Improving Accuracy in Imitating and Reading Aloud via Speech Visualization Technology

https://doi.org/10.3991/ijet.v15i08.11475

Xiaobin Liu (✉), Diying Wu, Yiwen Ye, Manfei Xu, Jianli Jiao
South China Normal University, Guangzhou, China
liuxiaobin@m.scnu.edu.cn

Wenheng Lin
Nansha Dongchong Middle School, Guangzhou, China

Abstract—This article reports on a teaching experiment that uses speech visualization technology to facilitate EFL (English as a foreign language) learners' ability of imitating and reading aloud (IRA). Traditional practice of "listen and repeat" used in pronunciation and intonation training often only hazily illustrates some salient speech features, such as pitch and intensity, etc., making it difficult for learners to accurately differentiate the features of pronunciation and capture the subtlety of intonation of the target language. As speech visualization transforms aural information into visual graphics and illustrates it in a clearer way, it can highlight the most important phonetic features of utterances. Such technology was used in a three-month teaching experiment carried out in a senior middle school in southern China for the purpose of reducing learners' phonetic errors and enabling them to perceive sense groups appropriately. The results indicated that this approach was effective in increasing accuracy in pronunciation while did not have a significant effect on the fluency and rhythm in speech.

Keywords—Speech Visualization Technology, English pronunciation, speech

1 Introduction

There has been a long-reached consensus that in language learning, reading aloud (RA) is closely related to listening, reading, and writing in EFL teaching, and that it has always been regarded as a basic and effective method for EFL learners to improve their speech accuracy. Madsen [1] points out that RA can test students' pronunciation, intonation, mastery of English phonetic rules, fluency, and grammatical competence. Halliday [2] mentions that oral reading helps learners deal with difficult texts with high information density. Remarkably, in the context of silent reading, "even skilled readers have a little voice running through their heads the whole time" [3]. Warner, Crolla, Goodwyn, Hyder and Richards [4] insist that "reading aloud is apparently an indispensable part of teaching". It can be concluded that what learners can learn from RA includes not only some certain language items, but also the whole system of the

target language [5]. Based on this idea, it is of high importance to strengthen learners' oral reading training and improve their abilities in pronunciation, intonation and other relevant aspects during daily teaching.

EFL teaching has emphasized that learners should be exposed to an abundance of English recordings of native speakers and receive a wide variety of corresponding imitating and reading aloud training. Imitating and reading aloud (IRA) which is a further version of RA focuses on learners' perception of the speech input and their speech imitation experience. They are supposed to manage to produce speeches as exactly as the original audio texts [6]. IRA remains one of the main approaches to train Chinese learners' English pronunciation. Normally, given a text and the related audio or video material, learners are required to imitate the speech, rehearse it several times and finally read it out. Through imitating and uttering aloud some sections or paragraphs of a text, or some short passages, learners can be encouraged to pay more attention to their pronunciation and intonation accuracy, and the process of their language acquisition (namely, chunking) can be also promoted [7]. In this way, learners' language becomes increasingly accurate and close to the original sound version. Eventually, learners are able to use the target language flexibly [8]. Therefore, the method of IRA is highly recommended to language teachers in their teaching of speech.

However, the traditional practice of "listen and repeat", which is still commonly found in language classes currently, tends to allow learners to aurally sense the pitch, intensity, and other salient speech features hazily, making it difficult for them to accurately perceive the varied pronunciation and intonation features of their target language. It fails to provide learners with more detailed and accurate guidance. Besides, the traditional method also indicates that students are in need of high-quality English teachers who often serve as language role models for students to follow. Such teachers are not only limited in number but also would encounter great difficulty in conducting speech training to large size classes having at least 40 students. To solve these problems, speech visualization technology supported by multimodal discourse theory may provide an alternative way.

2 Multimodality and Speech Visualization

Inspired by Maybury's [9] and Hill's [10] definition of modality in human-computer interaction in the field of computer science, researchers use modality to refer to the way in which humans interact with the external environment (such as humans, machines, animals, etc.) utilizing their senses (such as seeing, hearing, touching, etc.) [11]. Various modalities can be found in newspapers, magazines, advertisements, posters, storybooks, textbooks, encyclopedias, instructions, computer interfaces and even in people's daily communication with each other [12]. For example, while people are talking, they are vividly using body language, including facial expressions and gestures, indicating that people actually are applying some modalities in daily communication. [13].

Since the emergence of multimodality, its concept has been studied and further investigated in a great number of disciplines such as social semiotics, conversation or discourse analysis, interface design, human-computer interaction, visual design, mathematics, hypermedia, media studies, new literacy studies and so on [14, 15]. Some of the important issues in multimodal discourse research include as follows: Is there a "stronger" or a "weaker" modality in a multimodal discourse? Do they complement each other? Can one of the modalities strengthen or weaken the information provided by the other modalities? [12] Some scholars believe that the more modalities presented, the more information and learning experience people can obtain [16]. Therefore, multimodal discourse theory advocates that teachers can properly use images and sounds integrated with texts and other interactive ways to stimulate learners' multiple senses in language learning.

Speech visualization is one of the technical realizations of multimodal discourse theory. Technically illuminated by the traditional oscilloscope, speech visualization technology presents the acoustic information in the form of electronic graphics, images, text, data, charts, etc. Simply speaking, it converts auditory information into visual symbols. From the perspective of multimodality, this technology adds visual modality to the interaction originally limited to auditory modality in the process of human-computer interaction. In recent years, tools and systems of speech visualization have sprung up with the development of information technology. Overall, speech visualization can be divided into the following categories: (1) speech edition; (2) speech analysis; (3) speech recognition; and (4) visual speech for instance, audio editing tools, such as Cool Edit as well as specialized phonetic software such as Praat, are widely used in various relevant industries.

There are at least three advantages of speech visualization in language teaching. To begin with, employed in the design of teaching activities, speech visualization achieves dynamic presentation of visual interfaces (such as speech waveform, fundamental frequency, energy, etc.) and relevant speech data analysis, thereby closely integrating the planning and final implementation of teaching activities to improve the efficiency of teaching and learning [17]. Secondly, speech visualization allows teachers to use visual symbols to express abstract aural information, which enables learners to understand the phonetic features of the target language in a more visual and accurate way and provides effective guarantee for oral training. Thirdly, as most EFL classes in China are of large size, the visual form of learning can help learners practice and correct their pronunciation independently without teachers at their sides, relieving teachers' pressure of correcting every inaccurate response.

Up to now, a great number of scholars have studied the effect of visual tools on speech teaching. Among them, Hincks [18] mentions that teachers have been using signal analysis software to visualize speech for decades to support teaching and help learners perceive and produce target sounds of a second language. Levis and Pickering [19] and Chapelle [20] believe that the application of visualization technology (such as the use of WASP and Praat) is an important progress in intonation teaching. Godwin-Jones [21] points out that speech analysis and language learning software (for example, Visio-Pitch of KayPentax) are helpful for the accuracy of pronunciation in language learning. Gorjian, Hayati and Pourkhoni [22]

and Boersma and Weenik [23] explore the effectiveness of Praat in phonetics learning and helping learners master English prosodic features. By employing Praat, Brett [24] provides feedback on vowel and diphthong learning while Wilson [25] and Dixon [26] study the teaching of segmental sounds and suprasegmental features. Wallace and Lima [27] present the role of visual techniques such as the role of Audicity and Praat in pronunciation teaching. Martin [28] introduces WinPitch LTL II whose functions mainly include an automated visual comparison between learners' imitation and the model recordings, and the automated feedback of correcting learners' pronunciation on segmental and suprasegmental levels. Tamandani [29] uses Praat to study the stress and intonation teaching for undergraduate students of AMU. Chun [30] describes the role of Speech Analyser in speech teaching. Fox [31] investigates Speech Analyser and Praat in his related research of online phonetics course teaching. Levin [32] studies the practicability of Cool Edit in speech teaching. Derman, Bardakçı and Öztürk [33] use Praat and Cool Edit to assist the investigation of Arabic learners' reading speech in terms of stress and pause when they learn Turkish as a second language. Sariman and Çetin [34] use Cool Edit when studying the influence of computer-aided instruction on the skill of word stressing in teaching Turkish as a foreign language. Zoghbor [35] conducts research on how to teach multi-dialect L2 learners to produce oral English using Cool Edit.

In China, plenty of researchers have made use of various speech visualization tools to carry out speech teaching research. Sun [36] and Zhuang and Bu [17] use Praat to investigate college English majors' accuracy problems in pronunciation. Zhou and Zhang [37] undertake research on oral English fluency using Cool Edit, while Xie [38] focuses on increasing learners' ability of sound recognition and sensitivity to the differences between English and Chinese sounds by making use of Speech Analyser. However, the application and research of speech visualization tools for English IRA training have not been much explored yet.

3 Two Common Problems of IRA

According to David Nunan [39], accuracy and fluency are two important dimensions of oral ability. These two dimensions have then become two major reference designators for evaluating learners' speaking ability and quality. In order to reveal the current situation of English speech teaching, some researchers in China made use of the RA recording samples of 125 college level English learners from China for detailed analysis. By looking statistically at their various errors in RA, researchers found there were two main types of errors, namely, phonetic errors (43 per cent) and errors affecting oral fluency (31 per cent) [40]. From the perspective of error analysis, this study proves that accuracy and fluency are two major indicators and the most common problems with regard to EFL learners' spoken English ability.

The subjects in the research above are learners at an advanced stage of EFL learning. Do Chinese K-12 EFL learners also encounter similar problems? Since 2011, Guangdong Province in China has added Computer-based English Listening and Speaking Test (CELST) to the College Entrance Examination. CELST consists of

three tasks: Reading Aloud, Role Play, and Retelling. Reading Aloud task mainly requires candidates to firstly watch a one-minute video clip and imitate the pronunciation and intonation along its playing. Candidates can have a short rehearsal reading the subtitles of the video clip for about one minute. Then they listen to the speaker once again and have to read along with the subtitles provided. The testees should finally read the subtitles. Candidates are required to keep their pronunciation, intonation and speech speed as consistent as possible with what they hear from the clip. Such a task is referred as IRA in this paper. To inquire into high school graduates' English pronunciation and intonation, a random sampling of IRA recordings from 200 testees participating in the 2011 CELST was conducted and the types of their errors were described and classified. 1,682 errors in total were found by the authors and the specific distribution of errors is shown as follows (Table 1), while the definitions [40, 41] and examples of these errors are shown in Tables 2, 3 and 4.

Table 1. Error statistics of IRA task in CELST

Error types		Amount	Ratio (%)
Phonetic errors	Mispronunciation	442	26.3
	Addition	54	3.2
	Reduction	62	3.7
	Stress	136	8.1
	Subtotal	694	41.3
Fluency errors	Error on sense group	392	23.3
	Repetition	256	15.2
	Invalid mending	61	3.6
	Filling	20	1.2
	Towing	20	1.2
	Subtotal	749	44.5
Word deletion		113	6.7
Word addition		31	1.8
Intonation errors		95	5.6
Grand Total		1,682	100

Table 2. Definitions and examples of phonetic errors in IRA task in 2011 CELST

Error types		Examples	
		Original version	Incorrect sample(s)
Phonetic errors	Mispronunciation: Incorrect or inaccurate pronunciation, e.g. pronouncing a word into another one which shares a similar sound.	(1) For the South Pacific, this is a critical time. (2) So it may not stay healthy much longer.	(1) From the South Pacific, this is a critical time. (2) So it might not stay healthy much longer.
	Addition: Adding a segment or segments to the end of a word.	(1) We depend on it. (2) The South Pacific Ocean is on the surface of it still a healthy ocean.	(1) We depends on it. (2) The Southern Pacific Ocean is on the surface of it still a healthy ocean.
	Swallowing: Swallowing part of the sound of a word when uttering.	It's changing in ways that, if left unchecked, could develop into a global crisis.	It's chan(g)ing in ways that, if left unchecked, could develop into a global crisis.
	Stress:	Over 60% of the world's	Over 60% of the world's fish

	Misplacing the correct stress position of a word and stressing the wrong syllables.	fish catch comes from the Pacific [pəˈsɪfɪk].	catch come from the Pacific [ˈpəsɪfɪk].

Table 3. Definitions and examples of fluency errors in IRA task in 2011 CELST

Error types		**Examples**	
		Original version	Incorrect sample(s)
Fluency errors	Error on sense group: Inappropriate pause (e.g. a sense group is split by wrong pausing) or lack of pause (e.g. speaking in a non-stop way).	The South Pacific is on the surface of it still a healthy ocean.	The South Pacific is on the surface of it still a / healthy ocean.
	Repetition： Uttering a word again.	But like all oceans, it has little or no protection.	But like all oceans, it has little or no protection, protection.
	Invalid mending: Pronouncing a word or a segment as a correction or an addition.	But like all oceans, it has little or no protection.	But like all oceans, it has little or no protection, tation. But like all oceans, it has little or no protection, person.
	Filling: Using some meaningless sounds in between two words, such as eh, hm, ah, uh, er, en, etc.	The South Pacific is on the surface of it still a healthy ocean.	The South Pacific is on the surface of uh, uh, a healthy ocean.
	Towing: Drawling part of the sound of a word.	We depend on it.	We depeeend on it.

Table 4. Definitions and examples of other errors in IRA task in 2011 CELST

Error types	**Examples**	
	Original version	Incorrect sample(s)
Word deletion: Missing or skipping some words in the text.	And protecting the fish will ensure a healthy ocean for all the marine life of the Pacific.	And protecting the fish will ensure a healthy ocean for all the marine life (of the Pacific).
Word addition: Adding a new word to the text.	The South Pacific is on the surface of it still a healthy ocean.	The South Pacific is on the surface of a it still a healthy ocean.
Intonation errors: Pronouncing within inaccurate or incorrect tones, e.g. wrong rising or falling tone, or no tones.	So it may↗ not stay↘ healthy much longer↗.	So it may → not stay → healthy much longer →.

Table 1 shows that the learners' IRA errors are primarily made up of phonetic errors (41 per cent) and fluency errors (45 per cent). Such findings are basically consistent with Gao's study [40] quoted above. The present study adopts large random sampling and carries out corpus-based rigorous analyses (i.e., testees' recordings in the 2011 CELST were made into a recording corpus). The validity of the results is therefore guaranteed.

The analysis of the error types found in 2011 CELST indicates that the errors are also closely related to the learners' inability to grasp the phonetic features of the

materials they imitated. For example, they had difficulty pronuncing the vowel sounds which are related to the shape and size of speakers' oral cavity (such as coronal high vowel). Correct pronunciation of these types are difficult for learners to attain. In addition, the fluency errors are usually caused by learners' confusion about the sense groups, i.e., different components in a sentence which are divided according to meaning and structure, of the utterances. Learners' perception of sense groups directly determines whether they can pause appropriately when performing an IRA task. If learners can perceive the sense groups visually in such a task and receive sufficient uttering training, their accuracy of perceiving sense groups can therefore be improved.

4 The Experiment

4.1 Research questions

In view of the mentioned common errors made by Chinese EFL learners in IRA and in light of multimodal discourse theory, this study attempts to explore how speech visualization technology in teaching can help to reduce learners' phonetic errors and fluency errors in speaking. There are two specific questions to be answered: (1) Whether speech visualization in teaching can effectively reduce learners' phonetic errors? (2) Whether speech visualization in teaching can help learners perceive sense groups correctly so that they can improve their oral fluency?

4.2 Subjects

This study targeted two parallel classes of similar English level (according to their scores in the Senior High School Entrance Examination) as the research subjects. They were Grade-One students in a senior high school in Panyu District, Guangzhou, China. They were taught by the same English teacher who was also responsible for the grading throughout the experiment. There were 40 students respectively in these two classes, one serving as the experimental class and the other as the control class.

4.3 Instruments

The instruments used in this study mainly include: (1) Cool Edit Pro 2.0 (CE) which was used to present the recording waveform of the materials and produce teaching courseware for classroom teaching and students' self-learning; (2) language laboratory in which students conducted self-directed IRA training and Pronunciation Power 2 (PP2) with which they have their utterances recorded and converted into waveforms for comparisons with the standard recording waveforms; (3) SPSS (version 18.0) with which the authors analyzed the experimental data; and (4) two test items of equivalence for pre-test and post-test respectively.

4.4 Process

This experiment lasted for three months. The following is a detailed introduction of this experiment.

Pre-test: In order to confirm that the experimental class and the control class were at the same level of IRA before the experiment, a pre-test for the two classes was conducted. Participants read the same material out in the same language laboratory, which was recorded as data resources. Then the English teacher graded the recordings according to the scoring criteria (Table 5) of IRA task in CELST (with a full score of 20).

Table 5. The scoring criteria of IRA task in CELST

Rank	Pronunciation and Intonation		Speed and Content	
	Score	***Standard***	***Score***	***Standard***
A	8-12	(1) Fluent and accurate pronunciation (2) Correct and natural intonation (3) Fluent and coherent utterance	6-8	(1) Reading aloud at the originally suggested speed (2) Complete the whole content (no more than three words skipped)
B	4-7	(1) Roughly correct pronunciation (2) Roughly correct intonation (3) Roughly fluent utterance	3-5	(1) Reading aloud roughly at the originally suggested speed (2) Several words skipped
C	0-3	(1) Most morphemes are mispronounced (2) Incorrect and not natural intonation (3) Broken utterance	0-2	(1) Not reading aloud at the originally suggested speed (2) Skipping a whole sentence or more than 10 words

To better analyse statistically, the authors converted all the scores into the scale of zero to ten points. A scoring sample is shown in Table 6.

Table 6. A scoring sample of IRA in 2011 CELST

Scene	Original text	An IRA sample of Candidate A	Scoring
1	The South Pacific is on the surface of it still a healthy ocean.	A south *pacifical* is on the face of it *see* a helpful oc…/	Total score: 7 (Level B) Pronunciation and Intonation: 4 This sample presents many mispronounced words (e.g.'critical' is sounded as 'correctical'), mainly flat intonation, and roughly fluent utterance. Speed and Content: 3
2	We depend on it.	We depend on it./	
3	Over 60% of the world's fish catch comes from the Pacific.	Over 60% of the world's fish catch come from the Pacific[ˈpəsɪfɪk]./	
4	But like all oceans, it has little or no protection,	But like all ocean it's/ little of no protection,/	
5	so it may not stay healthy much longer.	so it may not say healthy/*not* longer./	
6	For the South Pacific, this is a critical time.	From the South of Pacific[ˈpəsɪfɪk],/ this is a *correctical* time./	
7	It's changing in ways that, if left unchecked, could develop into a global crisis.	It's changing in way that, if left *much*, /could *delop* into a *go by colicy*./	
8	Some of its residents have been through crisis before.	Some of its *resaident* have *be flau colicy* before./	

9	And protecting the fish will ensure a healthy ocean for all the marine life of the Pacific.	And protect the *fit* will inside a health ocean of/ the many life of the *pastical./*	This candidate can basically follow the original speed, but misses some words (e.g 'ocean' is voiced as 'oc'.).
10	It will require international commitment and co-operation.	It will require *intonation* commitment and cooperation.	

* The words of mispronunciation are italic.
* The words of swallowing are underlined.
* The words of wrong stress are boldfaced.
* The pauses are marked with "/".

To test whether the difference of the converted scores (with a full score of ten) in IRA task between the two classes was significant, SPSS 18.0 was used to conduct an independent sample T-test on the data. The results are shown in Table 7.

Table 7. The independent-samples t-test results of pre-test

	Experimental class (*N=40*)		**Control class (*N=40*)**		**t**	**p (2-tailed)**
	M	***SD***	***M***	***SD***		
Pre-test	6.235	1.742	6.168	1.830	-0.064	.148

*Significance: p< 0.05

It can be concluded from Table 7 that there is no significant difference in the mean score of IRA performance between the two classes before the experiment ($t<1$, $p>0.05$), so it can be determined that the two classes were at the same initial levels.

Experiment materials: The preparation before the experiment included making ten units of learners' textbooks into speech coursewares. The key task was to conduct and offer the visual analysis of every text recording using the CE waveform, aiming at deepening learners' perceiving and understanding of the suprasegmental features (e.g. pronunciation, intonation, sense groups, and so on) through intuitively demonstrating audio, images, and text together during the class. At the same time, learners practiced reading aloud and strengthened their oral competence through IRA training. Fig.1 is the screenshot sample of the courseware using CE waveform to visualize the recording "I say to myself: Aren't they lucky?" from English Book 4, Unit1, Women of achieve.

Fig. 1. Annotation of words and sound intensity data added to CE waveform of text recording (The underlined are stressed syllables. The size of the letters indicates the strength of the voice.)

Practical teaching: According to the experimental plan, the teacher had started the experimental teaching since the second week of the new term in February 2012. The practical teaching was mainly divided into two parts: classroom teaching and after-class self-directed learning. Firstly, the teacher demonstrated the experimental courseware in the experimental class and assigned the students to do self practice at home and review it by themselves. Secondly, the students of the experimental class also needed to keep on IRA training in the language laboratory after class. The training applied CE and PP2 to visualize the recordings for IRA practice and to compare such self-made visualization with the template waveform of the recordings. Examples are shown in Fig.2 and Fig.3.

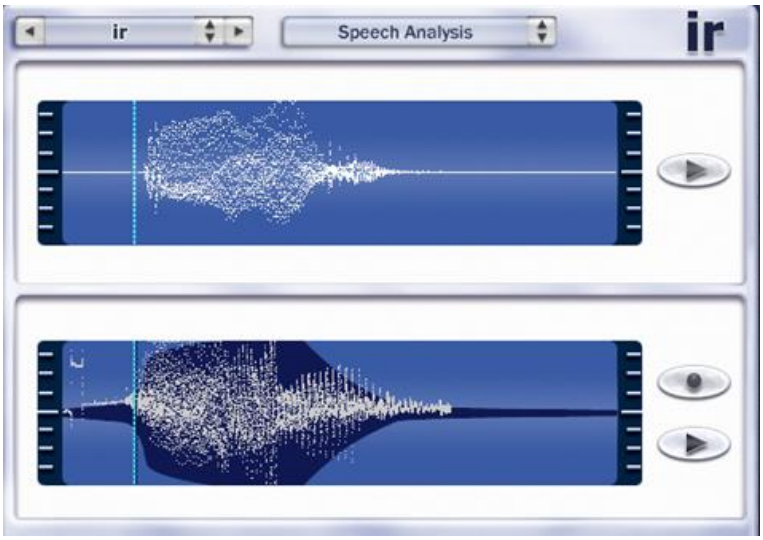

Fig.2. Learners used PP2 to make visualized comparison between the recordings

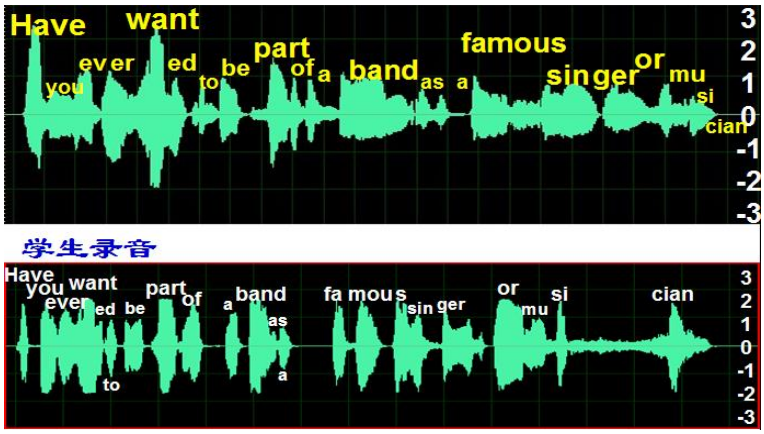

Fig.3. Learners used CE to make visualized comparison between the recordings (The upper one is the original recording. The lower one is the learners' imitation recording.)

Meanwhile, the control class did not adopt speech visualization approach and received training by means of merely repeating after the teacher in class and imitating the official recording materials after class. Each time a unit was finished, the teacher checked the students' IRA results through recordings and rated them based on three degrees (grade A, B, C, as illustrated in Table 5). At the same time, the teacher gave suggestions to the students for helping them solve their problems found from the recordings. Students were encouraged to discuss their learning strategies individually with the teacher.

After a twelve-week instruction, post-tests on both classes were carried out. The test content simulated the IRA task in CELST. Besides, the difficulty, speed, and duration of recording, and scoring criteria were all set to be equivalent to the requirements of the CELST. According to the research questions of this study, the key points of the test consisted of phonetic accuracy and speaking fluency. The evaluation of phonetic accuracy was mainly based on to what extent these subjects imitated the content, pronunciation and intonation of the text, while the assessment of speaking fluency was based on whether the subjects' imitation speed was appropriate and whether their division of sense groups was accurate. When the test was completed, the recordings of both classes were graded by the teacher and the relevant data collected.

Data collection: The 40 subjects in each of the two classes participated in the listening and speaking test by the end of the teaching experiment. Some recording samples were discarded due to failure of recording or poor quality in sound volume. 32 valid test samples of each class in the post-test were respectively collected.

Data analysis: This study compares the recording data of the experimental class and the control class in three aspects: content and accuracy in pronunciation, fluency and rhythm in speech, and the overall effect. Since the participants were students from two classes, the author used the independent sample T-test to compare the average value of each item of the two classes to test the significance of difference. The following is the analysis of the results.

Content and accuracy in pronunciation: As shown in Table 8, the mean score of the experimental class is higher than that of the control class in content and accurate pronunciation (4.59>4.22), and the difference is significant (p=0.027<0.05), which indicates that IRA assisted by visualization is conducive to train learners' accuracy.

Table 8. The independent-samples on content and accuracy in pronunciation of post-test

	Experimental class (*N=32*)		**Control class (*N=32*)**		**t**	**p (2-tailed)**
	M	***SD***	***M***	***SD***		
Post-test	4.59	.615	4.22	.706	2.265	.027

*Significance: p< 0.05

Fluency and rhythm in speech: Table 9 shows that in terms of fluency and rhythm (i.e., sense group), there is no significant difference in the mean score between the two groups (t=0.497<1, p=0.621>0.05), which reveals that the visualization method has not been proved to have a positive effect on learners' English speech fluency.

Table 9. The independent-samples on fluency and rhythm in speech of post-test

	Experimental class (*N=32*)		**Control class (*N=32*)**		**t**	**p (2-tailed)**
	M	***SD***	***M***	***SD***		
Post-test	3.72	.457	3.66	.545	.497	.621

*Significance: p< 0.05

Overall effect: Noted from Table 10, the mean score of the experimental group is higher than that of the other group (8.31>7.88) in the listening and speaking post-test, and the difference is significant (p=0.045<0.05). It represents that over all, the method of speech visualization adopted in this study can promote IRA training.

Table 10. The independent-samples on overall effect of post-test

	Experimental class (*N=32*)		**Control class (*N=32*)**		**t**	**p (2-tailed)**
	M	SD	M	SD		
Post-test	8.31	.896	7.88	1.271	1.052	.045

*Significance: p< 0.05

5 Reflections

This study found that speech visualization technology can significantly help learners improve accuracy in their English speech. However, it has less satisfactory results in training learners' fluency and rhythm in speech. Four aspects are considered when reflecting on the experiment process and the analysis of the data obtained. Firstly, the time to implement the experiment is not sufficient enough. Improvement of fluency and rhythm requires long-term training, especially the perception and

division of sense groups which require continuously longer time exposure to a larger amount of language input. Secondly, the participants usually need time to read the content on the screen before imitating and reading out the text, thus their voice cannot be synchronized with the content on the screen, which is consistent with most learners' behavior. Thirdly, learners are accustomed to listening to the recording first and then imitating in the way they were daily trained to do. As a result, they might have difficulty keeping up with the speed of the recordings. Lastly, the visual analysis and presentation of speech may increase the amount of information loaded in learners' IRA process, thereby objectively reducing learners' fluency.

By pinning down the aspects which account for a lack of improvement in terms of learners' speech fluency and sense group awareness, feasible ways to solve these problems can be obtained. The following two points are thus proposed.

5.1 Shadowing

As this study focuses on learners' independent use of speech visualization technology, not much guidance to IRA was given to learners during the experiment. Besides, learners' existing imitation habit makes it difficult for them to keep up with the switching speed of the screen text when their oral reading responses were recorded. In other words, learners are in need of scientific methods which may help them improve their oral fluency. According to the influence of speech visualization on speaking fluency and due to the fragmentary feature (for example, sentence by sentence) of speech visualization materials, a combination of shadowing method and speech visualization materials is highly recommended. Shadowing refers to the act or task of listening in which the learners track the speech they hear and repeat it as exactly as possible [39, 42]. When listening to a chapter of language learning recording, learners do not take notes but only follow and read with the speaker on the recording synchronously, trying their best to imitate the speakers' pronunciation, intonation and tongue [43]. This method helps learners practice both pronunciation and intonation and exercise their vocal organs in a steady way, and develops their habits of following the model recording sentence by sentence carefully and raise their awareness of identifying and corresponding with the original text. Shadowing is considered to bring about an obvious promoting effect on early-staged non-English-major learners. The visual materials in this experiment are presented sentence by sentence. Therefore, the combination of the visual materials and the shadowing method can improve learners' accuracy more effectively, and increase their fluency and ability to perceive sense groups efficiently.

5.2 Aural modality assisted with visual modality

Speech visualization is an implementation form of multimodal discourse theory. One of the basic premises of multimodal discourse theory is that the increase of modalities can bring more information and learning experience to learners at the same time. In this experiment, visual modality was added to the traditional single aural modal training. Learners were exposed to multimodalities which brought them more

information and experience. However, it also causes overload of information to learners in the process of IRA, which objectively making fluency more difficult to achieve. This problem occurred because of the neglect of interaction and differentiation among modalities during the designing of visual materials. Besides, the aim of this experiment, which targets at improving learners' accuracy, is also neglected to some extent. In fact, the "listen and repeat" mode is still the most significant activity in this experiment, which indicates that the aural modality should still be dominant while visualization process only performs as a subsidiary means. Therefore, more attention should be paid to the amount of information input when the designed materials are expected to reduce learners' burden of visual information load which may significantly affect learners' fluency in speaking. Some detailed information such as intensity data can be presented in the correction materials after learners' IRA practice. The results also show that in multimodal teaching, modalities strengthen or complement each other. A more flexible way should be further explored in dealing with the complex relationships among various modalities.

6 Acknowledgement

This work is supported by the Center for Language Cognition and Assessment, South China Normal University, and by Guangdong "13th Five Year" Plan Co-funded Project of Philosophy & Social Science (GD16XWW25).

7 References

[1] Madsen, H. S. (1983). Techniques in Testing. Oxford University Press.
[2] Halliday, M. A. K. (1994). An Introduction to Functional Grammar. London: Edward Arnold (Publishers) Limited.
[3] Pinker, S. (2015). The sense of style: The thinking person's guide to writing in the 21st century. Penguin Books. https://doi.org/10.5860/choice.191126
[4] Warner, L., Crolla, C., Goodwyn, A., Hyder, E., & Richards, B. (2016). Reading aloud in high schools: students and teachers across the curriculum. Educational Review, 68(2), 222-238. https://doi.org/10.1080/00131911.2015.1067881
[5] Wang, Z.Y. (2002). Reading aloud and English learning. Foreign Languages and Their Teaching 8: 51-52.
[6] Tian, F. (2018). Influences of the Accent Perception on Production in Oral Reading Speech. Foreign Languages and Their Teaching 3:55-64+144.
[7] Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. The handbook of second language acquisition, 14, 63. https://doi.org/10.1002/9780470756492.ch4
[8] Ding, Y. (2007). Text memorization and imitation: The practices of successful Chinese learners of English. System, 35(2), 271-280.https://doi.org/10.1016/j.system.2006.12.005
[9] Maybury, M. T. (1994). Intelligent multimedia interfaces. In Conference Companion on Human Factors in Computing Systems (CHI '94), Catherine Plaisant (Ed.). ACM, New York, NY, USA, 423-424. https://doi.org/10.1145/259963.260410

[10] Hill, D. R. (2000). Give us the tools: A personal view of multimodal computer–human dialogue. The Structure of Multimodal Dialogue II. Amsterdam: John Benjamins, 25-62. https://doi.org/10.1075/z.99.04hil
[11] Gu, Y. G. (2007). On multimedia learning and multimodal learning. Computer-assisted Foreign Language Education, 29(2), 3-12.
[12] Holsanova, J. (1999). Olika perspektiv pa språk, bild och deras samspel. Metodologiska reflexioner.(Different perspectives on multimodality. Methodological considerations) In: Inger Haskå & Carin Sandqvist (eds.): Alla tiders språk. Lundastudier i nordisk språkvetenskap A 55. Alla tiders språk. Lundastudier i nordisk språkvetenskap A, 55, 117-26. https://doi.org/10.7557/4.2482
[13] Hu, Z. L. (2007). Multimodalization in social semiotics. Language Teaching and Linguistic Studies, (1), 1-10.
[14] O'Halloran, K. L. (2011). Multimodal discourse analysis. Continuum companion to discourse analysis, 120-137.
[15] Jewitt, C., Bezemer, J. J. and O'Halloran, K. L. (2016). Introducing multimodality. London: Routledge. https://doi.org/10.4324/9781315638027
[16] Jiang, X. Q. and Ding Y. (2012). A theoretical framework of New College English Teaching Model. Computer-Assisted Foreign Language Education, 6.
[17] Zhuang, M. Q. and Bu, Y. H. (2011). Research on speech visualization teaching. e-Education Research, 2, 92-98.
[18] Hincks, R. (2003). Speech technologies for pronunciation feedback and evaluation. ReCALL, 15(1), 3-20. https://doi.org/10.1017/S0958344003000211
[19] Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. System, 32(4), 505-524. https://doi.org/10.1016/j.system.2004.09.009
[20] Chapelle, C. A. (2004). Technology and second language learning: expanding methods and agendas. System, 32(4), 593-601. https://doi.org/10.1016/j.system.2004.09.014
[21] Godwin-Jones, R. (2009). Speech tools and technologies. Language Learning & Technology, 13(3), 4-11. http://dx.doi.org/10125/44186
[22] Gorjian, B., Hayati, A., & Pourkhoni, P. (2013). Using Praat software in teaching prosodic features to EFL learners. Procedia-Social and Behavioral Sciences, 84, 34-40. https://doi.org/10.1016/j.sbspro.2013.06.505
[23] Boersma, P. and Weenik, D. (2017). PRAAT: A System for Doing Phonetics by Computer. v. 6.0. 31, retrieved August 22, 2017. Institute of Phonetics Sciences, University of Amsterdam.
[24] Brett, D. (2004). Computer generated feedback on vowel production by learners of English as a second language. ReCALL, 16(1), 103-113. https://doi.org/10.1017/S0958344004000813
[25] Wilson, I. (2008). Using Praat and Moodle for teaching segmental and suprasegmental pronunciation. In Proceedings of the 3rd international WorldCALL Conference: Using Technologies for Language Learning (WorldCALL 2008) (pp. 112-115).
[26] Dixon, T. (2018). Teaching Pronunciation in Integrated Skills Classes. The TESOL Encyclopedia of English Language Teaching, 1-6. https://doi.org/10.1002/9781118784235.eelt0691
[27] Wallace, L. R., & Lima, E. F. (2018). Technology for Teaching Pronunciation. The TESOL Encyclopedia of English Language Teaching, 1-7. https://doi.org/10.1002/9781118784235.eelt0443
[28] Martin, P. (2004). Winpitch LTL II, a multimodal pronunciation software. In InSTIL/ICALL Symposium 2004.

[29] Tamandani, K. K. (2017). Teaching Stress and Intonation to Improve Intelligibility of Undergraduate Students of AMU: A Call based Approach (Doctoral dissertation, Aligarh Muslim University).
[30] Chun, D. M. (2012). Computer-Assisted Pronunciation Teaching. The encyclopedia of applied linguistics. https://doi.org/10.1002/9781405198431.wbeal0172
[31] Fox, R. A. (2017). Teaching an online phonetics course: One approach. The Journal of the Acoustical Society of America, 142(4), 2617-2617. https://doi.org/10.1121/1.5014583
[32] Levin, M. H. (1999). Use of a soundcard in teaching audio frequency and analog modem communications. ACM SIGCSE Bulletin, 31(3), 79-83. https://doi.org/10.1145/384267.305868
[33] Derman, S., Bardakçı, M., & Öztürk, M. S. (2017). An investigation of read speech of Arabic students learning Turkish as a second language in terms of stress and pause. Dil ve Dilbilimi Çalışmaları Dergisi, 13(1), 215-231.
[34] Sariman, G. and Çetin, D. (2018). The Effect of Computer Aided Education on the Skill of Word Stressing in Teaching Turkish as a Foreign Language. Universal Journal of Educational Research, 6(8), 1701-1709. https://doi.org/10.13189/ujer.2018.060811
[35] Zoghbor, W. S. (2018). Teaching English pronunciation to multi-dialect first language learners: the revival of the Lingua Franca Core (LFC). System, 78, 1-14. https://doi.org/10.1016/j.system.2018.06.008
[36] Sun, X. P. (2008). Prominence of Prepositions in Chinese EFL Learners Read Speech. Journal of PLA University of Foreign Languages, 3, 92-98.
[37] Zhou, A. J., and Zhang, C. (2006). Measuring English Oral Fluency with COOL EDIT PRO. Media in Foreign Language Instruction, 4, 67-70. https://doi.org/10.3969/j.issn.1001-5795.2006.02.013
[38] Xie, P. (2007). Application of Speech Analyser to English Pronunciation Teaching. Media in Foreign Language Instruction, 6, 26-29. https://doi.org/10.3969/j.issn.1001-5795.2007.06.005
[39] Nunan, D. (2004). Practical English Language Teaching. Higher Education Press, 55&119.
[40] Gao, X. (2006). A Study of Chinese EFL Learner Oral Reading Miscues. Journal of PLA University of Foreign Languages, (5), 53-58. https://doi.org/10.3969/j.issn.1002-722X.2006.05.011
[41] Ying, X., & Yongqiang, Z. (2017). A Study of Pronunciation Features and Score Predictors for the Read-aloud Task. Journal of Xi'an International Studies University, (2), 14. https://doi.org/10.16362/j.cnki.cn61-1457/h.2017.02.014
[42] Mochizuki, M. (2006). Exploring the application of shadowing to Japanese education. Shicho kaku kyoiku [Audio-Visual Education], 6, 37-53.
[43] Song, J. F. (2000). Itemized Training Methods in Interpretation teaching. Foreign Language Teaching, 4, 35-39.

8 Authors

Xiaobin Liu is an associate professor of School of Foreign Studies, South China Normal University. His research interests include educational technology & EFL teacher development, computer assisted language learning.

Diying Wu is a master student of School of Foreign Studies, South China Normal University.

Yiwen Ye is a master student of School of Foreign Studies, South China Normal University.

Manfei Xu is an associate professor of School of Foreign Studies, South China Normal University. Her research interests include corpus application in language teaching and teacher education. She is also a corresponding author of this paper.

Jianli Jiao is a professor of School of Information Technology in Education, South China Normal University.

Wenheng Lin is an English teacher of Guangzhou Nansha Dongchong Middle School.

Article submitted 2019-08-07. Resubmitted 2020-01-31. Final acceptance 2020-02-03. Final version published as submitted by the authors.