Verbose : Designing a Context-based Educational System for Improving Communicative Expressions

AKASH CHAUDHARY, Indraprastha Institute of Information Technology, Delhi, India

MANSHUL BELANI* and NAMAN MAHESHWARI*, Indraprastha Institute of Information Technology, Delhi, India

AMAN PARNAMI, Indraprastha Institute of Information Technology, Delhi, India

ESL (English as a second language) speakers tend to follow the tone structure of their first language, making their speech difficult to understand for native speakers, thereby limiting their opportunities for education and employment. To address this problem, we build an interactive smartphone-based educational mobile application using the user-centered design process. This application teaches English intonations based on globally consistent pitch patterns through conversations with a trained chat assistant, which inculcates expert linguists' teaching principles. After co-designing the application's parameters with primary stakeholders and expert visual designers, we assess its effectiveness by measuring the pre and post-performance of the users after the system usage, using various quantitative measures, like intonation scores, SEQ, and SUS. Feedback from users suggests that ESL speakers find significant improvement in the perception of their vocal expressions, thereby highlighting the necessity of such a system in improving the quality of conversations that people have in general.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*; *Usability testing*; **Sound-based input / output**; **Auditory feedback**; *Natural language interfaces.*

Additional Key Words and Phrases: Learning application; Context-based learning; Stress-timed language; Intonations; Communicative expressions

ACM Reference Format:
Akash Chaudhary, Manshul Belani, Naman Maheshwari, and Aman Parnami. 2018. Verbose : Designing a Context-based Educational System for Improving Communicative Expressions. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

The world today has become a global hub of people migrating across the world searching for different opportunities. With the rampant increase in cross border migration lately [7], it has become important for people to have a common language for communication purposes. Including all migrants, around 400 million people worldwide speak English as their second language [5]. It has been proven that most of these ESL speakers have difficulty in articulation and perception of the English language ([12, 24]; [47], p. 81; [33], p. 283). This happens because non-native ESL speakers

*Both authors contributed equally to this research.

tend to speak English with a regular pattern of syllabic duration [26], which is monotonous and does not signify intentionality in communicative expression.

Languages can be classified either as stress-timed or syllable-timed ([40], p. 72-79; [20], p. 10; [27], p. 51-62). Although no language is completely stress-timed or syllable-timed, they fall towards any one end of the spectrum. This continuous classification spectrum scale that gives rise to the different expressions in various languages. English is a predominantly stress-timed language [26, 31, 48], which means that different parts of an English sentence are spoken by giving pitch-related stresses, called intonations. However, many varieties of English, are characterized as nearly syllable-timed language [48] due to the different native languages of people like French, Spanish, Hindi and Urdu, which are syllable-timed languages [31, 44]. This leads to the rise in a phenomenon called Diglossia [6], where people in a community speak different variants or dialects of English. This causes people with non-native English language to be perceived by others as having reduced intelligibility and understanding of what they speak ([12, 24]; [47], p. 81; [33], p. 283).

This becomes a barrier for non-native English speakers in communicating properly in English, thus hampering their opportunities in job interviews, public speaking, higher education, etc. Acquiring stress-timed rhythmic patterns, or intonations, is considered one of the most difficult aspects of grasping English pronunciation [39]. Further, the average cost of learning spoken English through standard organisations like the British Council involves a cost of around $ 200 [1], which is prohibitive and an obstacle for people in relatively economically poor countries. Therefore, there arises a need to design an effective and easily adaptable pedagogy that mobile application creators can use to make a system that teaches non-native English speakers so they can communicate effectively.

Currently, professional English teaching linguists teach intonations through an imitation-based teaching pedagogy[2]. We changed and adapted this existing imitation-based teaching pedagogy to conversation-based teaching pedagogy, by adding an extra element of context. This was done by adding extra sentences before any sentence to be taught and curate them as dialogues going on between two people to make them seem like part of a conversation. These precursor sentences set up the context and are spoken by a chat agent who is well trained in intonations in advance. This context or scope of a sentence is used to get hints for raising or dropping pitch at different parts in a sentence.

Our mobile application allows users to learn different types of English intonations and their application in different contexts based on the relative difference in pitch levels while speaking. We use an iterative design process to make this mobile application with an overall total of 34 participants from our university in 3 user studies, excluding dropouts. We test our system for intonation performance scores, usability, and overall experience of the participants through a quantitative and qualitative analysis. Following are our contributions through this research-

- We present a conversation-based mobile application, Verbose, to assist ESL speakers in the training of context-based pitch modulation, also known as intonations.
- We further provide a novel and effective visual representation for showing pitch/intonations mapped to the orthographical system of English writing.
- We finally present our learnings and insights from qualitative observations and interviews performed with users, which provide support for the effectiveness and usability of Verbose in training ESL speakers on intonations.

2 BACKGROUND

This section highlights the important language-related literature used in teaching English intonations through our system.

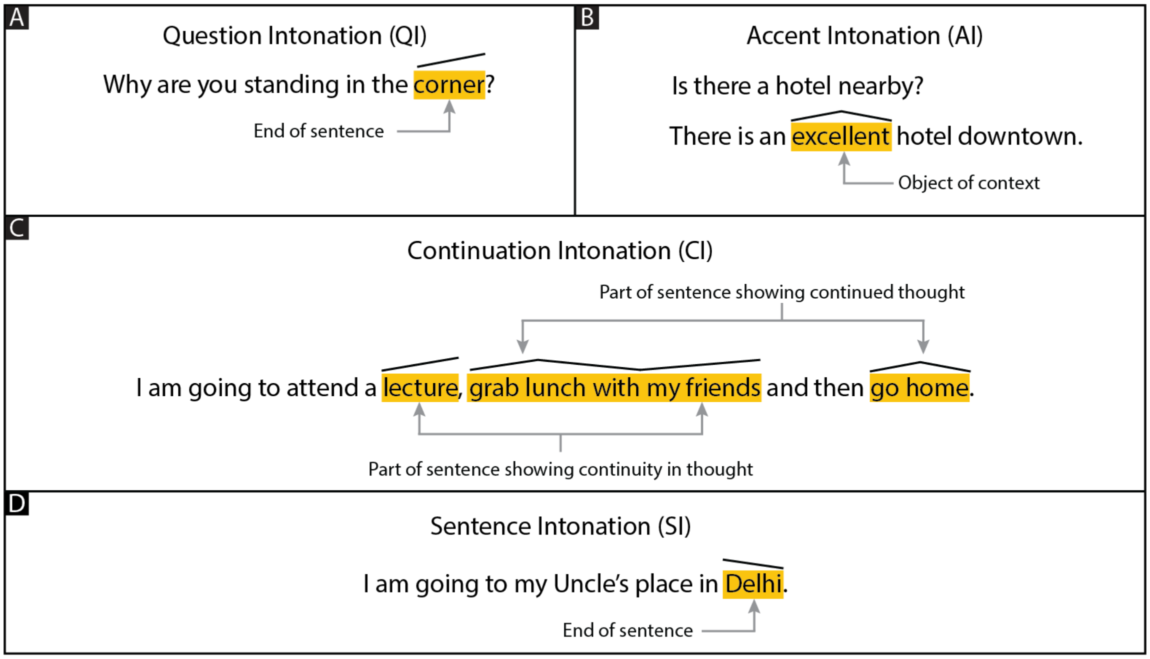

A
Question Intonation (QI)
Why are you standing in the corner?
End of sentence
B
Accent Intonation (AI)
Is there a hotel nearby?
There is an excellent hotel downtown.
Object of context
C
Continuation Intonation (CI)
Part of sentence showing continued thought
I am going to attend a lecture, grab lunch with my friends and then go home.
Part of sentence showing continuity in thought
D
Sentence Intonation (SI)
I am going to my Uncle's place in Delhi.
End of sentence

Fig. 1. Four different types of intonations shown along with the places (black line above the text) where they are used in a sentence. a) QI - Relative increase in pitch towards the end of sentence. b) AI - Relative increase and then decrease in pitch at object of context. c) CI - Relative increase in pitch before a continuity in thought, relative increase and then decrease in pitch at the continued part of previous thought, d) SI - Relative decrease in pitch at the end of sentence

2.1 Different types of pitch intonations

We explored the various communicative speech expressions used by people who speak English and found four intonations (patterns of pitch modulation, Figure 1) in the domain of pragmatics (branch of linguistics which works on practical findings of intentions in speech) which are globally accepted across various versions of English like Australian, British, American, etc. ([33], p. 292-296).

3 RELATED WORK

We found two types of system which provide feedback on the different parameters of communication. We classify these systems as synchronous and asynchronous feedback systems. Further, we list down the various ways in which attempts have been made to represent pitch visually.

3.1 Synchronous Feedback Applications

Prior HCI research work has investigated the development of various tools to assist users on speech production and semantic-related parameters synchronously for enhancing their social communication skills. Applications like Rhema [50], Logue [25], AwareMe [16], Aging and Engaging [9], ROCSpeak [55], RoboCOP [51], VoiceCoach [53], MACH [29], and Automated Social Skills Trainer [49] provide feedback on user speaking rate [17, 25, 29, 34, 49–51, 53], as well as on paralinguistic speech functions such as amplitude [9, 17, 50, 53, 55], pauses [29, 49, 53], and pitch and filler words [17, 29, 49, 51]. Our system provides feedback on the paralinguistic parameter of pitch. However, it differs from the above mentioned systems in that we don't just provide feedback on pitch, but on certain patterns of pitch, which represent contextual meaning in a sentence. These contextual meanings are accepted globally in almost all accents of English ([33], p. 292-296).

Other applications like Presentation Sensei [34], RoboCOP [51], Aging and Engaging [10], MACH [29], Aging and Engaging [10], ROCSpeak [55], and Presentation Trainer [46] also provide instantaneous feedback on parameters other than paralinguistics, like eye contact, smile, head gestures, and body movements.

All these speech systems are useful in improving a user's solo speech but are not very helpful in improving conversations with other people. For improving conversations, a system needs to provide contextual conversation-based training of speech production. Only the applications, Automated Social Skills trainer [49], and Aging and Engaging [9] provide conversation-based synchronous teaching. However, the quality of communication skills in these systems is determined by just the objective feedback of the speech functions, as it is difficult to give feedback on context synchronously. The context in which a person speaks shapes many of these speech functions which is missing from these feedback systems. Hence, it is difficult to employ a synchronous conversation-based improvement system in practice for improving communications. Therefore, we concentrate on providing accurate and reliable asynchronous feedback on the context-dependent feature of pitch intonation in our system.

3.2 Asynchronous Feedback Applications

Asynchronous applications are systems where an application is given enough time to pre-set a certain content which it can analyse efficiently and accurately. The user is supposed to speak the already curated pre-set content and receive feedbacks on it. The most famous asynchronous feedback app world-over for language learning is Duolingo [52]. Though Duolingo falls within the scope of language learning, it is usually used for teaching grammar and vocabulary within and across various languages. Here we investigate the systems that have been deployed in asynchronous modes for people to enhance their English speaking skills in intonations. Asynchronous learning applications have ranged from imitation-based systems to conversation-based systems.

CALL (Computer Assisted Language Learning) [18] involves speech recognition in employing imitation-based learning techniques to develop listening and speaking skills among users. MyET (My English Tutor) [37] is another imitation-based interactive learning application that allows learners to listen to teachers of different accents, imitate their teacher's pronunciation, and improve their English listening and speaking skills in the process. SLION [38], a karaoke application also shows imitation-based singing combined with automatic speech recognition (ASR) to have pronunciation enhancing potential for learning foreign languages.

However, conversation-based teaching technique has shown to be very effective in improving pronunciation learning amongst users [41]. Applications like TRACI (Teacher Ranging Across the Computer Interface) talks, Caroline in the City, CNN Interactive English, The Syracuse English Comprehensive Learning Series, Tell Me More Pro, and Encarta Interactive English Learning, all use role-play based activities and task-based conversations to improve user's speaking skills [18]. While these applications use a conversation-based teaching pedagogy to train the speakers on "how" and "where" to speak, none of them provide any feedback on "why" to speak in such a way. Our application uses a contextual conversation-based pedagogy to make the user understand the correct type, place and reason for using an intonation, therefore covering the "how", "why" and "when" of intonations. This is important for a holistic learning process where the educational technique covers all possible aspects of user speculation while learning, as evidenced through the study.

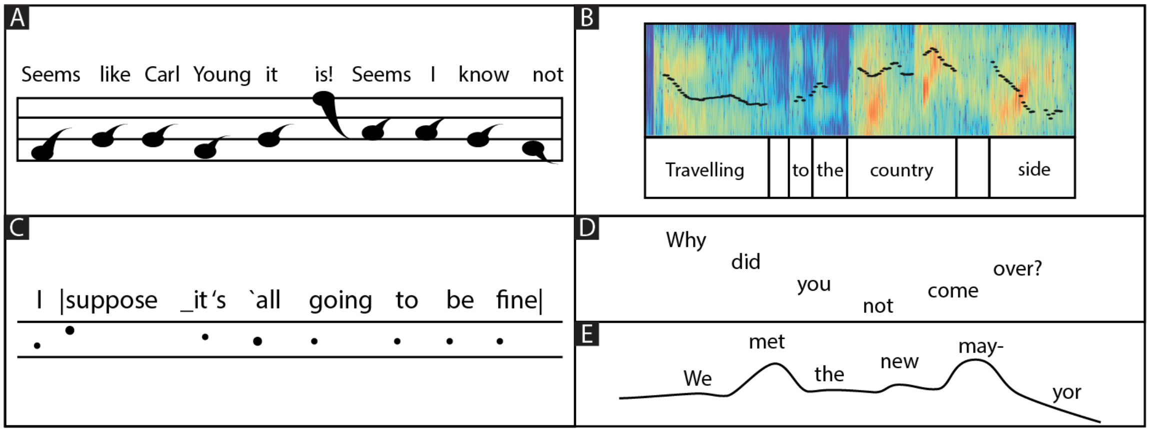

A
Seems like Carl Young it is! Seems I know not
B
Travelling
to
the
country
side
C
I |suppose _it's `all going to be fine|
D
Why
did
you
not
come
over?
E
We
met
the
new
may-
yor

Fig. 2. The various ways in which pitch has been represented textually. A) Notation presented by James Rush, B) Notation given by Lieberman tracing the 10th harmonic of a spoken sentence, C) Notation presented by Crystal, D) Notation presented by Bolinger, E) Notation presented by Ladefoged

3.3 Visual Representation of Pitch

It is important to find an intuitive representation of pitch for learning intonations because a visual representation provides hints on where and how to modulate your speech. This information can also be received by simply listening and imitating the raw audio, but this information can easily be lost if one doesn't concentrate hard enough. Hence, a visual representation complements the raw audio when trying to use intonations. Here, we investigate the various ways in which pitch has been visualized in previous works.

Although a universal agreement has been reached to represent all possible human speech sounds through International Phonetic Alphabet (IPA) [11], little consensus has developed for the representation of stress or pitch modulation ([35], 2006). One early representational approach is from James Rush ([42]), where he used the musical pitch scale to represent the pitch and transcription of spoken text written above it to represent the place of pitch (Figure 2A). Lieberman ([36], 1967) displayed pitch curves by tracing the 10th harmonics of audio signals on narrowband spectrograms by extracting their fundamental frequency. These types of pitch curves are commonly represented by continuous curves representing pitch modulation over time as these values change continuously while one speaks (Figure 2B). Crystal ([23]) used curved lines on pitch range spaces represented by icons to map the meanings of intonations (Figure 2C). He further utilized "large and small dots, capitalization, arrows, dashes, and two kinds of accent marks (grave and acute), along with curved lines placed in a vertical space" to represent the various intonations ([22]). One particular drawback of these types of representations is that they have at least two separate layers (one for pitch representation and one for words) within the representation to encode the meaning of speech which is not very cogent for reading.

Bolinger ([14]) used examples of sentences with prominent syllables and words written relatively up or down to represent pitch curves and their displacements (Figure 2D). Ladefoged ([35]) combined pitch curves obtained through pitch extraction algorithms with linguistic units of speech written along the curve. These spoken units were used to track intonations embedded over them (Figure 2E). Although the system provided by Bolinger used a single layer of visual representation to encode speech meaning, it is not a straight line system of representation and hence, not very feasible for reading large speech corpora. Similarly, the system provided by Ladefoged, though continuous and much more legible than the previous notations, again disrupts the flow of reading by not being in a straight line.

We systematically come up with various visual representations that primarily follow a straight line orthographic system with minimal iconic pitch representations embedded over it. We then determine the most effective representation out of these with the help of visual communication design experts as detailed in the study (Figure 3).

4 RESEARCH APPROACH

We hypothesize that a conversation-based feedback system that teaches people how, where and why to modulate pitch according to different contexts can help ESL speakers in terms of their speech expressions and communication with other people.

To explore the hypothesis, we developed a mobile-based application using a 3-step iterative process with the help of two formative studies and evaluate the various design aspects of our system with the help of a summative study. First, a visual representation study found an intuitive representation of pitch through the orthographic system of English writing. Second, a low-fidelity sketch-based mock-up study defined our content design and the basic app flow. Third, user testing on a mobile application evaluated the system on intonation performance scores, usability, and ease of use. Along with this, we use detailed insights of users to evaluate our system. We discuss these studies in detail in the following sections.

5 PREREQUISITE DESIGN REQUIREMENTS

As observed from the state of existing applications used in teaching communicative expressions, we find two prerequisite design requirements for our final mobile-based application. We denote the design requirements for the mobile application with notation Rx throughout the study.

R1 - **Conversation based pedagogy** - We chose a trained member of the research team as the human voice behind the Verbose (name of our mobile application) chat agent who could properly intonate and not a computer-synthesized voice as we wanted the interaction to be as engaging as possible.

R2 - **Context-based pitch modulation** - The design of our teaching pedagogy is directly adapted from the teaching methodology used by professional English teaching linguists [2] to teach intonations, wherein a sentence is divided into units, and taught through imitation. We changed this methodology from an imitation-based to a conversation-based teaching pedagogy, by adding precursor information before these sentences when used in conversations ([47], p. 61). This defines the context or scope of the sentence and this context-specific information is used to raise pitch at different parts in that sentence to know "how" and "where" to stress by imparting different paralinguistic meanings to the speech ([8, 21]; [47], p. 6, p. 9). For example, if Jane says, "Where are you going?", Martha says, "I am going to the *hotel* at noon. However if Jane says, "When are you going?", Martha says, "I am going to the hotel at *noon.*

Initially, we needed to represent the place and type of pitch modulations. Hence, we conducted a visual representation user study which informed the visual representation of intonations within the system.

6 VISUAL REPRESENTATION TESTING

The visual representation of the intonations informs the user on how to speak a certain part of the sentence without actually listening to it. We determined a novel and effective visual representation for different types of intonations with the help of a study conducted with design practitioners. This visual representation of intonations acts as an important cue for users to understand the place and type of intonation to be used in a particular context.

The following parameters were considered while designing the visual representations -

- We use an orthographic system of language representation and not a phonetic representation with the assumption that the users are more aware and responsive to such a representation.
- The representation system highlights the tonic unit as well as the tonic syllable of the required intonation. The tonic unit in a sentence determines the phrase where stress is to be used, while the tonic syllable within that unit determines the exact syllable that is required to be stressed, thereby indicating the type of intonation
- We consider using all the available highlighting options, namely text color, bold, italics, underline and background color which are available within any standard text editor like Google Docs, along with gradient text color, to highlight the pitch target part.
- We further consider using either relative change in text size and graphical representation superimposed on the orthographic representation, to show a relative change in pitch among the syllables of a word. This design criterion is assumed to help a user know the exact place of pitch modulation.

We developed 12 visual representations (Figure 3) from the combination of the above parameters to represent the stress or emphasis in a sentence for a user to intonate.

6.1 Recruitment

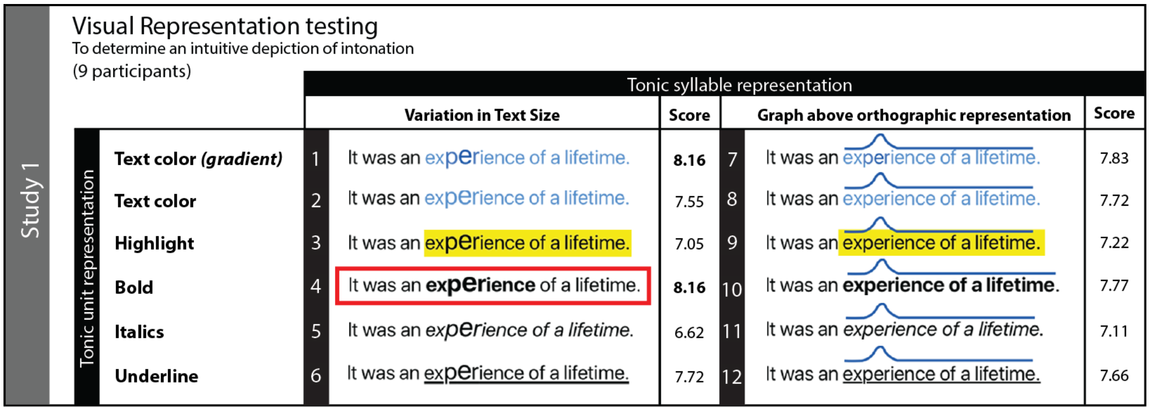
1 visual communication design expert with 30 years of experience from our university, assisted us in the development of the intonation representations by providing us the above mentioned four parameter for designing and 10 students studying in the same domain helped in their eventual evaluation.

6.2 Procedure

The 12 visual representations were given in a randomized order to 10 students of visual design along with their respective audio clips. They were then asked to rate the closeness of the audio recordings to each of the representations (on a scale of 1 to 10) and point out the mismatch between the two if any.

6.3 Results

Only 9 out of the 10 contacted students gave their feedback on the 12 representations.



Visual Representation testing
To determine an intuitive depiction of intonation
(9 participants)

Tonic unit representation		Tonic syllable representation				
		Variation in Text Size	Score		Graph above orthographic representation	Score
Text color *(gradient)*	1	It was an experience of a lifetime.	**8.16**	7	It was an experience of a lifetime.	7.83
Text color	2	It was an experience of a lifetime.	7.55	8	It was an experience of a lifetime.	7.72
Highlight	3	It was an experience of a lifetime.	7.05	9	It was an experience of a lifetime.	7.22
Bold	4	It was an **experience** of a lifetime.	**8.16**	10	It was an **experience of a lifetime**.	7.77
Italics	5	It was an *experience of a lifetime.*	6.62	11	It was an *experience of a lifetime.*	7.11
Underline	6	It was an experience of a lifetime.	7.72	12	It was an experience of a lifetime.	7.66

Fig. 3. The details of the formative study used to find visual representation of pitch. The figure also shows the average scores given by users when presented with the visual representations in a randomized order

1. The average rating for both bold/variation in text size and text color (gradient)/variation in text size both came out to be 8.16/10, which meant that they were the most intuitive representations for highlighting the tonic unit and tonic syllable within a sentence.

2. The lowest average rating (6.62/10) was for italics/variation in text size, which meant that the participants did not consider italicization to be a very intuitive way of highlighting pitch visually.

3. *Text size variation vs graphs* - 5/9 users found text size variation as a better way to highlight the place of stress in comparison to drawing a graph above the text. The uniform text size, alongwith the color of the tonic unit and a graph representing the change in pitch above the text confused the participants as they felt both of them to be conveying different levels of stress (Figure 3). Most users also found the graph as an additional element spoiling the overall reading experience requiring additional interpretation and found it to be unintuitive in terms of highlighting the exact letters to be stressed.

Additional comments -

4. 2 design students found the highlighting of the tonic unit (with any of the highlighting techniques-text color, bold, italics, underline, background color, color gradient) to be confusing since stress is supposed to be imparted only on certain syllables in the tonic unit and some of the highlighted tonic unit parts showed no pitch modulation as such. Our intention behind highlighting tonic unit was to indicate that the complete tonic unit is to be spoken continually with pitch modulation without taking any pause, however, highlighting did not convey the intention clearly. Therefore, we highlighted only the word whose syllables needed to be intonated and not the entire tonic unit, for accent, sentence, and question intonation. For continuation intonation, the stress usually starts on the syllables of one word and ends on the syllables of another word, and therefore for this intonation, we highlight from the starting word to the ending word to avoid any confusion.

5. 3 design students found highlighting the tonic syllable with any color (blue in our case as used by us) other than black to be confusing. They felt that blue being a lighter color than black can also imply more stress on the black text part (text that is supposed to have no intonation). Therefore, we only used black as our text color consistently.

We validated the above mentioned two changes to our representations obtained from the additional comments provided by the design students by consulting with the design expert.

6.4 Design requirements

R3 - Out of the two highest-rated representations (text size variation with - color gradient and bold), we chose the representation that uses **bold** to highlight the words that need to be intonated, and **variation in text size** to indicate the tonic syllable and type of intonation to be used in the system (Figure 3). To increase the contrast, we reduced the font-weight of the text not to be intonated in the final representation.

Before finally creating the mobile application, we prepared sketch-based mockups and tested them among university students to get our final design requirements.

7 LOW-FIDELITY TESTING THROUGH SKETCH-BASED MOCK-UPS

The following study informs the design elements involving the content design and basic application flow used to create the final system. We used wire-framing to develop the low-fidelity prototype.

After consenting to the study, the participants were requested to go through the flow of the application screen by screen, interacting while following a think-aloud protocol. At the end of the study, we asked participants some open-ended questions about their experience with Verbose. Each participant session lasted for 45 minutes on average.

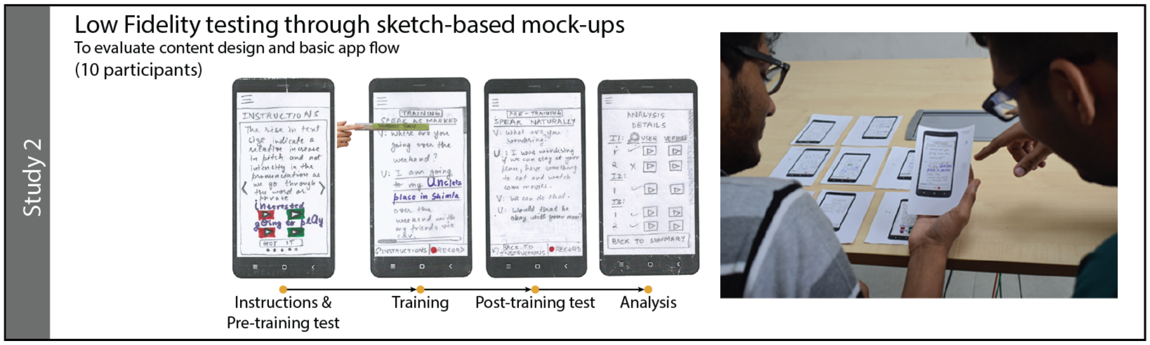

Study 2
Low Fidelity testing through sketch-based mock-ups
To evaluate content design and basic app flow
(10 participants)
Instructions &
Pre-training test
Training
Post-training test
Analysis

Fig. 4. The details of formative low-fidelity study done to evaluate the content design and basic app flow

We voice recorded all user sessions and transcribed them for analysis. We then open-coded all the transcriptions for identifying emergent themes.

7.1 Design requirements

From the transcribed user observations, we initially came up with 7 design requirement themes, but eventually converged to 5 (R4 to R8), as presented in Table 2.

Design Requirements		
Rx	***Description***	***Description***
R1	Conversation-based pedagogy	A trained member of the research team who can intonate properly speaks as the chat agent
R2	Context-based pitch modulation	Precursor sentences before a target sentence define context in a conversation between the participant and chat agent
R3	Visual representation of pitch	Bold highlights the words of intonation and variation in text size indicates the place and type of intonation
R4	Introductory/concluding screens	Introductory screen presents conceptual facts on intonations and concluding screen rationalizes why to use a particular intonation (7/10)
R5	Gradual increase in difficulty	Gradual increase in number of intonations and difficulty of a training sentence enable ease of learning (10/10)
R6	Gamify the learning process	Red and green visual circles provide instant feedback. (n+1) minimum attempts followed by a hint provide optimal challenge (5/10)
R7	Personalized training metrics	Participant's own result achieved in a previous session provides benchmark for future performances (6/10)
R8	Holistic analysis	An intonation-wise, sentence-wise and overall overview of the entire session provide a holistic analysis (5/10)

Table 1. Design requirements as observed from the formative studies. The information within the parenthesis for requirements R4 to R8 indicate the number of users out of 10 who suggested these requirements.

While inculcating the design requirements into the mobile application specifications, we met all the 8 requirements (Rx) above and mention them in parenthesis wherever necessary in the technical implementation and analysis of our system.

8 TECHNICAL IMPLEMENTATION

The technical implementation of our system contains four phases - voice input phase, pitch extraction phase, intonation analysis phase, and visual feedback phase.

Voice input phase - The iOS application initially records the users on two sentences, one in their natural style and one in which they imitate the Verbose chat agent. This is done to set the base, raised and lowered threshold limits of their intonation utterance according to their own personalized speaking limits (**R7**). Once the limits are set, all the user recorded audio files are sent to the server running Flask for intonation analysis.

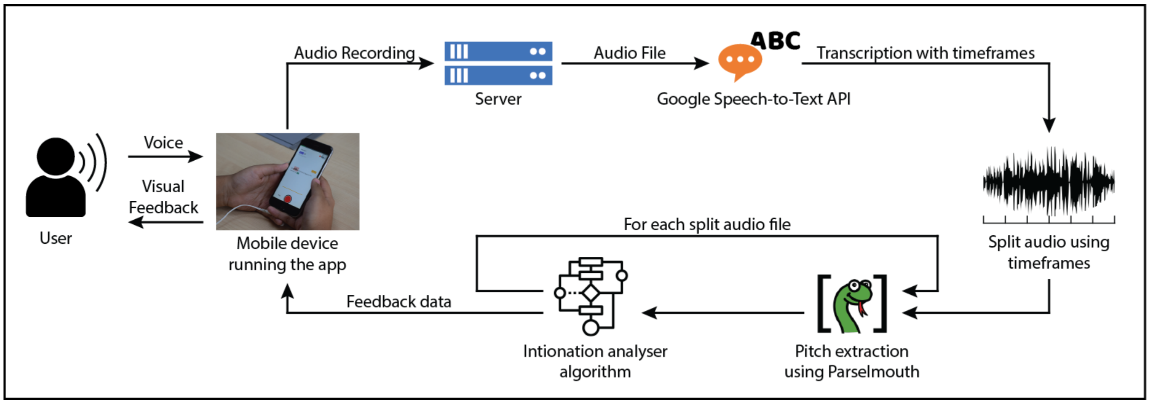

Audio Recording
Server
Audio File
ABC
Google Speech-to-Text API
Transcription with timeframes
Voice
Visual Feedback
User
Mobile device running the app
Split audio using timeframes
For each split audio file
Feedback data
Intonation analyser algorithm
Pitch extraction using Parselmouth

Fig. 5. System architecture

Pitch extraction phase - At the server, we maintain an individualized user log for each session. This is done so that in a repeat training session, the results for a user are compared from his/her performance over the last session (**R7**). We then extract pitch values from the audio file using Parselmouth [30]. The audio file is then sent to Google Cloud Platform for transcribing it from Speech to Text. The transcription from the Google Speech-to-Text API is converted into an array of individual words which are checked against the hardcoded array of words that should ideally be in the transcription. After running a pilot on 10 participants, we found out that if the difference of words between both the arrays is greater than 40% of the total words, one of two things is possible: (1) The user did not speak all the words while recording or (2) The Speech-to-Text API transcribed a few words incorrectly (a common occurrence with people who do not have English as a first language/people who have an accent). Hence, considering both the cases, users are prompted to record the audio again for both scenarios. Further, to ensure that we took the correct phrase for analysis, we hardcoded the condition for getting atleast one word of both, starting 2 words and ending 2 words of a tonic unit as correct. If the preceding condition is satisfied and the overall difference in transcription is less than 40%, the audio file is sent forward for a check on the pitch modulation of the spoken words.

Intonation analysis phase - The audio file is then sliced according to the timeframes of each word (with the help of Google Cloud Platform) and checked against the algorithm that contains user-specific base and threshold values for each intonation. From the averaged pitch values, it can be seen whether the buckets fall in the range of the base, upper or lower intonation, and subsequently it is decided which intonation it is.

Visual feedback phase - The system displays the final results of the analysed intonations through the visual representation design requirements (**R3**) as found in the visual representation user study.

9 MOBILE APPLICATION TESTING

We developed an iOS mobile application for testing the effectiveness of our learning application. We used a mixed bag approach of conducting a user study to evaluate the effectiveness of the developed system, using both qualitative and quantitative insights.

9.1 Mobile application structure

The main part of our mobile application is broadly classified into 4 sections that come sequentially in the following order - Pre-Training test, Training, Post-Training test, and Analysis. We carefully curated the content for the application so that the meanings mapped to the various intonations follow certain rules as described by Grice [28].

The Pre-Training test section has 2 modules, each module having 3 exercises, with an instruction for a user to speak naturally but with no visual cues for help. There are a total of 6 AI (accent intonations), 3 QI (question intonations), 9 CI (continuation intonations), and 2 SI (sentence intonations) embedded within the content of the 2 modules. The total number of each intonation in each exercise is restricted by the place at which each of these intonations is used within a sentence and the sentence structure and hence their number in the exercises is unequal. The second module within this section has bigger sentence lengths and more intonations which increases the difficulty of the second module as compared to the first one (based on requirement **R5**). Within each module, the content is set up in such a way that the same sentence has to be stressed at different words according to 3 different contexts in 3 exercises (based on **R2**).

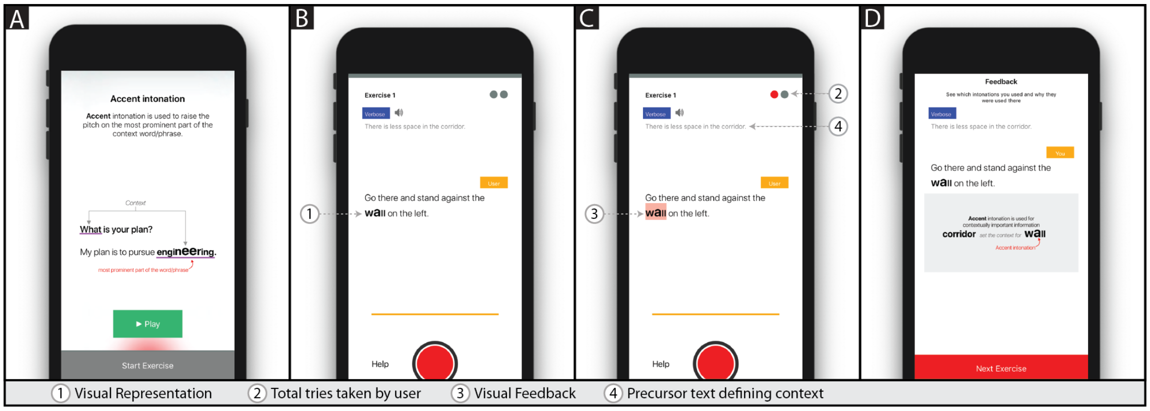

A
Accent intonation
Accent intonation is used to raise the pitch on the most prominent part of the context word/phrase.
Context
What is your plan?
My plan is to pursue engineering.
most prominent part of the word/phrase
▶ Play
Start Exercise
B
Exercise 1
Verbose
There is less space in the corridor.
User
Go there and stand against the wall on the left.
1
Help
C
Exercise 1
Verbose
There is less space in the corridor.
User
Go there and stand against the wall on the left.
2
3
4
Help
D
Feedback
See which intonations you used and why they were used there
Verbose
There is less space in the corridor.
You
Go there and stand against the wall on the left.
Accent intonation is used for contextually important information
corridor set the context for wall
Accent intonation
Next Exercise
1 Visual Representation
2 Total tries taken by user
3 Visual Feedback
4 Precursor text defining context

Fig. 6. Training session. A) Introduction page introducing an intonation with information and audio clip. B) Training page C) Training page with the evaluated tonic units. D) Feedback page explaining why an intonation is used at a specific place and in what context.

The training section first introduces each new intonation on the screen with its visual representation which the user is supposed to necessarily play and listen to (based on **R4**, Figure 6A). Next, we have a training exercise screen that emulates a conversation with the Verbose chat agent, which the user is supposed to participate in by looking at the visual representation and record his voice to get an evaluation on the tonic units (based on **R1, R3**, Figure 6C). Finally, we present a feedback screen that reinforces the intonation information introduced earlier, along with providing explanations of "why" the user was prompted to stress on certain words in different contexts (based on **R4**, 6D). The Training section has 4 modules, each module containing 3 exercises with different context but the same user content, in the same manner as designed in the Pre-Training test. We set up the training exercise in increasing order of number of intonations and readability index [3], which uses number of sentences, length of sentences, number of words, number of complex words, and average syllables per word (based on **R5**). The Training section has a total of 18 accent intonations, 7 sentence intonations, 9 continuation intonations, and 3 question intonations. Again, this unequal number of intonations is because of the placement of these intonations at specific places in sentences in the English language which are restricted by the sentence structure. The number of continuation intonations is highest as they are used for expressing continued thoughts and hence they come multiple times in one stretch. The number of accent intonations (context-related intonations) is more than question and sentence intonations as context-related information is usually present in each sentence, whereas question and sentence intonations may not. Visual representation on the training page indicates "where" and "how" to stress (based on **R2**, Figure 6B). The total number of tries that a user can take is displayed on the top right part of the training page with the total number of circles. This is designed as (n+1), n being the number of intonations within the exercise (based on **R6**). HCI gamification literature has discussed the relevance of inculcating a sense of challenge, difficult enough to merit a player's attention but at the same time not so difficult to leave them failing and disappointed [19]. Therefore, the user was provided with n+1 attempts. This is designed with the understanding that the more intonations in an exercise, the more number of tries a user might need to get the exercise as correct. However at the same time to ensure that it does not get too stressful for the users, we further introduced a bonus section after n+1 incorrect attempts where the user could take a hint and imitate the correct intonation. Another important gamification aspect to enhance a player's cognition and evaluation ability to present the user with instant, easy to understand feedback [32, 54]. To inculcate this, we decided to present each user attempt visually as correct or incorrect through red and green visual circles on the training screen. The red circle displays incorrect attempt, the green circle displays correct attempt and the grey circle shows the number of attempts left. The skip button gives the option to skip the exercise once the user has used at least one bonus try, which gets activated after (n+1) tries. The hint button gets activated after the user exhausts (n+1) number of tries (based on **R6**). It consists of the correct pitch modulation recording of the Verbose chat agent. The evaluated tonic units are then presented on the next screen to show the exact parts of the sentence that the user got right or wrong (Figure 6C). The help button is designed to reiterate the action part of the screen.

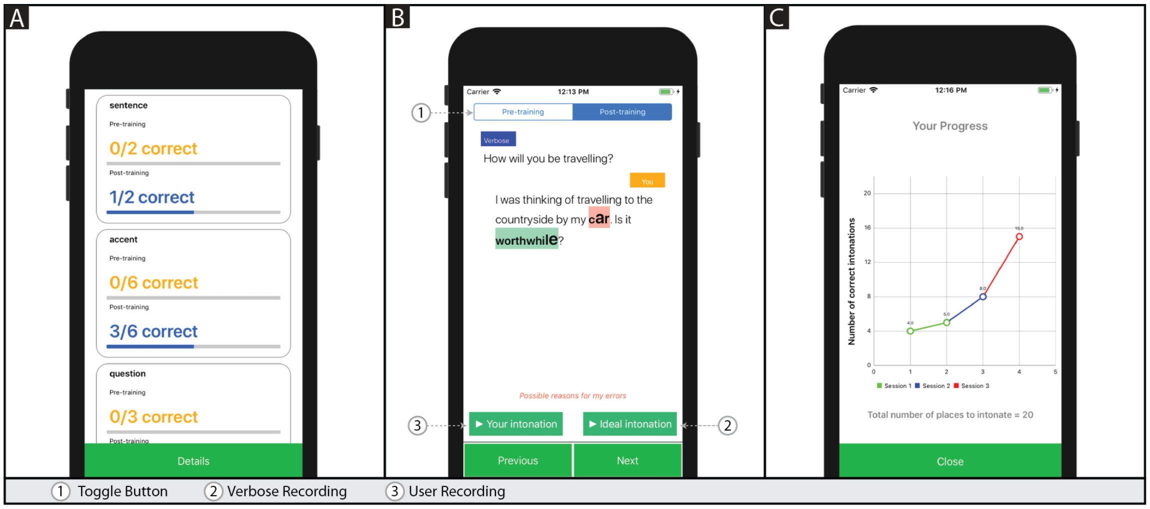
The Post-Training test section has the same content as the Pre-Training section, with no visual representation cues but with the instructions encouraging the participant to speak with skills learned in training.

Finally, an Analysis section (Figure 7) is designed to show the percentage change in user performance across sessions and across different intonations. All Pre and Post Training screens also appear again in the Analysis section with correct visual representations indicating the place and type of stress, the correct and wrong marked units, the audio recordings as performed by the user, and the ideal recordings of the Verbose chat agent (based on **R8**). This section is supposed to make the user aware of his/her overall performance and give details on the exact parts where he/she did mistakes.

9.2 Recruitment

17 students from our university were recruited over responses from email requests sent across the university to participate in the study, out of which 15 completed the study. No participant was provided any incentive to complete the study.

All the participants (18 years of age or above) chosen for the study were pass-outs from standard English teaching schools and were comfortable with English as a second language. These eligibility for recruitment were necessary as intonating skills are developed over and above the normal English speaking skills which include knowing the basic pronunciation of words and speaking with the right grammatical structure. We wanted to calculate the improvement across people not because of their ease with the English language but because of the skills taught through the application.



How will you be travelling?

I was thinking of travelling to the countryside by my car. Is it worthwhile?

Fig. 7. Analysis session. A) Intonation wise performance analysis. B) Audio clip comparisons of user recordings with Verbose recordings in Pre and Post Training sessions. C) Session wise overall performance analysis.

9.3 Procedure

The study consisted of 3 sessions spread over 3 consecutive days (Figure 8). This was done to observe the carry-over effects of the learning methodology and confirm that any changes observed are not because of the novelty effect of introducing a new teaching methodology. In all the 3 sessions, users were given earphones with an embedded mic in it to be used for listening to the Verbose chat agent and recording their voice while interacting with the smartphone application installed on iPhone 6 devices made available by the researchers. The sessions took place in a quiet room with negligible ambient noise. No other people were allowed to be present in the room except for the participant and the researchers.

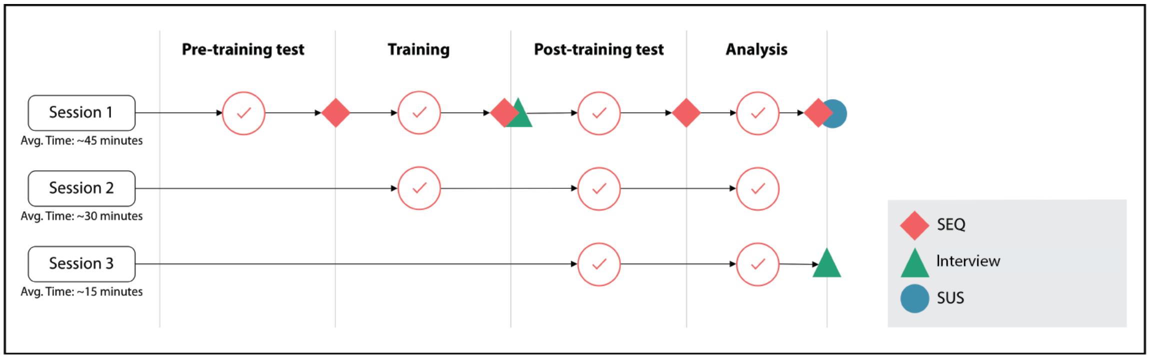


Fig. 8. The flow of mobile application testing

In **the first session** (on day 1), the researchers explained the users with the purpose of the experiment followed by them consenting to participate. They were then requested to interact with the complete application and in between fill the Single Ease Questionnaire (SEQ) [43], after each relevant task (one after Pre-Training test, second after Training, third after Post-Training test, and fourth after Analysis). This was done to get evaluation of the ease of use of participants with the teaching pedagogy and the application flow. We also conducted an open-ended semi-structured interview after the Training phase which was meant to evaluate the effectiveness of the teaching pedagogy taught in the Training phase through qualitative analysis of the user inferences on the mobile application. This interview was done right after the Training phase and not when the session ended as we expected a participant to better remember the training details right after he finishes it than at the end of the session. After finishing the complete study, we also administered the System Usability Scale (SUS) [15] to measure the application usability.

The second session (on day 2) did not have the Pre-Training test. The performance on the Post-Training test phase of day 1 is carried over to become the initial value of the Pre-Training test phase on day 2. The modules in the post-training test phase in each session are kept different from the previous session to avoid the practice bias that could potentially cause a change across the different sessions. However, the difficulty level of the modules is matched using the readability index [3] so that the change in performance is evaluated due to actual transferability of learning skills and not because of a practice bias on similar sentences. The users are made to interact with the entire application in this session in one go without any form filling in between.

The third session (on day 3) only required the users to complete the Post-Training test. Again, the modules in this session are different from the previous session but with a comparable difficulty level. This session is followed by a small semi-structured interview to understand the overall experience of the participant, their learning from the app, and other feedbacks.

9.4 Analysis and Results

The first session lasted about 45 minutes, the second session for about 30 minutes, and the third session lasted with an average duration of 15 minutes. Over the three sessions of this user study, we measured the intonation performance scores, usability, and ease of use of the complete system followed by a qualitative analysis of the user inferences on the mobile application. We inductively coded the themes and then, deductively coded all the interviews again, marking relevant converging themes. Following are the themes that we came up with after analysing the data from the user study.

Novelty of the system

All but 2 participants pointed out the novelty of the speaking style introduced by the system in the English speaking skills of ESL speakers. The mobile application gave them "awareness on using intonations" to express intentions in English. P02 said, "*I was unaware and didn't think specifically about stress (intonations) till now. I always thought more on clarity or time spent on a topic, rather than the stress in the pitch.* Specifically, continuation intonation (CI) was something new and useful for 11 users. P01 said,"*...I guess the continuation intonation (in the application) told (informed) me to emphasize each and every word while I was finishing a thought. I would really like to implement that.*"

Effectiveness of teaching pedagogy

We use the performance scores of users from session 1 pre-test training and session 3 post-test training to find the effectiveness of teaching pedagogy in the Verbose mobile application. The score of a user is measured by assessing the marked units in the Pre and Post Training and only if the user got both the type and place of intonation correct simultaneously, the tonic unit is marked as correct.

To find if the scores were significant, we performed a two-way repeated measures ANOVA test for the two factors of session (levels - 1 and 3) and intonation (levels - accent , sentence, question and continuation). The data followed normal distributions as tested through a Shapiro-Wilk and Kolmogorov-Smirnov test. However, our data violated sphericity for the factors - intonation and (intonation * session), as observed through the Mauchly's test of sphericity. Hence, we used Greenhouse Geisser corrections for reporting the final test values for these 2 factors.

Session -> F (1, 15) = 13.182, p = 0.003, pes = 0.485

Intonation -> F (1.329, 18.606) = 15.750, p < 0.001, pes = 0.529, (E = 0.443)

*Intonation * session* -> F (1.418, 19.852) = 3.969, p = 0.048, pes = 0.221, (E = 0.473)

We observe that the change in performance of user intonations is very highly significant. This could be because the users were unaware of how to intonate in the beginning of the study and hence didn't intonate much. We also observe that the overall change in user performance across sessions, and the change in performance with respect to each intonation was significant. Since we tested users in multiple sessions we can say that the overall change across session session 1 to 3 was not because of the novelty effect of teaching new skills to ESL speakers. Further, since the content taught in the Training exercises was different from the content in the Pre and Post Training Test phases within and across different sessions, we can say that there was no practice bias that could have affected the results. This affirms that the change in performance is because of the effectiveness of Verbose application in teaching intonations.

Teaching pedagogy qualitative validation

The "where and how" of intonations - All participants explicitly pointed out the role of learning pedagogy in the mobile application in making them understand intonations and their implementation in conversations. 12 users mentioned that they understood "where and how" they raise their tone on certain parts of a sentence after the app usage (**R2**). For example, P01 said, "*...I guess it (the mobile application) provides hints towards the modulation of how you speak. If someone is talking really fast, or if someone doesn't know what he or she is speaking, just by the modulation you can tell if that sentence is a question or a statement.* P02 said, "*...(the mobile application helps) in a way that you know the place where you stress and the other person knows where the real information is.*".

Conversation-based teaching pedagogy - The average time taken by the participants to complete an entire session in the mobile application from start to finish was about 45 minutes. Still, all participants found the experience of learning intonations as a pleasant one. This is mainly attributed by almost all participants to the nature of the conversation-based teaching pedagogy style (**R1**). P05 said, " *I found the experience to be a very pleasant one. I never felt at any moment that I was getting bored. In fact, I rather enjoyed conversing with the chat agent, mainly because the way he spoke sounded very nice. I could almost feel like a personalized trainer is taking me through the app and conversing with me, alongside teaching the necessary skills.*"

Gamification - 6 participants directly or indirectly also mentioned the role of gamification in the training exercises in making the experience a pleasant one (**R6**). P10 said, "*The design of the training exercises was very effective in my opinion. The screens were very interactive, it almost became like a challenge not to enter in the Bonus section and get the intonation right in the given number of tries*". Practicing the exercises until a minimum (n+1) number of times till a participant got a particular intonation as correct was a bit hectic for 3 participants. However, 5 participants explicitly mentioned that the Bonus section turned out to be very important in making them practice a minimum threshold number (based on the exercise difficulty) of times, before they had the option to skip the exercise.

Visual representation - Most users were able to finish successive training exercises in a lesser number of tries. This was attributed by 10 users to the visual representation (**R3**). P04 said, "*...(the context) was easy to understand because of the changed characters (in the visual representation).*

Conceptual facts on intonations - Participants also mentioned the importance of introduction and feedback pages for introducing new conceptual knowledge, helping in disambiguation between different intonations, understanding the 'why' of 'where, how and why' in intonations and finally in retention of the taught intonations (**R4**). P06 said, "*The introductory screen was very important as it introduced me with new conceptual knowledge, which made things relatively easy to understand when I had to perform in the exercises. Also, the feedback screen gave explanations that why depending on a particular question (context) asked by Verbose, the intonation of an answer for specific words will change.*". P12 stated, "*I also think that presenting a feedback of each exercise was also very helpful in concretizing and remembering these concepts*"

Context - In particular, 9 users better understood the importance of context through these introductory and feedback screens just after finishing an exercise (**R2**). P03 stated, "*It (feedback screen) is important because first when I read it, I wasn't connecting it to the question (context), but after I saw the feedback I saw (understood) why that thing (a word) was intonated according to the context*".

Gradual increase in difficulty - 13 participants said that introducing the 4 intonations and the aspect of 'how, why, and where' to speak in a conversation didn't seem to be at all overwhelming for any participant. This is attributed to the gradual introduction of intonations and the relative increase in readability index [3] of sentences in terms of number of sentences, length of sentences, number of words, number of complex words, average syllables per word and number of intonations in the app content (**R5**). P12 said, "*When the session finishes, it seems like you have learned a lot. It is a lot of new things, but I think the content has been designed intelligently. I didn't feel anytime that I have to learn a lot suddenly. It was a gradual process.*

Analysis section - The personalized analysis section and the display of the session performance into individual intonations and sentences, alongside presenting an overview of the entire session, helped 9 participants in self-analysis and increasing awareness of their current level of communicative intonations (**R7, R8**). P08 said, "*Comparing my performance with the way I performed before training helped me a lot, I could instantly get what was wrong with my communication process. It made me aware of how I used to speak earlier.*". P09 said, "*The analysis section was very neatly presented evaluating me for each sentence that I performed on, and then displaying my overall session performance. It was a very holistic way of presenting results.*".

Utility of the mobile application

Application helpful in conveying user intentions through intonations - 13 participants found the mobile application to be useful in learning how to express oneself better by conveying intentions through intonations and communicating effectively. P04 gave a very relatable example, "*You can't understand the formation of sentences in their (ESL speakers) speech, so you are in general confused (about context) what they are trying to stress on. At those moments you understand that it (context specific stress words) matters...like some people intonate words like "the" and "it" , and it just makes the sentence confusing to interpret.*" P04 summed it up very nicely, "*It's important because you want to convey the intention very clearly...my speech has improved. I am not a bland speaker anymore (laughs).*"

Relatable use cases of application - Upon prompting, almost all participants came up with some examples of people who intonate and are effective speakers. P06 gave the example of an acquaintance who uses intonations and speaks at toastmasters, "*I have a senior at college, she uses intonation or stress very well. She also went to the toastmasters.*" P07 stressed how such a system can be helpful for people preparing for job interviews. "For UPSC (Civil Services examination) aspirants, it is necessary to clear the interview...So these people who are preparing, they use many methods (to improve English) and this definitely could be one of them".

System usability and ease of use of the mobile application

We used the SUS questionnaire after session 1 to assess the usability of our system. Ten SUS statements were presented in sequence to all participants. For this, we created a simple online questionnaire where participants were asked to rate each statement using a five-item Likert scale ranging from strongly disagree to strongly agree. The participants rated the overall user-friendliness of our mobile application to be 82.3 (out of 100) on average, which is above the average benchmark SUS score of 68 [4]. Any score above 80.3 on the SUS scale is among the top 10 percentile and correlates to the adjective rating of good [13].

Users were also requested to rate the level of difficulty of the four main tasks in the mobile application using the Single Ease Questionnaire (SEQ) during session 1: one immediately after Pre-Training test, second after completing the Training, third after the Post-Training test and fourth after going through the Analysis section. The mean SEQ rating for the four tasks came out to be 6.06, 4.31, 5.12, and 6.25 (out of 7).

Since users were made to learn about new intonations, and why, how, and where to apply them in the training section, it was expected for this section to have a higher cognitive load for the users as compared to the other sections. A 4.31 rating (somewhere between average and somewhat easy), therefore seems justifiable. The Pre-Training and the Analysis section were rated between easy and very easy. The Post-Training section was rated by the users to be between somewhat easy and easy. The Post-training section was rated as more difficult than the Pre-Training section even though both the sections required the same content to be spoken by the user. From the interviews after day 1, we found out the main reason behind this as cited by most users was that since the post-training required them to apply what they had already learned during training, i.e., which intonation to use, where to use it based on context and continually practice in the Training phase to incorporate it into their usual speaking style, they found this section to be slightly easier as compared to Pre-Training phase.

10 DISCUSSION

Verbose is designed to provide ESL speakers with a conversation based teaching pedagogy of teaching intonations. These intonations can be very meaningful when used with digital personal assistants like Alexa, Siri, Cortana, and Google Assistant to increase the understanding of the user intentions while giving commands. Currently, these chat agents have limited capabilities in understanding the context. If users provide intonation-related cues while conversing with them, it can help the chat agents identify when a user finishes a sentence, when he stresses on important words, when there is a pause coupled with a continuation statement, as compared to a normal pause, and when is a question being asked. Our incisive interviews with the primary stakeholders provide support for the usability, effectiveness, and ease of use of Verbose in enabling ESL speakers to express their intentions clearly through globally observed intonations. However, we elaborate on some key issues in the design of the system that could potentially act as limitations.

10.1 Limitations

Firstly, The system only calculates the intonations where they are supposed to be present but doesn't detect where no intonation is supposed to be present. This could lead to potential misuse of the system by participants for getting better results by intonating at all possible places. However, the main purpose of the work was to present a design of a system that makes ESL speakers aware of how, why and where to use intonations, and not otherwise. Secondly, our system is susceptible to user tiredness and people can have modified thresholds while using the system. We assumed that the voice thresholds of people remain constant within a training session. Further, the three sessions are timed successively at 45, 30 and 15 minutes respectively. It could be possible that the increase in performance results of participants could have some effect due to the participants getting less tired successively in each session. Thirdly, although we evaluated the system over multiple sessions, change in performance scores of intonations might be a useful measure of finding its effectiveness over a longer period. Hence, we couldn't only rely on change in scores by itself and augmented our results with qualitative observations from the user study. Lastly, we only chose to map the function of pragmatic intentions and not emotional intentions for providing feedback in our system. This is done as the accuracy of detecting emotional intentions is around 60 %[45], whereas the accuracy of pragmatic intentions is 100%, as pragmatic linguistics tries to find out the meaning of an expression by removing the ambiguity in its meaning ([33], p. 292).

11 CONCLUSION

We designed a novel mobile-based application iteratively with the help of various stakeholders and design experts for teaching voice modulations to ESL speakers. The final evaluation of our system revealed that it was usable in its intended purpose i.e. it helped ESL speakers to learn pitch modulation in English. People got aware of a new skill in speech delivery that helps them express their intentions clearly and agreed to the utility of such a system by looking at various places where the taught pitch modulation skills are helpful, for example, in job interviews, teaching etc. Further, our system proved to be effective in terms of retaining the taught skills over multiple usage of the artifact.

ACKNOWLEDGMENTS

We would like to thank Google Cloud Platform for awarding us research credits to support this work.

REFERENCES

[1] 2019. *British Council Spoken english.* Retrieved May 15, 2019 from https://www.britishcouncil.in/english/courses-adults/spoken-english

[2] 2019. *British English Pronunciation.* https://englishpronunciationroadmap.com

[3] 2019. *Readability Test Tool.* Retrieved August 14, 2019 from https://www.webfx.com/tools/read-able/

[4] 2019. *System Usability Scale.* Retrieved May 15, 2019 from https://www.userfocus.co.uk/articles/measuring-usability-with-the-SUS.html

[5] 2020. *How Many People In The World Speak English?* Retrieved February 15, 2019 from http://www.stgeorges.co.uk/blog/learn-english/how-many-people-in-the-world-speak-english

[6] 2021. *Diglossia.* Retrieved February 03, 2021 from https://en.wikipedia.org/wiki/Diglossia

[7] 2021. *Global migration, by the numbers: who migrates, where they go and why.* Retrieved February 03, 2021 from https://www.weforum.org/agenda/2020/01/iom-global-migration-report-international-migrants-2020/

[8] David Abercrombie. 1968. PARALANGUAGE. *International Journal of Language & Communication Disorders* 3, 1 (1968), 55–59. https://doi.org/10.3109/13682826809011441 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.3109/13682826809011441

[9] Mohammad Rafayet Ali, Kimberly Van Orden, Kimberly Parkhurst, Shuyang Liu, Viet-Duy Nguyen, Paul Duberstein, and M. Ehsan Hoque. 2018. Aging and Engaging: A Social Conversational Skills Training Program for Older Adults. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18*. ACM Press, Tokyo, Japan, 55–66. https://doi.org/10.1145/3172944.3172958

[10] Mohammad Rafayet Ali, Kimberly Van Orden, Kimberly Parkhurst, Shuyang Liu, Viet-Duy Nguyen, Paul Duberstein, and M Ehsan Hoque. 2018. Aging and engaging: A social conversational skills training program for older adults. In *23rd International Conference on Intelligent User Interfaces*. ACM, 55–66.

[11] International Phonetic Association et al. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge University Press.

[12] Peter L Auer, Peter Auer, Elizabeth Couper-Kuhlen, Frank Müller, et al. 1999. *Language in time: The rhythm and tempo of spoken interaction.* Oxford University Press on Demand.

[13] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[14] Dwight Bolinger and Dwight Le Merton Bolinger. 1986. *Intonation and its parts: Melody in spoken English.* Stanford University Press.

[15] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[16] Mark Bubel, Ruiwen Jiang, Christine H Lee, Wen Shi, and Audrey Tse. 2016. AwareMe: addressing fear of public speech through awareness. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems.* ACM, 68–73.

[17] Mark Bubel, Ruiwen Jiang, Christine H. Lee, Wen Shi, and Audrey Tse. 2016. AwareMe: Addressing Fear of Public Speech through Awareness. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16.* ACM Press, Santa Clara, California, USA, 68–73. https://doi.org/10.1145/2851581.2890633

[18] Hao-Jan H Chen. 2001. Evaluating five speech recognition programs for ESL learners. In *ITMELT 2001 Conference, Hong Kong. http://elc. polyu. edu. hk/conference/papers2001/chen. htm.*

[19] Dongseong Choi, Hoyoung Kim, and Jinwoo Kim. 1999. Toward the construction of fun computer games: Differences in the views of developers and players. *Personal Technologies* 3, 3 (1999), 92–104.

[20] Alan Cruttenden. 1997. *Intonation* (2 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139166973

[21] David Crystal. 1974. *Paralinguistics.* Retrieved May 14, 2019 from http://www.davidcrystal.com/?fileid=-4166

[22] David Crystal. 1975. *The English tone of voice: essays in intonation, prosody and paralanguage.* Hodder Arnold.

[23] David Crystal. 2003. *The Cambridge Encyclopedia of the English Language.* Cambridge University Press.

[24] F Cummings and R Port. 1998. Rhythmic constraints on stress timing in English. *Journal of Phonetics* 26 (1998), 145–171.

[25] Ionut Damian, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems.* ACM, 565–574.

[26] Tanusree Das, Latika Singh, and Nandini C Singh. 2007. Rhythmic structure of Hindi and English: new insights from a computational analysis. *Progress in brain research* 168 (2007), 207–272.

[27] Rebecca M Dauer. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of phonetics* (1983).

[28] H Paul Grice, Peter Cole, Jerry L Morgan, et al. 1975. Logic and conversation. *1975* (1975), 41–58.

[29] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing.* ACM, 697–706.

[30] Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics* 71 (2018), 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

[31] Mohammad Owais Khan. 2011. Rhythm and Intonation Patterns in English and Urdu-A Contrastive Analysis. *Language in India* 11, 5 (2011).

[32] VKMK Kondraju. 2003. An exploration of HCI design features and usability techniques in gaming. *IOSR J. Comput. Eng* 15, 3 (2003), 53–57.

[33] Jody Kreiman and Diana Sidtis. 2011. *Foundations of voice studies: An interdisciplinary approach to voice production and perception.* John Wiley & Sons.

[34] Kazutaka Kurihara, Masataka Goto, Jun Ogata, Yosuke Matsusaka, and Takeo Igarashi. 2007. Presentation sensei: a presentation training system using speech and image processing. In *Proceedings of the 9th international conference on Multimodal interfaces.* ACM, 358–365.

[35] Peter Ladefoged and Keith Johnson. 2006. A Course in Phonetics (5th). *Thomson Wadsworth* (2006).

[36] Philip Lieberman. 1967. Intonation, perception, and language. *MIT Research Monograph* (1967).

[37] Yi-Jing Lin and Chialin Chang. 2017. MyET and English Pedagogy. (2017).

[38] Dania Murad, Riwu Wang, Douglas Turnbull, and Ye Wang. 2018. SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In *2018 ACM Multimedia Conference on Multimedia Conference.* ACM, 1679–1687.

[39] Mikhail Ordin and Leona Polyanskaya. 2014. Development of timing patterns in first and second languages. *System* 42 (2014), 244–257.

[40] Peter Roach. 1982. On the distinction between 'stress-timed'and 'syllable-timed'languages. *Linguistic controversies* 73 (1982), 79.

[41] Sean Robertson, Cosmin Munteanu, and Gerald Penn. 2018. Designing Pronunciation Learning Tools: The Case for Interactivity against Over-Engineering. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* ACM Press, Montreal QC, Canada, 1–13. https://doi.org/10.1145/3173574.3173930

[42] James Rush. 1833. The philosophy of the human voice. (1833).

[43] J Sauro. 2012. 10 things to know about the Single Ease Question (SEQ). *Measuring U, 2012* (2012).

[44] SR Savithri, M Jayaram, D Kedarnath, and S Goswami. 2007. Speech rhythm in Indo Aryan and Dravidian languages. In *Proceedings of the International Symposium on Frontiers of Research on speech and music.* 170–174.

[45] Klaus R Scherer. 1986. Vocal affect expression: A review and a model for future research. *Psychological bulletin* 99, 2 (1986), 143.

[46] Jan Schneider, Dirk Börner, Peter van Rosmalen, and Marcus Specht. 2015. Presentation Trainer, Your Public Speaking Multimodal Coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15).* Association for Computing Machinery, New York, NY,

USA, 539–546. https://doi.org/10.1145/2818346.2830603

[47] Bjorn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing* (1st ed.). Wiley Publishing.

[48] Hema Sirsa and Melissa A Redford. 2013. The effects of native language on Indian English sounds and timing patterns. *Journal of phonetics* 41, 6 (2013), 393–406.

[49] Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2015. Automated social skills trainer. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 17–27.

[50] M Iftekhar Tanveer, Emy Lin, and Mohammed Ehsan Hoque. 2015. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 286–295.

[51] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. 2017. RoboCOP: A Robotic Coach for Oral Presentations. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 27.

[52] Roumen Vesselinov and John Grego. 2012. Duolingo effectiveness study. *City University of New York, USA* 28, 1-25 (2012).

[53] Xingbo Wang, Haipeng Zeng, Yong Wang, Aoyu Wu, Zhida Sun, Xiaojuan Ma, and Huamin Qu. 2020. VoiceCoach: Interactive Evidence-based Training for Voice Modulation Skills in Public Speaking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[54] Hou Wenjun and Bai Xiudong. 2006. HCI in real-time strategy games: a study of principals and guidelines for designing 3D user interface. In *2006 7th International Conference on Computer-Aided Industrial Design and Conceptual Design*. IEEE, 1–6.

[55] Ru Zhao, Vivian Li, Hugo Barbosa, Gourab Ghoshal, and Mohammed Ehsan Hoque. 2017. Semi-Automated 8 Collaborative Online Training Module for Improving Communication Skills. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 32.

A LOW-FIDELITY SCREENS

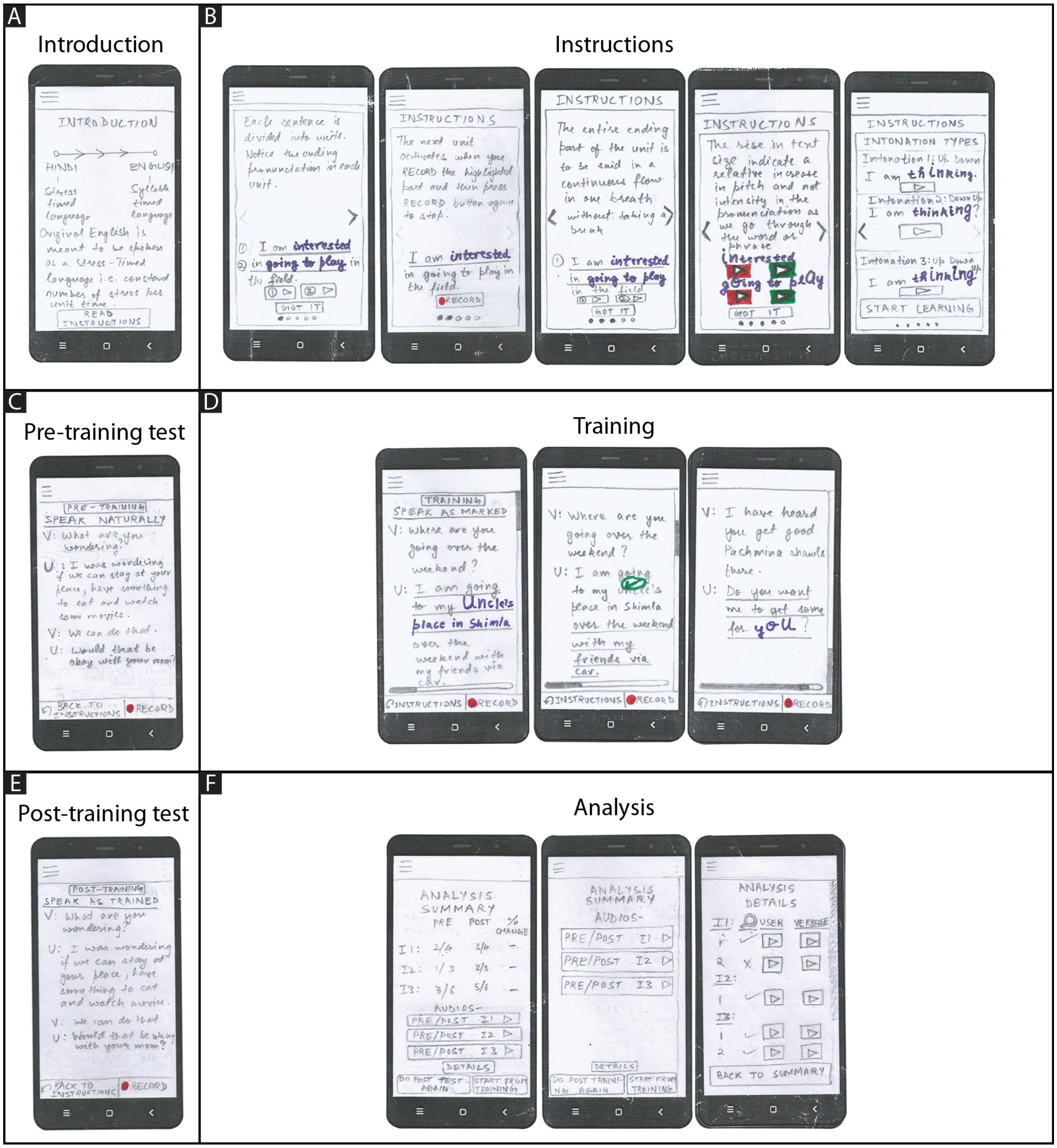
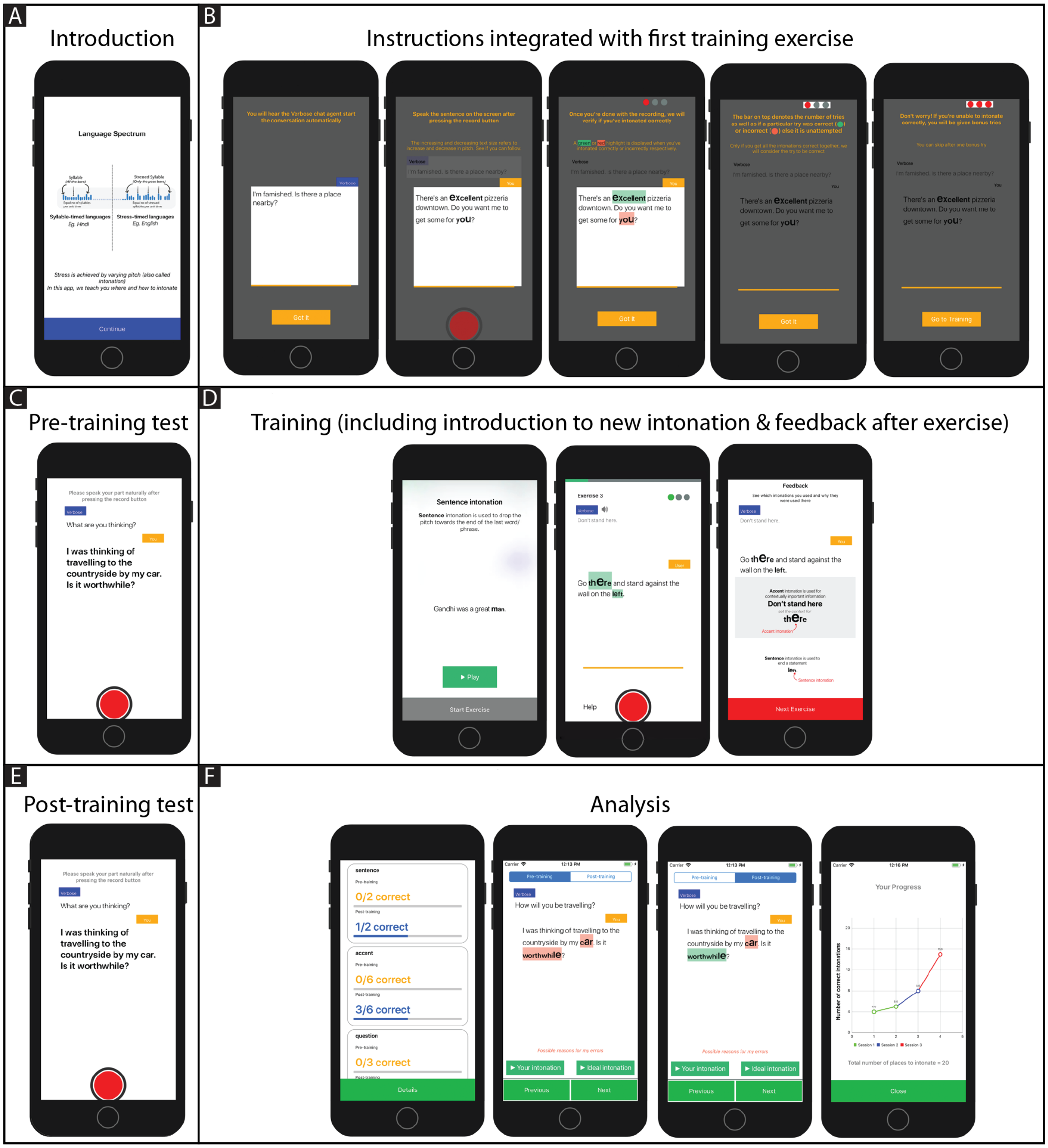


Fig. 9. Low-fidelity prototype sample screens

B SMARTPHONE APP SCREENS


A
Introduction
B
Instructions integrated with first training exercise
C
Pre-training test
D
Training (including introduction to new intonation & feedback after exercise)
E
Post-training test
F
Analysis

Language Spectrum

Syllable
(All the bars)
Equal no of syllables per unit time
Syllable-timed languages
Eg. Hindi
Stressed Syllable
(Only the peak bars)
Equal no of stressed syllables per unit time
Stress-timed languages
Eg. English

Stress is achieved by varying pitch (also called intonation)
In this app, we teach you where and how to intonate

You will hear the Verbose chat agent start the conversation automatically

I'm famished. Is there a place nearby?

Speak the sentence on the screen after pressing the record button

The increasing and decreasing text size refers to increase and decrease in pitch. See if you can follow.

I'm famished. Is there a place nearby?

There's an **excellent** pizzeria downtown. Do you want me to get some for **you**?

Once you're done with the recording, we will verify if you've intonated correctly

A green or red highlight is displayed when you've intonated correctly or incorrectly respectively.

I'm famished. Is there a place nearby?

There's an **excellent** pizzeria downtown. Do you want me to get some for **you**?

The bar on top denotes the number of tries as well as if a particular try was correct (●) or incorrect (●) else it is unattempted

Only if you get all the intonations correct together, we will consider the try to be correct

I'm famished. Is there a place nearby?

There's an **excellent** pizzeria downtown. Do you want me to get some for **you**?

Don't worry! If you're unable to intonate correctly, you will be given bonus tries

You can skip after one bonus try

I'm famished. Is there a place nearby?

There's an **excellent** pizzeria downtown. Do you want me to get some for **you**?

Please speak your part naturally after pressing the record button

What are you thinking?

I was thinking of travelling to the countryside by my car. Is it worthwhile?

Sentence intonation

Sentence intonation is used to drop the pitch towards the end of the last word/ phrase.

Gandhi was a great **man**.

Exercise 3

Don't stand here.

Go **there** and stand against the wall on the **left**.

Feedback

See which intonations you used and why they were used there

Don't stand here.

Go **there** and stand against the wall on the **left**.

Accent intonation is used for contextually important information

Don't stand here

set the context for

there

Sentence intonation is used to end a statement

left.

Please speak your part naturally after pressing the record button

What are you thinking?

I was thinking of travelling to the countryside by my car. Is it worthwhile?

sentence
Pre-training
0/2 correct
Post-training
1/2 correct

accent
Pre-training
0/6 correct
Post-training
3/6 correct

question
Pre-training
0/3 correct
Post-training

How will you be travelling?

I was thinking of travelling to the countryside by my **car**. Is it **worthwhile**?

Possible reasons for my errors

How will you be travelling?

I was thinking of travelling to the countryside by my **car**. Is it **worthwhile**?

Possible reasons for my errors

Total number of places to intonate = 20

Fig. 10. Smartphone app sample screens