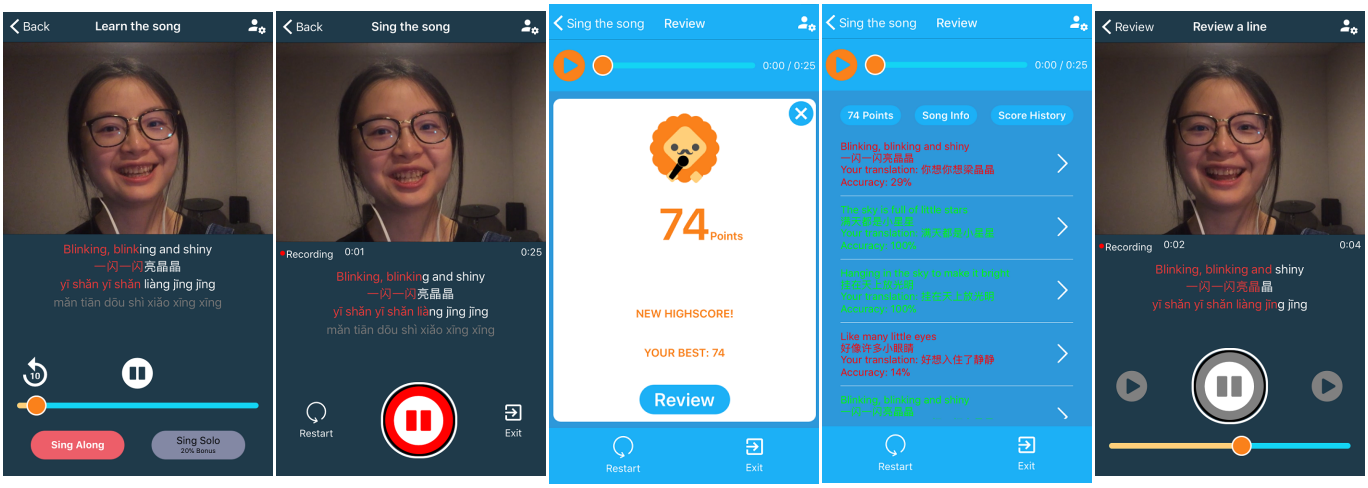
After establishing these goals, we followed the user-centered design process that began with a design review of popular karaoke and language learning applications. We then conducted multiple rounds of creating paper prototypes, whiteboard wireframes, and interactive mockups. After each round, we received feedback from experts both in language learning and human-computer interaction (HCI) as well as tested them with non-expert users. In this section, we describe some of the key ideas and outcomes of this design process while introducing our SLIONS application.

3.1 Design Considerations

Based on our discussions with experts and target users, we identified a list of important features for SLIONS:

- **Multi-Modal Instruction** Based on the different types of learners (e.g., visual, auditory, reading), different modalities are used for language learning instruction within SLIONS: audio of the vocal and backing music tracks, video of a professional singer to show vocal articulation [12], and translated text of lyrics in both foreign and native languages.
- **Speech Recognition** One of the most important features is the accuracy of our automatic speech recognition (ASR) system. While testing the singing-to-text, commercial Siri [11] and Google [12] speech recognition systems, we found the commercial Google Cloud Speech service to be more accurate during our initial testings with audio samples from the Stanford Digital Archive of Mobile Performances (DAMP) Sing! Karaoke dataset[13]. We passed good, medium and bad pronunciation singing clip samples through the ASR and observed the transcription results. Accurate transcription results were produced with good pronunciation clips while low accuracy was achieved with bad clips. More pronunciation errors were observed in the non-native speakers and even though they seemed to be aware of the melody of the song, they were not able to pronounce well due to unfamiliarity with the words used in the song. The speech accuracy for the Google Cloud Speech system is reported to have a 4.9% word error rate [26] for normal speech-to-text applications. However, we found that its performance varies for singing-to-text conversion. In particular, the performance is low when considering vocals with elongated vowels (e.g., Adele) or very fast lyrics (e.g., hip-hop) but high for a normal paced song (e.g. Let It Go - English version). Before including a song in our library, we have a professional singer record the song on a laptop or mobile phone to mimic a realistic recording environment. We then compare the output of the Google Cloud Speech system on this recording with the ground truth lyrics for the song to ensure that the word recognition accuracy is sufficiently high (i.e., less than 5% word error rate.)
- **Scoring** SLIONS provides a quantitative evaluation of pronunciation in the form of individual lyric line scores and an overall

[7] https://lyricstraining.com/
[8] https://www.lyricsgaps.com/
[9] http://www.mylingoapp.com/
[10] http://tubeoke.com/
[11] https://www.apple.com/ios/siri/
[12] https://cloud.google.com/speech/
[13] https://ccrma.stanford.edu/damp/



(a) Listen (b) Record (c) Scoring Feedback (d) Review Lines of Song (e) Review Individual Line

Figure 1: SLIONS Karaoke Application: The user first selects his or her native and foreign language (not shown). After selecting the section (e.g., chorus, verse) of a song, the user listens and learns the song through repeated listening (far left). Then the user records a karaoke performance (middle left). Based on the ASR-based word accuracy, feedback is provided to the user in the form of overall score (middle). The user then reviews the scores for each lyric line (middle right) and can select individual lines to practice and master (far right).

section[14] score. The ASR technology converts the user's acoustic signal to text and compares it against the text that has been generated from the professional singer's reference signal. The greater the match of the user's utterance with the reference utterance, the higher the score will be and is roughly equivalent to the word error rate. This gives the user a meaningful and interpretable score.

- **Feedback presentation** The users see their overall section score and individual lyric lines. They can then drill into a specific (low-scoring) line to hear the professional singer's version and their own recording of the line. They are then encouraged to practice this line (with ASR feedback given each time) until they have mastered it. Once the users have mastered all of the lines, they are encouraged to record the entire section in order to improve the overall section score.

These features taken together are intended to provide the user with an interactive experience that is both easy to comprehend and helpful in mastering the task of singing in a foreign language.

3.2 User Interface Design

SLIONS is designed to be simple and engaging to appeal to a broad audience. Users begin by creating an account, followed by choosing a language to learn. For example, a native English speaker interested in learning Chinese selects 'Learning Chinese for English Speakers'. Before proceeding to learn or practice a song in a foreign language, they can preview the song, which in turn helps the users to get familiar with the melody and difficulty level.

After selecting a particular song, users are provided with an option to select any *section* of the song they want to listen to and learn. This enables them to master the song in a systematic manner. Users then listen to a *section* sung by a professional singer along with the video of the singer singing (Figure 1a To allow users to learn the song in their own pace, they can scan the section forward or backward and play/pause it. Users who are already familiar with the song also have the option of recording the whole song without listening to it first.

Once the users listen to and learn any *section* of the song, they record a karaoke performance by selecting either one of the two modes: *'Sing Along'* and *'Sing Solo'* (Figure 1a). 'Sing Along' mode has accompanying professional vocals along with the backing track, whereas in the harder 'Sing Solo' mode, users perform with only the backing track. An additional incentive of 20% point multiplier is provided to users if they select the 'Sing Solo' mode. This encourages them to reattempt the song to get higher scores and allows them to test the ability to reproduce the song accurately without any assistance.

After selecting the mode, users can perform and record their song as shown in Figure 1b. As the users sing along, the application autoscrolls through the lyrics along with the video of the professional singer singing. The current individual words are also highlighted as the song advances. While users are singing, their utterance is evaluated line by line by comparing with a time aligned ground truth lyrics file.

Since feedback is an important element of gamification and motivation, scores are calculated in the form of word accuracy percentage. The comparison detects the number of words uttered correctly and calculates a corresponding score for it. The scoring mechanism is explained further in detail in section 3.3. These scores are then shown to the users as in Figure 1c along with the users' best score

[14] A section is a chorus, verse, bridge, etc. that consists of between 3 and 8 lines of lyrics.

with the purpose of incentivizing them to practice and improve further.

The application not only calculates the overall section score but also provide feedback for the individual lyric line. After the users receive feedback in the form of overall percentage score, they can preview each lyric line (Figure 1d) where the lines with word accuracy score less than 73% (determined empirically) are highlighted in red while the rest are highlighted in green. The application also displays the word accuracy score and the users' translation of individual lines. Users can then select an individual line for further review (Figure 1e). This lets them compare their voice next to the professional singer's voice to illuminate subtle differences in pronunciation. Users can then practice and re-record the line, after which the application provides an updated score for the line based on the new voice input from the users.

3.3 Automatic Scoring Mechanism

The scoring mechanism works by comparing the users' pronunciation results with the professional singer's corpus. We define the singing-to-text conversion of native professional singer's utterance as the *source string*, while the results for users' pronunciation is referred to as the *test string*. Source strings are generated by the Google ASR offline and stored in the database.

To compare test string to a source string, each word is first encoded with a unique numerical identifier. We then compute the Levenshtein edit distance [16] between the two sequences. The distance is increased if a word is added, removed or replaced in the test string. To calculate a score, we convert Levenshtein distance to string similarity score:

$$score(source, test) = \frac{max(len(source), len(test)) - lev(source, test)}{max(len(source), len(test))} \quad (1)$$

where $len()$ is the length of the string, $lev()$ is the Levenshtein distance, and $max()$ is a maximum of two numbers. This gives a rough approximation of the percent of correct words in the test string with a score of 1 if the strings are identical and 0 if they have no words in common.

3.4 Music Corpus

Songs are selected based on the clarity of vocals and lyrics in a song. Qualitatively, we preferred songs with lots of word repetition, simple or familiar melodies, and background music that is not overly distracting. To select the songs that can be effective for foreign language learning and specifically pronunciation improvement, the following two approaches are employed:

(1) **Choice of Songs** We incorporate three genres to target young adults for our user study: Kids, Pop, and Standards. We selected 4 English songs for English language learners: *Twinkle Twinkle Little Star* (Nursery Rhyme), *Row Row Row Your Boat* (Nursery Rhyme), *Let It Go* (Pop) and *Perfect* (Pop), and 4 Chinese songs for Chinese learners: *Twinkle Twinkle Little Star* (Nursery Rhyme), *Two Tigers* (Nursery Rhyme), *Tian Mi Mi* (Standard) and *Let It Go* (Pop). Most of the songs we selected are slow and repetitive. However, *Let It Go - Chinese version* is a faster song and was specifically included as a challenge for more advanced students.

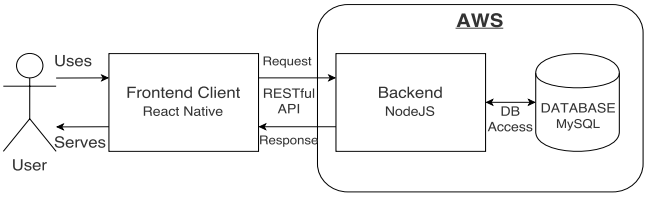


Figure 2: SLIONS Architecture Diagram

(2) **Singing transcription by ASR**: In this method, the song sung by the professional singer is transcribed by the ASR, and the transcription results are evaluated and compared with the ground truth set of lyrics. We pick songs in which a professional singer's reference vocal tracks achieve a high transcription accuracy to ensure that our scoring mechanism works well. This resulted in slow or moderate singing speeds without many elongated vowels. The degree of recognition accuracy returned by ASR is represented by the confidence values. The higher the confidence value, the higher the utterance matches with the native speaker model that corresponds to the ground truth set of lyrics. By carrying out sufficient tests with the DAMP dataset and professional singer's recordings, we ensure that the word recognition error rate is sufficiently low (i.e. < 5%).

4 IMPLEMENTATION

Figure 2 shows the architectural diagram of the SLIONS karaoke application. As with most web and mobile applications, it uses the three-tier client-server software architectural pattern, comprising the presentation tier, the application tier, and the data tier.

4.1 Presentation Tier (Front-end)

The presentation tier is the user interface layer that the user interacts with. In SLIONS, this is the mobile application that the user downloads and directly interacts with. It is implemented using React Native which allows us to build mobile applications that work on both mobile platforms (Android and iOS) using only JavaScript. A React Native application is indistinguishable from a native mobile application as the JavaScript generates the same fundamental UI views as native mobile applications, giving us the performance of a native application. Redux, an open source JavaScript library, is used to manage the application state. Most of the logic such as navigation and basic computation is performed on the client-side for the sake of performance, leaving only heavy processing and data access to the server.

4.2 Application Tier (Back-end)

The application tier is the middleware layer that handles AJAX HTTPS requests from the client while interacting with the database whenever necessary. We adopt the RESTful architectural style using ExpressJS, a lightweight and efficient NodeJS framework, to build an API server which handles functionalities such as user authentication and user data access. The API server is hosted on an Amazon EC2 instance which provides high reliability and ease of scaling. The client also interacts with a third party server, the Google Cloud Speech API, which handles the translation of users' recorded singing to text.

P. ID	Age	Gender	L1	Proficiency in L2
EN01	18-25	Male	Chinese	Intermediate
EN02	18-25	Female	Chinese	Advanced
EN03	25-34	Female	Chinese	Intermediate
EN04	18-25	Female	Chinese	Advanced
EN05	25-34	Female	Chinese	Intermediate
EN06	18-25	Female	Chinese	Advanced
EN07	18-25	Male	Chinese	Advanced
CH01	18-25	Female	Vietnamese	Intermediate
CH02	25-34	Female	English	No experience
CH03	18-25	Female	English	Beginner
CH04	18-25	Female	English	Intermediate
CH05	25-34	Female	English	No experience
CH06	18-25	Female	English	Beginner
CH07	18-25	Female	Indonesian	Advanced
CH08	18-25	Female	Vietnamese	Beginner

Table 1: Participants Demographics

4.3 Data Tier

Our data tier comprises a MySQL database and the Amazon S3 Cloud storage service. MySQL, a popular relational database, is used to store the general data such as song lyrics as well as user-specified data such as scores of past singing recordings and application usage statistics for analytical purposes. An Amazon S3 bucket is used to store media files, which includes the songs' videos and images as well as users' recordings (audio files). These static assets are distributed through Amazon CloudFront, which delivers the content through a worldwide network of data centers to ensure low latency regardless of users' geographical locations.

5 USER STUDY

We conducted a one-week pilot study to explore whether SLIONS is both engaging as application and has potential as a tool for language learning. In this section, we describe our study and share our results.

5.1 Participants

We recruited university students currently enrolled in English and Chinese courses. Seven students were learning English and eight were learning Chinese. The students were paid a small amount of money for their time and effort. Table 1 provides information about the participants. ENxx and CHxx represents the participants ID (P. ID) of English and Chinese language learners respectively, whereas the proficiency in foreign language (L2) is divided into four categories: No experience; Beginner (One semester or less); Intermediate (2-3 semesters); Advanced (4 or more semester). L1 indicates the native language of the participants.

5.2 Procedure

The following steps were performed in order to conduct the one-week user study:

Language	p	F	Mean of first trial (St. Dev.)	Mean of last trial (St. Dev.)
English	<0.00001	24.08	0.71 (0.16)	0.87 (0.15)
Chinese	0.00026	14.37	0.58 (0.24)	0.74 (0.20)

Table 2: One-way ANOVA test results for pronunciation improvement

(1) **Pre-Evaluation Vocabulary Test:** Participants were provided with a pre-evaluation vocabulary test which consisted of 20 receptive vocabulary multiple choice questions (MCQs) and each question was worth a point. The vocabulary questions were specifically designed for bilinguals in which participants choose the translation (in L1) of the provided words (in L2). We selected the vocabulary words (both correct answers and distractors) from the lyrics of the songs and the tests were reviewed by language teachers. For example, a Chinese vocabulary assessment question consists of: *Question*: Meng; *Multiple choices*: Dream; Wind; Bright; Howling.
(2) **Application Usage:** After the pre-evaluation vocabulary test, participants were provided with an installation link to the application through which they downloaded it in their mobiles from Google Play Store and Apple App Store for Android and iOS respectively. They were allowed to use the application on their own time and pace. We collected the users' data remotely by logging their clicks and activities in the database while they use the application. The logs consist of: Duration, timestamp and the number of times the user listens to the particular section; Number of times and scores of the *section* user attempts; Number of times user reviews and practices individual lyrics.
(3) **Post-evaluation Vocabulary Test:** After a week of using the application, participants were provided with a post-evaluation vocabulary test which was similar to the pre-evaluation test.
(4) **User Experience Survey:** We also asked participants to self-report about their experience with SLIONS using an online survey with both MCQs and open-ended questions about their user experience. We also interviewed a subset of the test subject to elicit additional qualitative feedback.

5.3 User Performance Evaluation

We evaluated the users' performance based on improvement in pronunciation and receptive vocabulary for each user.

Pronunciation. The system logs the pronunciation scores for each sections of the song that the participants attempted. To evaluate the pronunciation performance, we compared the first and last attempt of the participants for all the sections. The sections which the participant attempted only one time were discarded.

Vocabulary. The scores of the pre- and post-evaluation vocabulary test was compared for each user to measure the vocabulary improvement.

Table 2 shows one-way ANOVA test results that compare the pronunciation scores of the first and last attempts for each participant who used the application. As the table shows, the overall score improvement of the participants in each language is statistically significant (p<0.05). On average the English and Chinese

Language	p	F	Mean of pre-test (St. Dev.)	Mean of post test (St. Dev.)
English	0.147	2.40	19.71 (0.49)	20 (0)
Chinese	0.343	0.97	13.25 (5.28)	15.62 (4.34)

Table 3: One-way ANOVA test results for vocabulary improvement

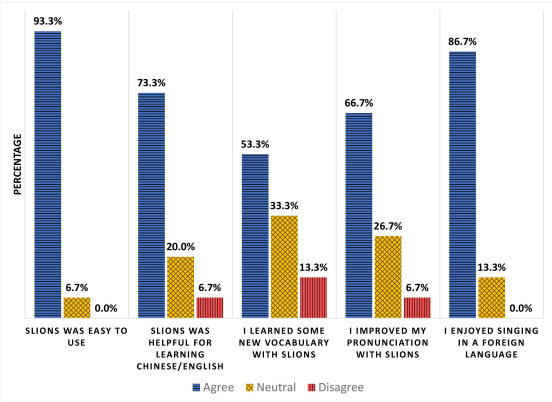


Figure 3: User Experience Results

learning participants improved 18.5% and 21.62% respectively, thus showing that there is a potential in using technological tools for learning a foreign language through songs. Though the improvement rate among the participants varies, all of the 15 participants improved their scores by the last attempt. The score improvement is statistically significant (p<0.05) for 7 out of the 15 participants.

Table 3 shows one-way ANOVA test results comparing the pre- and post-evaluation vocabulary scores for each participant. The mean scores for the pre- and post-evaluation vocabulary test show that the participants improved 1.5% and 7.5% for English and Chinese languages respectively. However, these improvements are not statistically significant, though we note that we made the English vocabulary too easy, resulting in little room for improvement. The lack of significance could also be due to participants focusing more on improving the pronunciation while singing rather than attending to vocabulary. In section 6, we describe how we plan to change our application design to focus more attention on vocabulary learning.

5.4 User Experience

Overall Experience. To enhance the user experience, SLIONS was evaluated on four main criteria: *ease of use*, *engagement*, *motivation* and *enjoyment*. After a week of using SLIONS, the participants filled out a user experience survey to rate their experience using a three-point Likert scale: Disagree, Neutral and Agree. The following questions were included in the survey:

- *SLIONS was easy to use*
- *SLIONS was helpful for learning Chinese/English*
- *I learned some new vocabulary with SLIONS*
- *I improved my pronunciation with SLIONS*
- *I enjoyed singing in a foreign language*

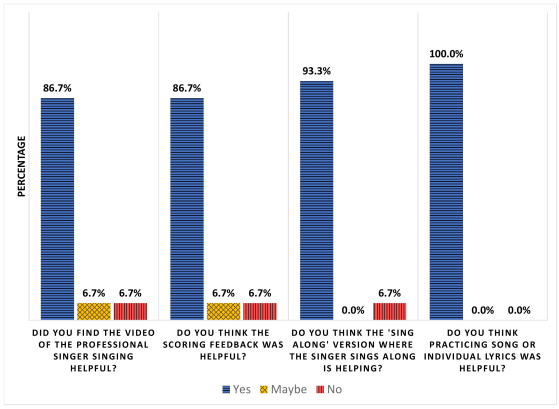


Figure 4: Feature Analysis Results

Figure 3 shows the quantitative results of all 15 participants. It shows that the majority of the participants found SLIONS to be easy to use and they enjoyed singing in a foreign language. Most participants reported that SLIONS helped them to improve pronunciation in a foreign language while fewer reported that SLIONS helped with vocabulary learning. Upon investigating the participants who thought that it did not help with vocabulary learning, they highlight that they focused more on improving pronunciation rather than learning the vocabulary through the lyric translations.

Net Promoter Score. We also calculated the Net Promoter Score[15] (NPS). Each user was asked *'how likely are you to recommended SLIONS to a colleague or friend'* and they answered using an 11-point Likert scale. The results show that 6 of the participants were detractors (0-6), 4 were passives (7,8), and 5 were promoters (9,10). This results in an NPS of -6.7 with a mean rating of 7.00 and the standard deviation of 2.34. This slightly negative NPS suggests that the SLIONS has potential but will need to be improved before we can expect it to become successful in terms of growing the user base through word-of-mouth recommendation.

Feature Analysis. On the survey, we also asked the users about specific features that we introduced in our application to make language learning effective. Figure 4 shows the overall results of the application feature analysis.

13 out of 15 participants found the video of the professional singer singing helpful. It was easy for them to follow the movement of the lips and improve their pronunciation by copying him or her. They further described the experience by stating *'It helped to enunciate words by copying them (lip-reading).'*, *'I can follow her mouth shape/movement to practice my pronunciation.'*, *'It was helpful especially for the songs we are unfamiliar with and for articulation learning.'*. The overall response from the participants supports our claim to add this feature to the application as speech perception improves substantially in the presence of a congruent face[12].

13 out of 15 participants reported that the scoring feedback was helpful. They were able to easily identify their mistakes through the feedback provided in the form of scores and translation which helped them to correct their pronunciation errors. Their overall experience with the scoring feedback was further described as *'The*

[15]https://en.wikipedia.org/wiki/Net_Promoter

scoring is quite accurate to reflect the correctness of my pronunciation. When I tried it again and again, the score improved and I finally knew the right way to pronounce a word. I am also very glad to see my improvement.'. In addition to the positive comments, we also received some recommendations on improving the feedback presentation as *'If the feedback can highlight the wrongly pronounced words, it would be better.'*. We plan to incorporate this feature in the future.

Based on the usability tests, another feature that we introduced in our application was the *'Sing Along'* version in which the users sing along with the professional singer while they perform the song. 14 out of 15 participants found this feature helpful especially when they are listening to a song for the first time and getting familiar with the melody. Some of the responses that we further received for this feature are: *'It's like an intermediate version between 'Sing alone' and 'listening'. It's helpful when I'm not familiar with the melody.'*, and *'It is helpful for me to follow the singer and learn a new song.'*.

We asked the participants to suggest the type of songs they think are suitable for beginners for language learning. Based on their experience while performing songs with varying difficulty levels, most of them suggested songs that are: Slow, have clear pronunciation, repetitive words, and familiar melody. As per their suggestions, they stated: *'Nursery rhymes, slow songs with repeating words'*; *'Slow, clear, Easy melody.'*; and *'familiar/more popular songs, like the 'Let It Go' is quite interesting because most people would know the original version.'*.

SLIONS also allows participants to practice parts of the songs and individual lyrics repetitively. All the 15 participants found this feature to be very helpful in improving their pronunciation. Along with providing a practicing platform, practicing individual lyric lines saved their time as compared to practicing the whole song. It also provided them a progressive learning opportunity as stated by one of the participants: *'Individual lyrics helped in zooming into specific problematic sections and practicing them, thus helping in improving the entire song.'*.

Singing for Language Learning. Finally, we asked the participants if singing in a foreign language is helpful for language learning independent of their experience with SLIONS. 13 of the 15 participants agreed and indicated that singing in a foreign language has additional benefits. Some found that singing helped overcome the barriers of speaking out loud. Singing also made it easier to learn and recall the words since they get associated with the tune of the song. Some of the participants found that singing songs in a foreign language made the learning process fun and enjoyable and students reported gaining more interest in learning the new language. As stated by the participants: *'Singing is a fun way of learning. I feel I will practice more by singing, than in the normal way. But at the same time, if the song is too tough or fast, it is distracting. So singing with simple and repetitive tunes are more effective, at least for an absolute beginner like me.'*; and *'Learning foreign language might be boring, karaoke makes it more fun.'*.

6 DISCUSSION

Overall, the participants reported a positive experience with SLIONS in terms of ease-of-use, enjoyment, and educational potential. They confirmed that many of our specific design ideas (e.g., feedback mechanisms, multi-modal instruction, music corpus) were beneficial and should continue to be included in the future versions of our application. We also received valuable feedback and helpful recommendations for our future work. For example, most of the participants felt that the fast song 'Let It Go' in Chinese was too difficult and not a good choice for beginners learning a foreign language. Others thought that many of the (English) songs were too easy and repetitive. Having a designation of 'Easy', 'Medium', and 'Hard' might fix this problem. Almost all of the participants requested that we have a larger music corpus which is also something that we plan to do in future iterations.

Our results suggest that singing combined with ASR-based computer-aided language learning (CALL) has the potential to help improve pronunciation. While vocabulary did not show significant improvement, we suggest that this is because our interactive design only passively made lyric translations available to users. Our future work includes incorporating vocabulary games as is typical in gamified language learning applications so that users can also improve the vocabulary along with pronunciation. However, our vocabulary words would come from the song lyrics and may be easier to recall due to the memorable acoustic cues that come from the singing [19, 31].

The major limitation of our work is that we conducted a small, short-term pilot study on a prototype application. In particular, testing the efficacy of the language learning would require a longer-term study with a large population of test subjects. In addition, we would want to use an experimental design that tests language learning outside the context of our application [11, 30]. For example, we would want human evaluators to assess the speaking ability of each test subject before and after our testing period. We would also need to compare our singing-based approach with other approaches (e.g., gamified language learning application, traditional paper-and-pencil homework) to test efficiency. We suspect that our approach may be slightly less efficient but perhaps more motivating.

Our next step is to refine SLIONS, grow our music corpus, and develop a cloud-based dashboard that teachers can use to monitor a class of students. We plan to conduct a long-term study with primary and secondary students in a classroom setting. The idea is to have teachers assign SLIONS as an alternative to typical homework assignments. We would then measure differences in pronunciation, vocabulary, and enthusiasm for language learning. Finally, we plan to make SLIONS widely available for public use so that we can collect additional feedback from a larger and more diverse set of users.

ACKNOWLEDGMENTS

The authors would like to thank Brenda Yuen Pui Lam from Centre for English Language Communication (NUS) and Lin Chiung Yao from Centre for Language Studies (NUS) for their valuable suggestions and support in recruiting language learning students. We are also grateful to Jovin Lew for his contribution in the UI design and Shuqi Dai for her participation in the singing recording. The work is supported by the ALSET from National University of Singapore under the grant R-252-000-696-113. D. Turnbull is supported by NSF grant IIS-1615679.

REFERENCES

[1] Sima H Anvari, Laurel J Trainor, Jennifer Woodside, and Betty Ann Levy. 2002. Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of experimental child psychology* 83, 2 (2002), 111–130.

[2] Kenneth Beare. 2017. How Many People Learn English? Web Article. (18 September 2017). Retrieved April 07, 2018 from https://www.thoughtco.com/how-many-people-learn-english-globally-1210367

[3] Tsuo-Lin Chiu, Hsien-Chin Liou, and Yuli Yeh. 2007. A study of web-based oral activities enhanced by automatic speech recognition for EFL college learning. *Computer Assisted Language Learning* 20, 3 (2007), 209–233.

[4] David Coniam. 1999. Voice recognition software accuracy with second language speakers of English. *System* 27, 1 (1999), 49–64.

[5] Farzad Ehsani and Eva Knodt. 1998. Speech technology in computer-aided language learning: Strengths and limitations of a new CALL paradigm. (1998).

[6] Dwayne Engh. 2013. Why Use Music in English Language Learning? A Survey of the Literature. *English Language Teaching* 6, 2 (2013), 113–127.

[7] Maxine Eskenazi. 1999. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. (1999).

[8] Judith Weaver Failoni. 1993. Music as Means To Enhance Cultural Awareness and Literacy in the Foreign Language Classroom. *Mid-Atlantic Journal of Foreign Language Pedagogy* 1 (1993), 97–108.

[9] Douglas Fisher. 2001. Early language learning with and without music. *Reading Horizons* 42, 1 (2001), 39.

[10] Jorge Francisco Figueroa Flores. 2015. Using gamification to enhance second language learning. *Digital Education Review* 27 (2015), 32–54.

[11] Arla J Good, Frank A Russo, and Jennifer Sullivan. 2015. The efficacy of singing in foreign-language learning. *Psychology of Music* 43, 5 (2015), 627–640.

[12] Ken W Grant and Philip-Franz Seitz. 2000. The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America* 108, 3 (2000), 1197–1208.

[13] Chitralekha Gupta, Haizhou Li, and Ye Wang. 2017. Perceptual Evaluation of Singing Quality. In *Proceedings of APSIPA Annual Summit and Conference*, Vol. 2017. 12–15.

[14] Debra M Hardison. 2004. Generalization of computer assisted prosody training: Quantitative and qualitative findings. (2004).

[15] Rebecca Hincks. 2003. Speech technologies for pronunciation feedback and evaluation. *ReCALL* 15, 1 (2003), 3–20.

[16] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. 707–710.

[17] Karen M Ludke, Fernanda Ferreira, and Katie Overy. 2014. Singing can facilitate foreign language learning. *Memory & cognition* 42, 1 (2014), 41–52.

[18] Erin McMullen and Jenny R Saffran. 2004. Music and language: A developmental comparison. *Music Perception: An Interdisciplinary Journal* 21, 3 (2004), 289–311.

[19] Suzanne L Medina. 1990. The Effects of Music upon Second Language Vocabulary Acquisition. (1990).

[20] Susan Bergman Miyake. 2004. Pronunciation and music. *Sophia Junior College Faculty Bulletin* 20, 3 (2004), 80.

[21] Carmen Foncesa Mora. 2000. Foreign language acquisition and melody singing. *ELT journal* 54, 2 (2000), 146–152.

[22] Ambra Neri, Catia Cucchiarini, and Wilhelmus Strik. 2003. Automatic speech recognition for second language learning: how and why it actually works. In *Proc. ICPhS*. 1157–1160.

[23] Murray Newlands. 2016. Smule Has Changed The Music Industry Completely: Here's How. Web Article. (20 September 2016). Retrieved April 07, 2018 from https://www.forbes.com/sites/mnewlands/2016/09/20/smule-has-changed-the-music-industry-completely-heres-how/

[24] Hyacinth S Nwana. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review* 4, 4 (1990), 251–277.

[25] Pandaily. 2017. With over 400 Million Users, How did Tencent's WeSing Make a Fortune? Web Article. (20 August 2017). Retrieved April 07, 2018 from https://pandaily.com/with-over-400-million-users-how-did-tencents-quanmin-k-ge-make-a-fortune/

[26] Emil Protalinski. 2017. GoogleâĂŹs speech recognition technology now has a 4.9% word error rate. Web Article. (17 May 2017). Retrieved April 07, 2018 from https://venturebeat.com/2017/05/17/googles-speech-recognition-technology-now-has-a-4-9-word-error-rate/

[27] Andrés Roberto Rengifo. 2009. Improving pronunciation through the use of karaoke in an adult English class. *Profile Issues in TeachersProfessional Development* 11 (2009), 91–106.

[28] Daniele Schön, Maud Boyer, Sylvain Moreno, Mireille Besson, Isabelle Peretz, and Régine Kolinsky. 2008. Songs as an aid for language acquisition. *Cognition* 106, 2 (2008), 975–983.

[29] Wang Shudong and Michael Higgins. 2005. An online pronunciation training support system designed for Japanese learners of English. In *Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on*. IEEE, 171–173.

[30] Roumen Vesselinov and John Grego. 2012. Duolingo effectiveness study. *City University of New York, USA* (2012).

[31] Wanda T Wallace. 1994. Memory for music: Effect of melody on recall of text. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20, 6 (1994), 1471.

[32] Hongyan Zhao. 2017. Study on the Effectiveness of the ASR-Based English Teaching Software in Helping College Students' Listening Learning. *International Journal of Emerging Technologies in Learning (iJET)* 12, 08 (2017), 146–154.