

PTeacher: a Computer-Aided Personalized Pronunciation Training System with Exaggerated Audio-Visual Corrective Feedback

Yaohua Bu*

Academy of Arts & Design, Tsinghua University

Tianyi Ma*

Department of Computer Science and Technology, Tsinghua University

Weijun Li

School of Information Science and Technology, Northeast Normal University

Hang Zhou

The Chinese University of Hong Kong

Jia Jia†

Department of Computer Science and Technology, Tsinghua University

Shengqi Chen

Department of Computer Science and Technology, Tsinghua University

Kaiyuan Xu

Department of Computer Science and Technology, Tsinghua University

Dachuan Shi

Department of Computer Science and Technology, Tsinghua University

Haozhe Wu

Department of Computer Science and Technology, Tsinghua University

Zhihan Yang

Department of Computer Science and Technology, Tsinghua University

Kun Li

SpeechX Limited

Zhiyong Wu

Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Shenzhen International Graduate School, Tsinghua University

Yuanchun Shi

Department of Computer Science and Technology, Tsinghua University

Xiaobo Lu

Academy of Arts & Design, Tsinghua University

Ziwei Liu

S-Lab, Nanyang Technological University

ABSTRACT

Second language (L2) English learners often find it difficult to improve their pronunciations due to the lack of expressive and personalized corrective feedback. In this paper, we present Pronunciation Teacher (*PTeacher*), a Computer-Aided Pronunciation Training (CAPT) system that provides personalized exaggerated audio-visual corrective feedback for mispronunciations. Though the effectiveness of exaggerated feedback has been demonstrated, it is still unclear how to define the appropriate degrees of exaggeration when interacting with individual learners. To fill in this gap, we interview 100 L2 English learners and 22 professional native teachers to understand their needs and experiences. Three critical metrics are proposed for both learners and teachers to identify the best exaggeration levels in both audio and visual modalities.

Additionally, we incorporate the personalized dynamic feedback mechanism given the English proficiency of learners. Based on the obtained insights, a comprehensive interactive pronunciation training course is designed to help L2 learners rectify mispronunciations in a more perceptible, understandable, and discriminative manner. Extensive user studies demonstrate that our system significantly promotes the learners' learning efficiency.

CCS CONCEPTS

- Human-centered computing → User studies.

KEYWORDS

Computer-Aided Pronunciation Training System; Audio-Visual Corrective Feedback; Language Learning; Exaggerated feedback

1 INTRODUCTION

Pronunciation plays a crucial role in English learning for second-language (L2) learners. However, a majority of L2 English learners encounter difficulties in improving their pronunciation. On the one hand, they are prone to pronounce L2 words in the tongue of their first language, which is called the negative influence of language transfer [17, 21, 35, 60, 79]. This causes multiple inconspicuous mispronunciations that can hardly be rectified by themselves. On the other hand, professional native-speaking English teachers, who

*Equal contribution.

†Corresponding author, jjia@tsinghua.edu.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '21, May 8–13, 2021, Yokohama, Japan

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8096-6/21/05...\$15.00

<https://doi.org/10.1145/3411764.3445490>

can diagnose pronunciation problems and give corrective feedback [51], are in pressing demand. According to statistics provided by the British Council, while approximately 1.5 billion English learners [18] exist, only 250 thousand native speakers are qualified to serve as English teachers. It is also reported [20] that 67% of ESL teachers survey in Canada have no training in pronunciation instruction. The insufficiency of native English teachers is even worse in underdeveloped areas.

Driven by the demand, quite a number of Computer-Aided Pronunciation Training (CAPT) systems have been proposed [2, 33, 57–59, 67] with feedback from both audio and visual modalities [46, 54, 55, 86, 87]. While most of them focus on synthesizing natural speech, their systems are designed to show only the pairwise differences between the correctly pronounced phonemes and the mispronounced ones. The teaching effect of such strategy proves to be less significant due to L2 learners' weaker perceiving ability [22]. Particularly, the difficulty for perceiving and correcting pronunciation grows along with age [39, 73]. As a result, more *perceptible* and *distinctive* feedback is in demand.

In offline English classes, exaggerated feedback has been a effective feedback strategy for teachers to rectify the pronunciation of learners [62, 73]. Alghamdi, et al. [1] have proven that visually exaggerated speech is capable of promoting the perceptual ability of learners. However, this direction has rarely been explored in the field of human-computer interactions. Only Zhao et al. [89] propose a basic pipeline for audio-visual exaggeration, but focusing more on implementing than evaluation. No practical training paradigm is constructed. Consequently, the problem of what compromises a good exaggeration system remains unsettled.

In this paper, we provide systematical studies towards designing personalized exaggerated feedback in a CAPT system named Pronunciation Teacher (*PTeacher*). We focus on constructing a participatory exaggerated pronunciation corrective system that emphasizes enlarging the teaching effectiveness in a *user perception* point of view. Specifically, our system evaluates learners' pronunciation with real-time mispronunciation detection and diagnosis (*MDD*) algorithm [49]. Then exaggerated feedback is given in the form of audio and articulated animation. Importantly, we point out that through *extensive human evaluation*, the following key issues can be thoroughly discussed: 1) how to define the fine-grained parameters of exaggeration for both audio and visual modalities so that corrective feedback can be distinctive enough; 2) how personalized responses can be made to L2 learners with different degrees of proficiency; and 3) how much the proposed designed training course can positively affect the learning efficiency of L2 learners compared with traditional CAPT system.

To this end, a total number of 100 L2 learners together with 22 professional native English teachers have participated in our studies. Among them 30 learners and 22 teachers are responsible for determining the set of exaggeration parameters which renders the best feedback performances on three key aspects: *distinguishability*, *understandability* and *perceptibility*. Afterward, we leverage comprehensive user studies on 30 learners to connect different levels of exaggeration feedback with different levels of proficiency, which has never been discussed before. We group and evaluate the

learners through both objective scores reported from *MDD* and subjective evaluations from native teachers. Then the best personalized feedback level suitable for each learner can be determined.

Equipped with these necessary data and analysis, we include an interactive pronunciation training course into the *PTeacher* system, aiming to improve learners' engagement. Along with the course, our system evaluates the English proficiency of each learner in a life-long manner and provides flexible exaggerated audio-visual corrective feedback. Therefore, personalized exaggeration can be given according to real-time mispronunciation detection from *MDD* as well as accumulated evaluation. User studies on the system demonstrate that our *PTeacher* with the exaggerated feedback enhances the learners' pronunciation accuracy by 14.19% and 27.55% for learners with a higher and lower degree of proficiency respectively, within a short time of learning (1 hour).

The contributions of our work are listed as follows:

- We define and identify the most suitable set of parameters in exaggerated corrective pronunciation feedback in both audio and visual modalities from three critical aspects.
- We propose personalized exaggerated feedback according to English proficiency of the learner.
- We design the audio-visual corrective CAPT system, *PTeacher*, which includes a pronunciation training course. The course can dynamically evaluate learners' English pronunciation proficiency in a life-long manner.
- We support all of our findings and analysis with extensive user studies conducted on 100 second-language English learners and 22 professional native teachers. Comprehensive results demonstrate the advantage of our proposed exaggerated training system as well as the effectiveness of each module.

2 BACKGROUND

Computer-aided or assisted learning is an important research area in both human-computer interactions (HCI) [10, 66] and language learning [27, 61]. We focus on design a Computer-Aided Pronunciation Training System (CAPT) with exaggerated audio-visual feedback. In this section, we will first present the learning theories behind our audio-visual exaggerated feedback design (Section 2.1). Then we demonstrate the connections between our work and the recent advances in HCI (Section 2.2). Finally, we review and discuss the related works in the specific area of CAPT (Section 2.3).

2.1 Learning Theories

Theories on Audio-Visual Feedback. Information processing in speech and language communication is bi-modal. For example, language learners not only listen to the speaker but also observe the speaker's articulatory movements [12, 40]. In the visual modality, phoneticians have summarized that articulatory phonetics are strongly correlated with the manners and places of articulation [19, 30, 68, 82], articulators [38] and airflow [26]. In the auditory modality, the principles of phonology and phonetics dictate and explain the ways humans make sounds [72, 78, 83]. These theories lead to our design of providing both audio and visual information as feedback. While previous methods show improvements in L2 learners' pronunciation abilities through training with articulatory

animations [3, 36], our user study shows that with exaggerated feedback from both modalities, the learning efficiency can be further improved compared to providing only exaggerated audio feedback, which supports the hypotheses of the theories.

Exaggerated Feedback. Numerous studies have suggested that many L2 speech production (pronunciation) difficulties are rooted in perception [15, 22, 23, 60, 76]. Moreover, it has been exemplified that reinforcing the perception ability of learners can significantly contribute to the speech production ability automatically [4, 8, 43, 44, 70, 80]. Exaggerated audiovisual feedback is a particular kind of perception reinforcement, which corrects the pronunciation by strengthening the user's visual or auditory attention. For example, by enhancing the duration of the audio of a nasal consonant, brain plasticity at the perceptual and pre-attentive neural levels can be strengthened [13, 14]. Increasing the movement of the animation, the user's visual perception of graphics can be enhanced. Specifically, exaggerated movement can lead to more memorable perception than non-exaggerated movement [28]. Therefore, we propose the exaggerated-feedback to improve the perceptual effect of the target phonemes in both audio and visual modalities.

2.2 Language Learning and Exaggerated Feedback in HCI

As there are rarely any studies that target exaggerated feedback in language learning specifically in HCI, we look into the studies that contribute mostly to the sub-area of language learning and exaggerated feedback. Previous works on language learning mainly focus on the effectiveness of computer aided training and the components that matters. Ambra et al. [66] investigated whether a language learning system can help young L2 learners improve word-level pronunciation skills. They also provided a fundamental evaluation of the effectiveness of computer-assisted language learning systems. Robertson et al. [77], advocated that new interactive designs supporting collaboration can be used to overcome engineering limitations. In our work, we also propose interactive courses in our system. Hailpern [29] introduced Spoken Impact Project (SIP) to examine the effect of audio and visual feedback on vocalizations in ASD children. Their experimental results suggested that individual customization is in demand given the children's varied preferences on different styles of feedback. This inspires us to provide personalized feedback. As for exaggerated feedback, Antti et al. [28] contributed a controlled experiment of exaggerating the teaching avatar's flexibility in a kicking task. The experimental results demonstrate that users prefer exaggerated results over original ones. Our work shares similar insights with their design.

Based on these studies and the previous learning theories on exaggerated feedback as discussed above, we focus on constructing a participatory exaggerated computer-aided pronunciation training system that emphasizes enlarging the teaching effectiveness from a user perception perspective. Specifically, we study how to identify suitable exaggerated feedback to learners with different demands or behaviors in language learning (i.e. different pronunciation proficiencies) and how to define the best set of feedback. Our idea of involving exaggerated feedback and our system of determining the best set of exaggeration parameters can be beneficial for the

general area of computer-aided language learning [25, 29, 66, 77] and exaggerated feedback [28] systems in HCI.

2.3 Computer-Aided Pronunciation Training Systems

Development of CAPT. Computer-aided pronunciation training (CAPT) system was introduced in the 1960s. The first CAPT system was developed by Kalikow and Swets [34]. They developed a system that used visual feedback to teach English pronunciation for Spanish learners. From 2000 to 2010, most English CAPT systems utilized speech recognition technology, but few of them offered instruction or feedback to learners [37]. From 2010 to the present, researches incorporated diverse technologies into the CAPT system, including pronunciation training method [65], Automatic Speech Recognition (ASR) [64], Mispronunciation Detection and Diagnosis (MDD) [45], speech synthesis [57, 57, 67], visual-speech synthesis [87] and application system design [87]. Several kinds of online pronunciation corrective feedback in the form of different modalities are proposed in CAPT [2, 6, 53, 86, 87], such as audio feedback, text feedback, articulatory feedback [86], etc. Yuen et al. [87] proposed a comprehensive method to produce audio-visual feedback in CAPT. In the next step, they extended their work to a distributed text-to-audio-visual-speech synthesizer (TTAVS) to design a CAPT system with the interactivity on a mobile platform [46]. Pennington [71] found that phonology knowledge was not well considered in most CAPT systems. Inspired by the phonology research, our work finds a suitable adjusting range for expressive speech in audio-visual corrective feedback.

Exaggerated Audio-Visual Feedback in CAPT. The above discussed methods fail to increase awareness of learners towards their mispronunciation. Thus, identifiable and perceptible feedback is still urgently needed. According to [76] [85], in the offline English class, exaggeration is a critical feedback method for the teachers to rectify the pronunciation of learners. Typical exaggerating methods include speaking louder and slower, and showing the movements of mouth clearly to learners. Exaggerations in the form-focused instruction [85] have been verified to be beneficial for inexperienced L2 learners. Alghamdi et al. [1] investigated that exaggeration of the visual speech improved the audio-visual recognition of many phoneme classes. Exaggeration methods were used in CAPT systems to assist L2 learners in perceiving stress patterns. Their work provided a fundamental theory, which discussed the effectiveness of exaggeration methods. For the exaggerated audio-visual feedback in CAPT, Zhao et. al [89] proposed an audio-visual exaggeration method to provide more perceptible corrective feedback. Both exaggerated audio and exaggerated articulatory animation were provided for learners to rectify their pronunciation. However, several problems of audio-visual exaggeration in CAPT remain unsolved. In our PTeacher system, we systematically discuss these problems and present a practicable solution through extensive user studies.

3 PTEACHER SYSTEM FOR PRONUNCIATION TRAINING

In this section, we introduce *PTeacher*, a CAPT system that helps L2 learners correct their pronunciation by incorporating exaggerated audio-visual feedback into a pronunciation course. The whole

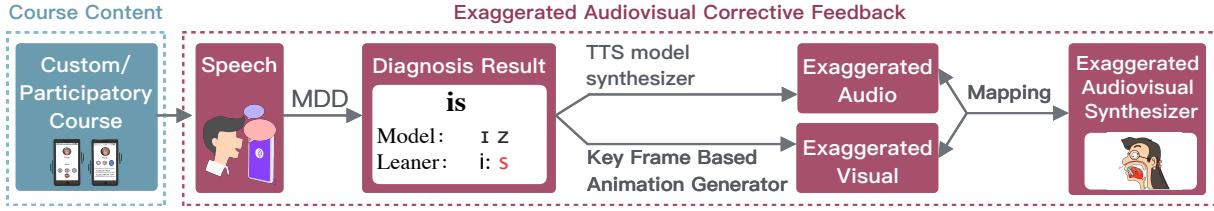


Figure 1: The whole working flow of PTeacher system. It consists of two key components: (1) exaggerated audio-visual corrective feedback generator, and (2) pronunciation training courses. Notably, the system is first served as the platform for user interactions, then the users' feedback also identifies the detailed design of the system.

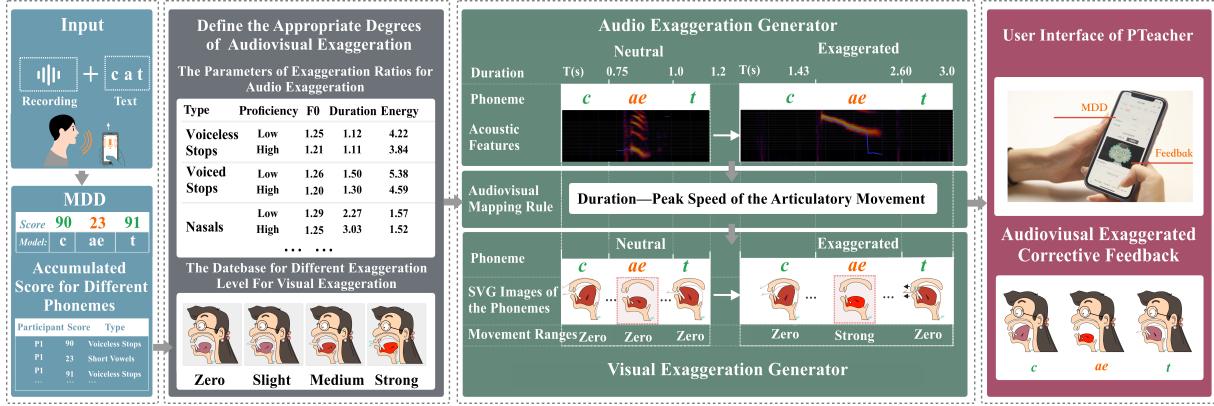


Figure 2: The working flow of the exaggerated audio-visual corrective feedback. Mispronunciation detection and diagnosis (MDD) systems are employed to diagnose pronunciation mistakes at sentence, word, and phoneme levels. Afterwards, personalized feedback are provided by the *audio* and *visual* exaggeration generators based on the MDD results.

pipeline of the system is illustrated in Figure 1. It consists of two key components: (1) exaggerated audio-visual corrective feedback generator, and (2) personalized pronunciation training courses. Notably, the system is first served as the platform for user interactions, then the users' feedback also identifies the design of the system.

3.1 Exaggerated Audio-Visual Corrective Feedback

Exaggerated audio-visual feedback is generated through an audio and a video exaggeration generator based on the different pronunciation situations of each users, which is the key feature of our system. As shown in Figure 2, Mispronunciation Detection and Diagnosis (MDD) systems [47–49] are firstly employed to detect and diagnose pronunciation mistakes at sentence, word, and phoneme levels. Afterwards, the MDD results are taken by both the *audio* and *visual* exaggeration generator to generate exaggerated feedback.

3.1.1 Accumulated Pronunciation Diagnosis. Different from previous systems [89], the historical MDD results are also considered to determine the exaggeration level. The MDD results are exponentially decayed and accumulated using the following equation:

$$R = (1 - \alpha)^n R_n + \sum_{0 \leq i \leq n-1} \alpha(1 - \alpha)^i R_i \quad (1)$$

where n is the number of all results, R_0 is the latest result, R_i is the i -th historical result and α is the decay ratio. In our research, α is

set to 0.9. After accumulated, the phoneme with the lowest score in one sentence will be selected and exaggerated by the audio and video exaggeration generators.

3.1.2 Audio Exaggeration Generator. The framework of the audio exaggeration generator is depicted in Figure 3. The exaggerations are generated by applying adjustment of proper exaggeration level to synthesized speech based on the MDD result.

A pre-trained Text-To-Speech (TTS) model [74] is used to synthesize high-quality, neutral speech with the given text. Montreal Forced Alignment (MFA) [56] algorithm is leveraged to locate the position of the selected phoneme in the synthesized speech. Then pitch, duration and energy of the selected phoneme are exaggerated with the parameters of the corresponding exaggeration ratios with PyWorld [31]. The exaggeration level for the selected phoneme is determined by the accumulated score as described in Section 3.1.1. Two exaggeration ratios are used in our research. The scores in [0, 50] are projected to the "Low Proficiency" exaggeration ratios and scores in [50, 100] are projected to the "High Proficiency" exaggeration ratios. The parameters for each exaggeration ratios are determined by personalized audio exaggeration experiment 4.2.1.

3.1.3 Visual Exaggeration Generator. We adopt the design of articulatory animation [11, 42, 75] to provide visual exaggeration and further increase the ability of expression in the visual domain. Our visualization plots three components: articulatory movement,

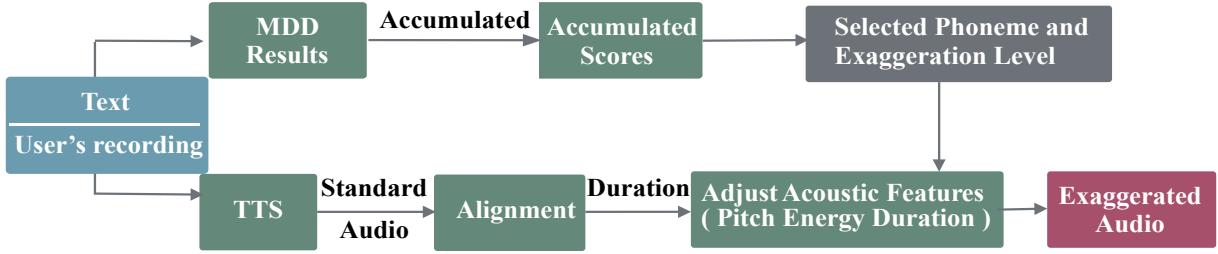


Figure 3: The working flow of the audio exaggeration generator. Text-To-Speech (TTS) model is used to synthesize neutral speech with given text. Montreal Forced Alignment (MFA) algorithm is leveraged to locate the position of the selected phoneme in the synthesized speech. The exaggeration level is determined by the accumulated score in the MDD results. Then pitch, duration and energy of the selected phoneme are exaggerated with the parameters of the corresponding exaggeration level with PyWorld.

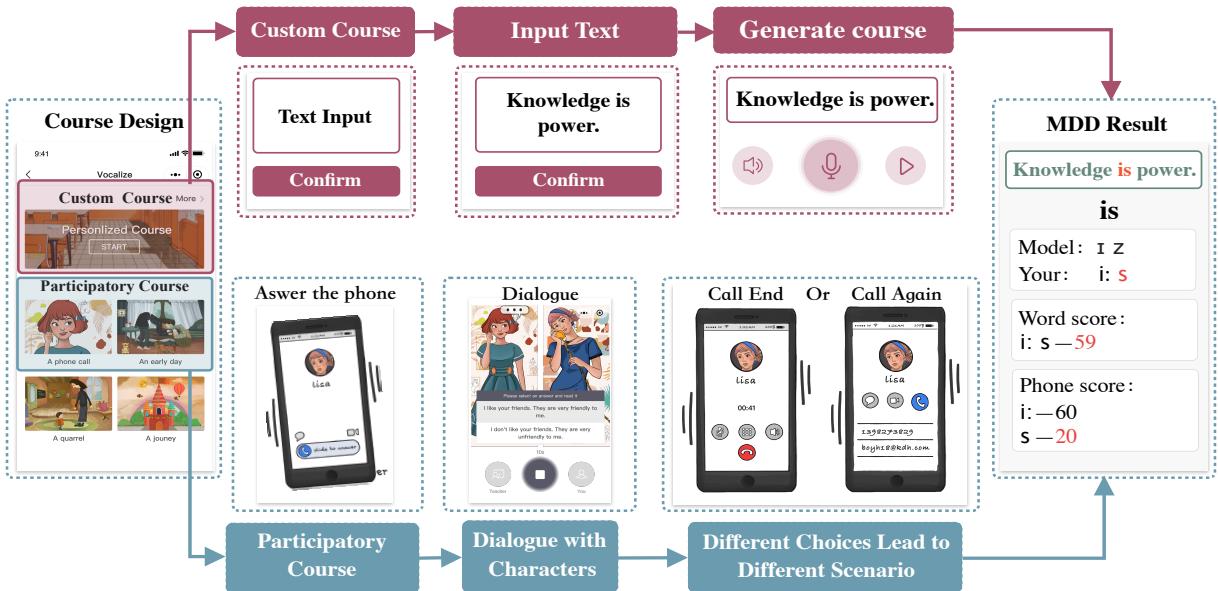


Figure 4: The workflow of the interactive courses and the custom course. In the interactive courses, user learn the drama courses. In the custom course, the system will generate a flexible course according to the English text which the user input.

tongue color, and auxiliary sign. For the articulatory movement, we plot the key parts of the articulatory actions (oral cavity), while irrelevant parts (*oesophagus, epiglottis, nasal, etc.*) are simplified. Four degrees of exaggerated side-view and front-view viseme components are designed under the guidance of articulatory phonemeticians and animation designer. Having obtained the articulatory plots of each phoneme, we leverage the Viseme Blending [24] to interpolate the overall animation.

We generate the exaggerated articulatory animation using the following methods: (1) increase the amplitude of key articulatory movement. (2) We modify the color of the tongue from low purity to high purity when exaggeration is needed, to draw the attention of learners. (3) Auxiliary graphics (e.g. arrows, airflow) and supplemental texts (e.g. *manner of articulation*), are finally added to help learners better understand the pronunciation through visualization, as shown in Figure 5. It is noteworthy that we update the design

in the amplitude of articulatory movements according to human interactions. Different degrees of exaggerated will also be provided to different users in a personalized manner. Details are illustrated in Section 4.3.

3.2 Personalized Course Content Design

We propose to personalize the pronunciation training through our novel course design to build the connection between general CAPT systems and individual users. As shown in figure 4, two types of courses, namely *Custom Course* and *Participatory Course* are designed.

The *Participatory Course* is actually a particular type of Interactive Participatory Drama (IPD) [32] in language learning. In this type of course, learners play active roles [7] in pre-programmed

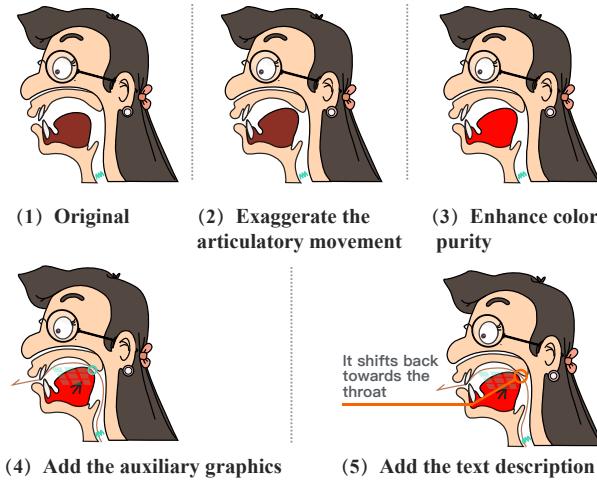


Figure 5: We exaggerate the mouth movement, the color of the key organs and add auxiliary graphics as well as multiple text descriptions.

scenarios by haptic and voice interaction (such as chatting, painting, etc). The course includes multiple storylines, depending on the choice of learners in the story. The primary purpose is to stimulate learners to do more perception (hear) and productive training (speak) in a participatory manner. While in *Custom Course*, learners could customize the content of the lessons based on their particular needs.

Both courses take recorded audios from users as inputs, and leverage the aforementioned exaggerated audio-visual feedback for users to recognize and rectify their mistakes. The learners' proficiencies are determined by *MDD* scores. The system assesses users' pronunciation accuracies on 14 types of phonetic symbols and makes a comprehensive judgment according to the user's immediately performance and historical proficiency. Thus personalized feedback can be dynamically generated in a life-long manner. We expect users to actively participate in pronunciation learning by noticing their improvements through reading the comprehensive reports from *PTeacher*.

4 USER-PARTICIPATED EXPERIMENTS

In general, with the users-participated experiments, we determine the optimum audio exaggeration ratios, find the most appropriate visual exaggeration level and verify the effectiveness of *PTeacher*. Participant are interviewed during the experiments and grounded theory approach [16] is leveraged to analyze the interview data.

- In audio experiments, we firstly define four audio exaggeration levels and determine the exaggerations ratios for four levels (Section 4.2.1). For each audio exaggeration level, we evaluate a) Distinguishability (Section 4.2.2); b) Understandability (Section 4.2.3); and c) Perceptibility (Section 4.2.4). Based on the result of the experiments mentioned above, we apply nonlinear fitting to determine the optimum exaggeration ratios (Section 4.2.5).

- In visual experiments, we first define multiple exaggeration levels for articulatory movements and tongue colors (Section 4.3.1). Then, we find the optimum exaggeration level with the highest user perceptibility rate (Section 4.3.2).
- In supplementary experiments, we evaluate the effects of user engagement and user experience between participatory course and custom course which is described in the supplementary materials (Chapter 1).
- Finally we verify the effectiveness of our system by comparing training effects of our system with other systems (Section 4.4).

4.1 Participants

30 L2 learners, including 15 high proficiency learners and 15 low proficiency learners are invited to participate two audio exaggeration experiments which are audio distinguishability experiment (Section 4.2.2) and understandability experiment (Section 4.2.3). The ages of 30 learners range from 20 to 32. Then 22 native English speakers together with 30 L2 learners in the previous experiments are invited to participate in the audio perceptibility experiment (Section 4.2.4). Among 22 native speakers, 10 are from South Africa, 5 are from the United States, 4 are from the United Kingdom and 3 are from Canada. The ages of the native speakers range from 24 to 42. All of the native speakers are certificated with TESOL (Teaching English to Speakers of Other Languages) issued by Ascentis, which is an officially recognized TESOL certificate authority. The 30 L2 learners from the previous experiments are further invited to participate in the visual perceptibility experiment (Section 4.3.2). Then, we invite 20 new L2 learners to take part in the supplementary experiments. 80 L2 learners, including 50 new L2 learners, and 30 learners in the previous experiment, are invited to participate in the *PTeacher* effectiveness verifying experiment (Section 4.4). The 80 learners contain 40 high proficiency learners and 40 low proficiency learners. The ages of 80 learners range from 20 to 32.

In each experiment, 20% of the participants are randomly selected to participate in face-to-face interviews based on the following criteria: (a) the ratio of low proficiency learners to high proficiency is 1:1; (b) the ratio of male to female is 1:1; (c) the participants are aged between 20 and 32. We choose 20 participants with an average age of 24. Grounded theory approach [16] is conducted with interview data analysis. All interview data are recorded by Google doc and processed by MaxQDA3 [69] for qualitative analysis. Through open coding, 100 codes [52] are produced. We collaboratively synthesized the interview content into higher-level themes through axial coding [52], including learning challenges, learning experiences, learning efficiency and learning effects. We also discussed the internal connections between these themes and generated an interviewer report.

4.2 Experiments on Audio Exaggeration

Audio exaggeration experiments are conducted to optimize the exaggeration ratios of the audio exaggeration generator in terms of distinguishability, understandability and perceptibility. We first synthesize the experimental audios with four exaggeration levels based on exaggerated speeches from a pronunciation training

expert. Then we test distinguishability, understandability and perceptibility for learners in the low proficiency and high proficiency groups w.r.t. each exaggeration level. Based on the results, we apply a non-linear fitting to determine the optimum exaggeration ratios for learners with low proficiency and high proficiency.

4.2.1 Material Preparation. First, a pronunciation training expert is asked to read 350 representative words from the Oxford Dictionary with four exaggeration levels: *zero*, *slight*, *medium* and *strong*, respectively. We calculate the exaggeration ratios of the pitch, duration and energy for each exaggerated phonemes. The expert then adjusts the exaggeration ratios. The exaggeration ratios for each exaggeration level w.r.t. different type of phonemes are shown in the supplementary material. Finally, we synthesize 800 speeches with different exaggeration level as test materials.

4.2.2 Experiment on Distinguishability of Exaggerated Speech. Distinguishability rate [50], which indicates whether a listener can easily distinguish the exaggerated phoneme in the exaggerated audio, is a crucial evaluation index to verify the effect of exaggerated expression. In our case, given a speech with one exaggerated phoneme, it evaluates whether the participants can discern the exaggerated phoneme from the speech. The distinguishability rate is defined as the accuracy of whether the user can discern the exaggerated phoneme.

The distribution and average of distinguishability rates are shown in Figure 6. It demonstrates that higher exaggeration level produces higher distinguishability rate. The average distinguishability rate increases from 87.11% to 98.89% for learners with high proficiency. The average distinguishability rate increases from 79.33% to 93.78% for learners with low proficiency. The result also indicates that learners with lower proficiency need audio with higher exaggeration level to discern the exaggerated part as easily as those with higher proficiency. We carry out single tail t-test between groups of the same proficiency with different exaggeration levels and different proficiency with different exaggeration levels. The P results mainly range from 0.0013 to 0.013, indicating significant differences. T-test result P between test groups with *medium* and *strong* exaggeration for low proficiency learners reaches 0.0678. The value is acceptable since it is close to 0.05.

The user interview also demonstrates the experiment result. A learner with low proficiency says “*Audio exaggeration helps me locate where I need to pay attention. It was quite easy for me to locate the exaggerated part for level 3 and level 4 exaggerated audios.*”

4.2.3 Experiment on Understandability of Exaggerated Speech. Understandability rate indicates whether the learner can still easily understand the exaggerated speech without confusion [9, 62, 63, 84]. Given two similar phoneme and a speech with one of the phonemes being exaggerated, participants are asked to choose which phoneme appears in the word. Similar to distinguishability rate, understandability rate is defined as the accuracy of whether the user can still recognize the exaggerated phoneme.

The distribution and average of the understandability rates are illustrated in Figure 7. Unlike the result of the experiment on distinguishability, a higher exaggeration level does not lead to the increase of understandability rates. For the learners with high proficiency, the optimum exaggeration level is *slight*, with the highest

understandability rate at 94.93%. For learners with low proficiency, the optimum exaggeration level is *medium*. The highest understandability rate is 90.72%. With the *strong* exaggeration, the average understandability rates drop to 85.07% and 86.38% for high and low proficiency learners, respectively. The reason is that *strong* exaggeration can cause distortion, which impedes learners from understanding it correctly, especially for learners with higher proficiency. We carry out single tail t-test between groups of the same proficiency with different exaggeration levels and different proficiency with different exaggeration levels. Most of the P-values range from 0.0017 to 0.0086, indicating a significant difference. The resulted P between the group with *medium* and *slight* exaggeration for high proficiency learner is 0.2480, indicating almost no difference. Also, P-value between the group with *medium* and *slight* exaggeration for low proficiency learner is 0.1198, indicating a minor difference. Based on the t-test results, we confirm that both *medium* and *slight* exaggeration is acceptable for learners with different proficiencies.

The participant interview further also demonstrates part of the experiment results. A learner with high proficiency comments that “*The voices break in some E4 exaggerated stops, and the phoneme ‘b’ sounds like ‘p’.*” Another learner with high proficiency comments that “*The ‘r’ sounds very strange in E4. I cannot tell you what it is.*” A learner with low proficiency comments that “*I cannot recognize the ‘th’ sound in audio with level two exaggeration, because ‘th’ and ‘sh’ are just too similar.*”

4.2.4 Experiment on Perceptibility of Exaggerated Speech. Perceptibility score, which evaluates exaggerated speech’s perceptual quality, is a significant indicator to check the intensity level [88] of perception from the perspective of hearing. Participants are asked to give opinion scores ranging from 0 to 5 (0 for too weak to perceive, 5 for too strong) in terms of perceptual exaggeration level. The perception scores are calculated with:

$$P = 2.5 - |2.5 - Score|, \quad (2)$$

where Score is the opinion score and P is the perception scores.

The distribution of perceptibility scores for different exaggeration level is illustrated in Figure 8. We find that the distribution is similar to that of the understandability rates. This result indicates that neither *slight* exaggeration nor *strong* exaggeration is good enough.

The participants’ interview further confirms our inference. A learner with low proficiency comments that: “*I felt that the ‘slight’ exaggerations are so slight sometimes that I have difficulties realizing them.*” A learner with high proficiency comments that: “*I don’t think the slight exaggeration is good enough since it is too weak. Also, I found severe distortions in the ‘strong exaggeration’ version. So I don’t think it is good enough, either.*” A native speaker comments: “*I think the medium exaggeration version is very cool. I must have used a similar exaggeration method in my class.*”

4.2.5 Optimizing Exaggeration Ratios with Non-linear Fitting. The experiments on distinguishability rates, understandability rates, and perceptibility scores apply a non-linear fitting to find the optimum exaggeration ratios w.r.t. different phoneme types. The optimum

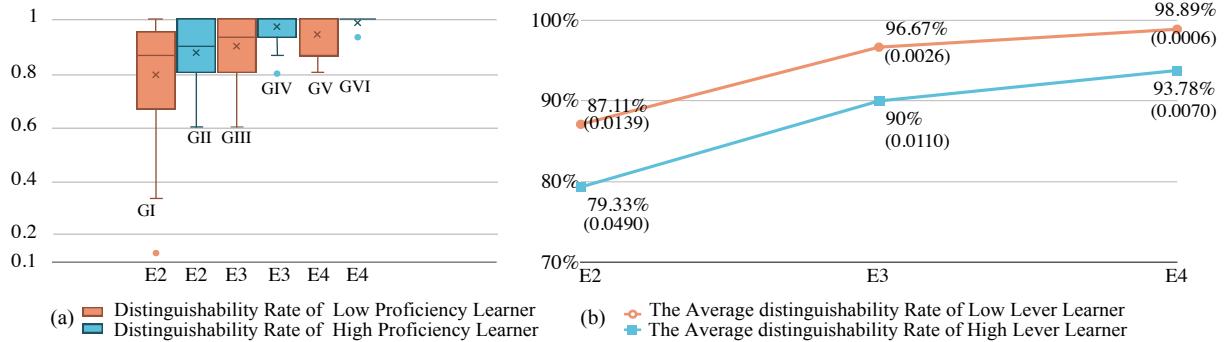


Figure 6: The distribution and average of the distinguishability rates are shown in figure (a) and (b), respectively. The variances of the distinguishability rates is in the brackets in figure (b). E1, E2, E3 and E4 represent zero, slight, medium and strong audio exaggeration respectively.

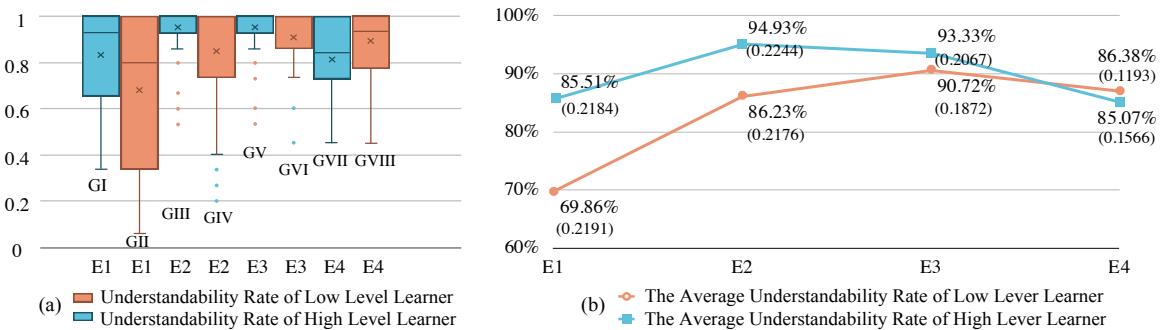


Figure 7: The distribution and average understandability rates are illustrated in figure (a) and (b), respectively. The variance of the understandability rates is in the brackets in figure (b).

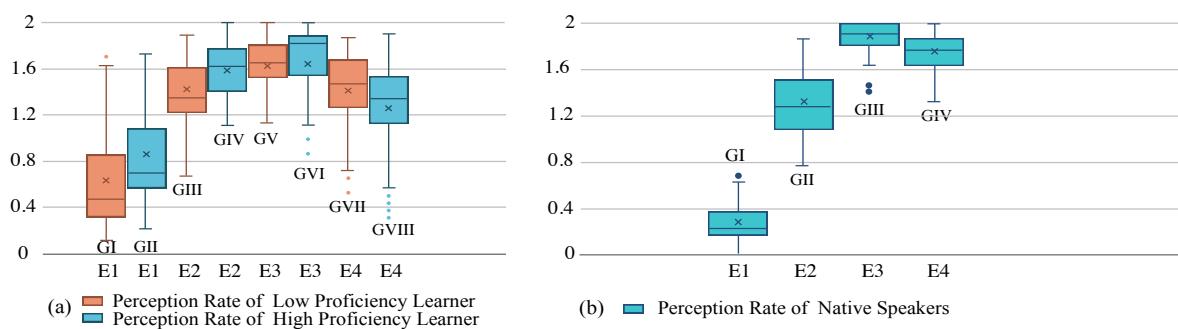


Figure 8: The perceptibility rates distribution w.r.t. four exaggerations for the learner with different proficiency are depicted in figure (a). The distribution of perceptibility rates assessed by the teacher is shown in the figure (b).

exaggeration ratios can achieve the highest value defined with:

$$V = DR + UR + \frac{PS}{2.5}, \quad (3)$$

where V is the optimizing target for non-linear fitting, DR , UR and PS are distinguishability rates, understandability rates and perceptibility scores respectively.

The result is shown in Figure 9. Almost all the best exaggeration ratios fall into the range of medium exaggeration. We notice that the exaggeration ratios needed for high proficiency learners are lower than those low proficiency learners. The optimum exaggeration ratios are used in the following experiments.

4.3 Experiments on Visual Exaggeration

Visual exaggeration experiments are conducted to optimize the articulatory movement exaggeration level and tongue color exaggeration level in terms of perceptibility. We first synthesize the experiment video with different articulatory movements exaggeration level and tongue color exaggeration level. Then we test perceptibility for learners in the low proficiency and high proficiency groups w.r.t. different exaggeration level. Based on the results, we determine the optimum articulatory movement exaggeration level and tongue color exaggeration level for the learners with low proficiency and high proficiency.

4.3.1 Material Preparation for Visual Exaggeration. Four exaggeration levels of articulatory movement, which are named *zero*, *slight*, *medium*, and *strong*, are manually designed. Three exaggeration level of tongue colors, which are named *zero*, *medium*, and *strong* respectively, are also manually designed. We synthesize the animations of 52 words, equally covering the 13 phoneme types, with different exaggeration levels w.r.t. articulatory movement and tongue color. An example of the animations is shown in Figure 10.

4.3.2 Experiment on Perceptibility of Exaggerated Animations. We also conduct perceptibility experiment on animations with visual exaggeration. We directly take the definition of opinion score and perceptibility score in the experiment on audio exaggeration (Section 4.2.4). Participants are first asked to give opinion scores on animations with zero tongue color exaggeration and different articulatory movement exaggeration. Then, we find optimum articulatory movement exaggeration with the highest perceptibility score. Following that, participants are asked to give opinion scores on animations with optimized articulatory movement exaggeration and different exaggeration. Finally, we find optimum tongue color exaggeration with the highest perceptibility score.

The distribution and average of perceptibility scores on different articulatory movement exaggeration and tongue color exaggeration are shown in Figure 11. We find that a higher articulatory movement exaggeration level or tongue color exaggeration level does not lead to the increase of understandability rates. The optimum articulatory movements exaggeration level is a *slight* exaggeration for learners with high proficiency, *strong* exaggeration for learners with low proficiency. The perceptibility scores are 2 and 1.96, respectively. The optimum tongue color exaggeration is *slight* for learners with high proficiency. They obtain a perceptibility score of 2. While the optimum tongue color exaggeration for learners with low proficiency is a *strong* exaggeration. The perceptibility score is 1.99. We carry out single tail t-test between groups of the same proficiency with different exaggeration levels and different proficiency with different exaggeration levels. Most of the P-values range from 0.00011 to 0.037, indicating a significant difference. The t-test result P between the group with *medium* and *strong* articulatory movement exaggeration for low proficiency learner is 0.1466,

indicating almost no difference. Also, P-value between the group with slight and strong tongue color exaggeration for high proficiency learner is 0.1378, indicating a slight difference. Based on the t-test result, we find that both *strong* and *medium* articulatory exaggeration is acceptable for learners with low proficiency. We also see that both *slight* and *strong* articulatory exaggeration is acceptable for learners with high proficiency. Still, we choose the optimum articulatory movement exaggeration and tongue color movement mentioned above as the visual exaggeration setting in the following experiments.

The participant interview further confirms our conclusion. A learner with low proficiency comments: “*I cannot see any difference with the low exaggeration level on articulatory movement. Same to the tongue color exaggeration. It is not very eye-catching.*” A learner with high proficiency comments: “*The high-level exaggeration may be too much for me.*”

4.4 Experiment on Effectiveness of PTeacher

To determine if the personalized feedback is effective, we compare our system to 3 other systems. The first system is set with no exaggeration. The second system is set with no personalization, which means the exaggeration ratios are fixed to the average of the exaggeration ratios for low proficiency and high proficiency. The third system uses feedback from pronunciation training experts. Participants are equally divided into four groups and are asked to test their pronunciation accuracy before and after one-hour training with different systems. Their pronunciation accuracy is annotated by pronunciation training experts with percentile. The improvement rate of learners is calculated with:

$$I = \frac{S_{\text{after}} - S_{\text{before}}}{S_{\text{before}}} \quad (4)$$

where I is the improvement rate, S_{before} and S_{after} is pronunciation accuracy of the learner before and after training respectively.

The result is shown in Figure 12. We find that the improvement rate of learners who are trained with *PTeacher* is much higher than those who are trained with the non-exaggeration system or the non-personalizing system is comparable to those who are trained with the system with human feedback. We carry out t-test between different groups. Most of the P-values range from 0.003 to 0.015, also indicates a big or enormous difference between these groups. The mean differences in improvement rate between learner trained with *PTeacher* and human feedback system is less than 2.5%. The P results of the t-test between them are 0.29 and 0.42 for learners with high and low proficiency, indicating a minimal difference. The result confirms that the effectiveness of exaggerated audiovisual feedback and personalization mechanism.

Participant interview is conducted during the experiments mentioned above. The feeling of participants about the system is asked and recorded. More than 85% of the learners and 80% of the teachers praise the audio exaggeration. A native speaker says: “*As an English teacher, I think exaggerated audio feedback is useful. It reminds me of how I teach a learner to say a word right in class.*” A learner with high proficiency tells us: “*The audio feedback reminds me of my high school English teacher. She would rectify our mispronunciations when we made mistakes. This method is quite useful for me.*” A learner with low proficiency also praises the mechanism: “*I am so happy with*

Phone Type	Leaner's Level	Pitch	Duration	Energy	Phone Type	Leaner's Level	Pitch	Duration	Energy
Voiceless Stops	Low Proficiency	1.26	1.13	4.22	Nasals	Low Proficiency	1.32	3.42	1.60
	High Proficiency	1.21	1.11	3.84		High Proficiency	1.25	3.03	1.52
Voiced Stops	Low Proficiency	1.26	1.49	5.29	Laterals	Low Proficiency	1.12	2.78	3.45
	High Proficiency	1.20	1.30	4.59		High Proficiency	1.09	2.37	2.88
Voiceless Fricatives	Low Proficiency	1.22	2.76	3.78	Retroflexes	Low Proficiency	1.13	2.43	2.04
	High Proficiency	1.15	2.27	3.14		High Proficiency	1.10	2.11	1.85
Voiced Fricatives	Low Proficiency	1.51	1.64	4.63	Semivowels	Low Proficiency	1.24	2.16	2.57
	High Proficiency	1.38	1.52	3.90		High Proficiency	1.19	1.93	2.17
Voiceless Affricates	Low Proficiency	1.19	1.19	3.58	Long Vowels	Low Proficiency	1.14	1.96	2.82
	High Proficiency	1.11	1.11	2.12		High Proficiency	1.11	1.76	2.16
Voiced Affricates	Low Proficiency	1.70	1.87	4.64	Short Vowels	Low Proficiency	1.38	2.15	2.15
	High Proficiency	1.49	1.55	3.87		High Proficiency	1.30	2.01	2.01
Diphthongs	Low Proficiency	1.10	2.11	2.00	Diphthongs	High Proficiency	1.07	1.74	1.68

Figure 9: The optimum audio exaggeration ratios are shown in the figure. We apply non-linear fitting to find the exaggeration ratios with the highest distinguishability rates, understandability rates and perceptibility rates.

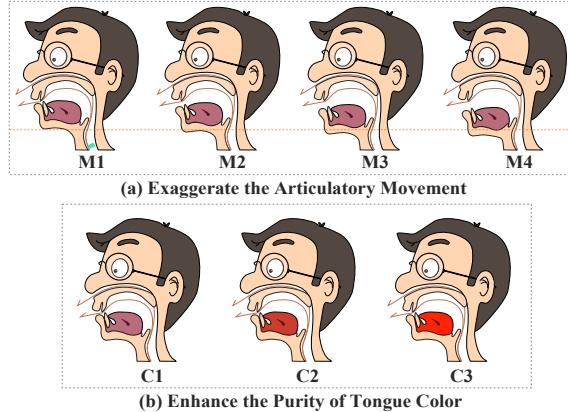


Figure 10: The figure illustrates examples of animations with exaggeration. M1, M2, M3 and M4 represent zero, slight, medium and strong articulatory movement exaggeration, respectively. C1, C2 and C3 represent zero, slight and strong tongue color exaggeration, respectively.

this audio exaggeration. It points out the mistake that I can hardly notice."

More than 70% of the learners mention visual exaggeration. A learner with low proficiency says: "Exaggerated visual feedback is helpful to me. I can imitate the animation and rectify my pronunciation. I do hope that you can try to combine virtual reality technology with visual feedback. That will be even more helpful to me."

About 75% of the learners say that the interactive course can raise their learning interest. A learner with low proficiency says "The interactive course is fascinating. I am willing to spend more time rectify my pronunciation with it. Please design more courses in the future."

It is also worth noticing that about 35% of the learners mention that our system helps deal with educational inequity. A learner with low proficiency tells us "Before I went to college, I had been living in a small town, where the education resource is limited. I didn't have

any chance to get a foreign teacher to teach me. Thus, my English pronunciation training is relatively inadequate. With the PTeacher, I can learn the pronunciation effectively anytime, anywhere. I can't be more willful to introduce it to the children in my town."

5 DISCUSSION

5.1 Contributions to HCI

Based on these studies and the previous learning theories on exaggerated feedback as discussed in Section 2, we focus on constructing a participatory exaggerated computer-aided pronunciation training system, *PTeacher*. It emphasizes enlarging the teaching effectiveness from a user perception perspective with two critical aspects: 1) we study how to identify suitable exaggerated feedback to learners with different demands or behaviors in language learning (i.e. different pronunciation proficiencies), and 2) how to define the best set of feedback. Our idea of involving exaggerated feedback and our system of determining the best set of exaggeration parameters would be an important finding to the general area of computer-aided language learning [25, 29, 66, 77] and exaggerated feedback [28] systems in HCI.

We further deepen the discussion to general educational systems. We point out that such an idea can be easily migrated to not only other language learning systems but also any imitation-learning system where certain degrees of exaggeration in educational feedback would enhance learners' perceptibility. For example, under dancing and piano teaching scenarios, we can exaggerate the teaching effect such as increase the music's expression for piano or the expressiveness of dance animation. The method for how to exaggerate the audio or visual modality can give us much inspiration. However, maybe it is not applicable to logical induction systems such as Math teaching.

5.2 Limitations

The Limitations of MDD. Though the accuracy of existing MDD is already very high according to [47–49], the feedback is still likely to contain two kinds of errors [2]: false accepts (FA) which means that the pronunciation is accepted although it is actually incorrect;

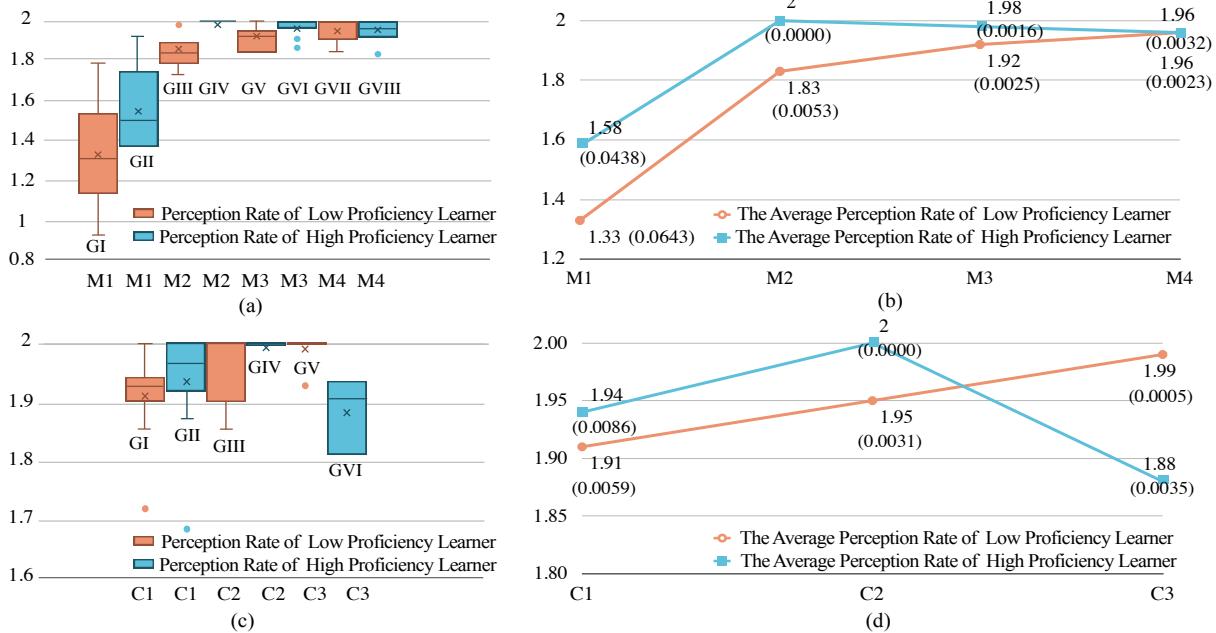


Figure 11: The distribution and average of the perception rates for different articulatory movement exaggerations are shown in figure (a) and figure (b), respectively. The distribution and average of the perception rate for different tongue color exaggerations are shown in figure (c) and figure (d), respectively.

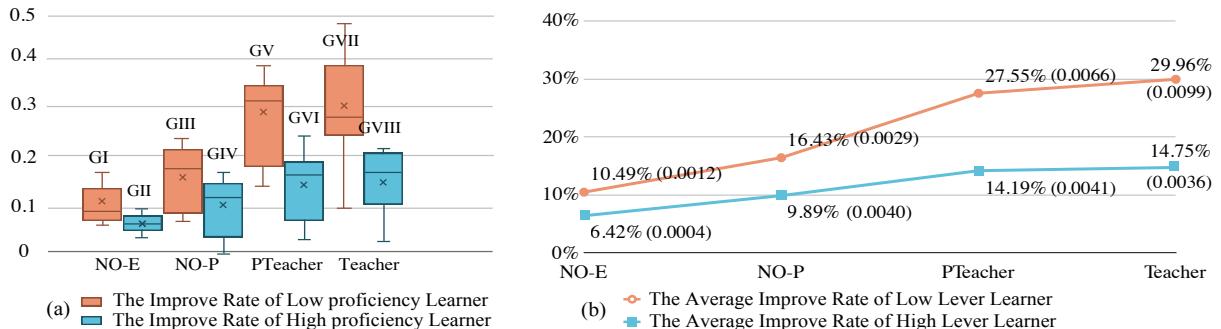


Figure 12: The distribution and the average of the improvement rates are illustrated in figure (a) and (b), respectively. The NO-E, NO-P, PTeacher and Teacher represent the system (1) without exaggeration feedback (2) without personalizing (3) personalized exaggeration feedback (4) pronunciation training experts, respectively. PTeacher is almost as efficient as teacher-aid training, far exceeding systems without personalized exaggeration feedback.

and false rejects (FR) means that the pronunciation is rejected although actually correct. As a result, the *MDD* may miss detect mistakenly spoken phonemes or mark correct ones as incorrect, which will affect our system.

There are two traditional mechanisms of pronunciation feedback. One is that the system directly tells the user which sound is mispronounced based on the *MDD* results. The other is to provide an *MDD* score. PTeacher will not directly tell the user which sound is mispronounced but presents the mispronounced phoneme by exaggeration. As a result, anytime the *MDD* makes a mistake of FR,

the learners will receive an exaggerated feedback, or the system will not exaggerate the target mistakenly pronounced word if it is a FA. So the negative impact is mostly that the learner cannot receive the exaggeration in the system's feedback.

The Limitations of PTeacher. The levels of exaggeration are defined mainly in a hand-crafted manner by consulting with English pronunciation education experts, animation designers then leveraging the theory of speech articulators [5, 11, 38, 41, 81]. As a result, on the one hand, the designing procedures for exaggerations are time-consuming. On the other hand, the manually defined

exaggeration levels can only be limited to a relatively small scale. Moreover, they are not continuously changeable. In the future, we can derive automatic procedures with the help of deep learning technologies [90, 91] from the collected data. As for the methodology, only 2 levels of proficiency are defined at phone-level. The personalized feedback can be improved by involving more detailed modelling on the feedback levels of proficiency.

6 CONCLUSION

In this paper, we present *PTeacher*, a pronunciation training system with personalized exaggerated audio-visual corrective feedback and practical training courses. Importantly, we uncover how to define the appropriate degree of exaggerations through extensive user-participated experiments. The optimum set of exaggerations can thus be identified for each individual learner. Moreover, interactive training courses are proven to be efficient in improving users' English proficiency.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China under Grant No. 2020AAA0108600 and No. 2019YFB1405700, the state key program of the National Natural Science Foundation of China (NSFC) (No.61831022) and Tiangong Institute for Intelligent Computing, Tsinghua University. We would like to thank Professor Helen Meng, Associate Research Fellow Chun Yu and Jingbei Li.

REFERENCES

- [1] Najwa Alghamdi, Steve Maddock, Jon Barker, and Guy J Brown. 2017. The impact of automatic exaggeration of the visual articulatory features of a talker on the intelligibility of spectrally distorted speech. *Speech Communication* 95 (2017), 127–136.
- [2] Pierre Badin, Atef Ben Youssef, Gérard Bailly, Frédéric Elisei, and Thomas Hueber. 2010. Visual articulatory feedback for phonetic correction in second language learning. In *Second Language Studies: Acquisition, Learning, Education and Technology*.
- [3] Heather Bliss, Jennifer Abel, and Bryan Gick. 2018. Computer-assisted visual articulation feedback in L2 pronunciation instruction: A review. *Journal of Second Language Pronunciation* 4, 1 (2018), 129–153.
- [4] Ann R Bradlow, David B Pisoni, Reiko Akahane-Yamada, and Yoh'ichi Tohkura. 1997. Training Japanese listeners to identify English/r/and/l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America* 101, 4 (1997), 2299–2310.
- [5] Catherine P Brown and Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49, 3-4 (1992), 155–180.
- [6] Matthew I Brown and Avi E Cieplinski. 2020. Device, method, and graphical user interface for providing audiovisual feedback. US Patent 10,599,394.
- [7] Yaohua Bu, Jia Jia, Xiang Li, Suping Zhou, and Xiaobo Lu. 2018. IcooBook: when the picture book for children encounters aesthetics of interaction. In *Proceedings of the 26th ACM international conference on Multimedia*. 1260–1262.
- [8] Yaohua Bu, Weijun Li, Tianyi Ma, Shengqi Chen, Jia Jia, Kun Li, and Xiaobo Lu. 2020. Visual-speech Synthesis of Exaggerated Corrective Feedback. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4521–4523.
- [9] Eva Cerviño-Povedano and Joan C Mora. 2010. Investigating Catalan learners of English over-reliance on duration: Vowel cue weighting and phonological short-term memory. *Achievements and perspectives in the acquisition of second language speech: New Sounds* (2010), 53–64.
- [10] Pierre Chalfouf and Claude Frasson. 2011. Subliminal cues while teaching: HCI technique for enhanced learning. *Advances in Human-Computer Interaction* 2011 (2011).
- [11] Bay-Wei Chang and David Ungar. 1993. Animation: from cartoons to the user interface. In *Proceedings of the 6th annual ACM symposium on User interface software and technology*. 45–55.
- [12] Tsuhan Chen and Ram R Rao. 1998. Audio-visual integration in multimodal communication. *Proc. IEEE* 86, 5 (1998), 837–852.
- [13] Bing Cheng, Xiaojuan Zhang, Siying Fan, and Yang Zhang. 2019. The role of temporal acoustic exaggeration in high variability phonetic training: A behavioral and ERP study. *Frontiers in psychology* 10 (2019), 1178.
- [14] Bing Cheng, Xiaojuan Zhang, and Yang Zhang. 2019. Temporal exaggeration facilitates second language phonetic training: The case of syllable-final nasal contrast. *The Journal of the Acoustical Society of America* 146, 4 (2019), 2844–2844.
- [15] Laura Colantoni, Jeffrey Steele, Paola Escudero, and Paola Rocio Escudero Neyra. 2015. *Second language speech*. Cambridge University Press.
- [16] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.
- [17] Nuria Calvo Cortés. 2005. Negative language transfer when learning Spanish as a foreign language. *Interlingüística* 16 (2005), 237–248.
- [18] British Council. 2013. The English Effect. *Retrieved March 22 (2013)*, 2015.
- [19] David Crystal. 2011. *A dictionary of linguistics and phonetics*. Vol. 30. John Wiley & Sons.
- [20] Tracey M Derwing and Murray J Munro. 2005. Second language accent and pronunciation teaching: A research-based approach. *TESOL quarterly* 39, 3 (2005), 379–397.
- [21] Tracey M Derwing and Marian J Rossiter. 2002. ESL learners' perceptions of their pronunciation needs and strategies. *System* 30, 2 (2002), 155–166.
- [22] Paola Escudero. 2001. The role of the input in the development of L1 and L2 sound contrasts: language-specific cue weighting for vowels. In *Proceedings of the 25th annual Boston University conference on language development*, Vol. 1. Citeseer, 250–261.
- [23] Paola Rocio Escudero Neyra. 2005. *Linguistic perception and second language acquisition: explaining the attainment of optimal phonological categorization*. Ph.D. Dissertation. Utrecht University & LOT.
- [24] Tony Ezzat and Tomaso Poggio. 2000. Visual speech synthesis by morphing visemes. *International Journal of Computer Vision* 38, 1 (2000), 45–57.
- [25] Christina Garcia, Mark Kolat, and Terrell A Morgan. 2018. SELF-CORRECTION OF SECOND-LANGUAGE PRONUNCIATION VIA ONLINE, REAL-TIME, VISUAL FEEDBACK. In *PRONUNCIATION IN SECOND LANGUAGE LEARNING AND TEACHING CONFERENCE (ISSN 2380-9566)*. 54.
- [26] Patrick H Geoghegan, C Spence, Wei H Ho, X Lu, M Jermy, P Hunter, and J Cater. 2012. Stereoscopic PIV measurement of airflow in human speech during pronunciation of fricatives. In *16th International Symposium of Laser Techniques to Fluid Mechanics, Lisbon, Portugal, 9th-12th July*.
- [27] Ewa M Golonka, Anita R Bowles, Victor M Frank, Dorna L Richardson, and Suzanne Freynik. 2014. Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer assisted language learning* 27, 1 (2014), 70–105.
- [28] Antti Granqvist, Tapio Takala, Jari Takatalo, and Perttu Hämäläinen. 2018. Exaggeration of Avatar Flexibility in Virtual Reality. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*. 201–209.
- [29] Joshua Hailpern, Karrie Karahalios, and James Halle. 2009. Creating a spoken impact: encouraging vocalization through audio visual feedback in children with ASD. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–462.
- [30] Morris Halle, Bert Vaux, and Andrew Wolfe. 2000. On feature spreading and the representation of place of articulation. *Linguistic inquiry* 31, 3 (2000), 387–444.
- [31] CC Hsu. [n.d.]. Python-wrapper-for-world-vocoder.
- [32] Philip Hubbard. 2002. Interactive participatory dramas for language learning. *Simulation & Gaming* 33, 2 (2002), 210–216.
- [33] Yurie Iribe, Silasak Manosavanh, Kouichi Katsurada, Ryoko Hayashi, Chunyue Zhu, and Tsuneo Nitta. 2011. Generating animated pronunciation from speech through articulatory feature extraction. In *Twelfth Annual Conference of the International Speech Communication Association*.
- [34] D Kalikow and J Swets. 1972. Experiments with computer-controlled displays in second-language learning. *IEEE Transactions on Audio and Electroacoustics* 20, 1 (1972), 23–28.
- [35] Natalia Kartushina and Ulrich H Frauenfelder. 2014. On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation. *Frontiers in psychology* 5 (2014), 1246.
- [36] Natalia Kartushina, Alexis Hervais-Adelman, Ulrich Hans Frauenfelder, and Narly Golestan. 2015. The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The journal of the acoustical society of America* 138, 2 (2015), 817–832.
- [37] Tatsuya Kawahara, Masatake Dantsuji, and Yasushi Tsubota. 2004. Practical use of English pronunciation system for Japanese students in the CALL classroom. In *Eighth International Conference on Spoken Language Processing*.
- [38] Gerald Kelly. 2006. *How To Teach Pronunciation (With Cd)*. Pearson Education India.
- [39] P Khul, K Williams, F Lacerda, and K Lindblom Stevens. [n.d.]. B.(1992). Linguistic Experience Alters Phonetic Perception in Infants by 6 Months of Age. *Science* 255 ([n. d.]).
- [40] AJ King and AR Palmer. 1985. Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental brain research* 60, 3 (1985), 492–500.
- [41] Valeri Aleksandrovich Kozhevnikov and Liudmila Andreevna Chistovich. 1967. *Speech: articulation and perception*. Vol. 30. US Department of Commerce, Clearinghouse for Federal Scientific and

- [42] John Lasseter. 1987. Principles of traditional animation applied to 3D computer animation. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*. 35–44.
- [43] Andrew H Lee and Roy Lyster. 2016. The effects of corrective feedback on instructed L2 speech perception. *Studies in Second Language Acquisition* 38, 1 (2016), 35.
- [44] Bradford Lee, Luke Plonsky, and Kazuya Saito. 2020. The effects of perception-vs-production-based pronunciation instruction. *System* 88 (2020), 102185.
- [45] Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8132–8136.
- [46] Wai-Kim Leung, Ka-Wa Yuen, Ka-Ho Wong, and Helen Meng. 2013. Development of text-to-audiovisual speech synthesis to support interactive language learning on a mobile device. In *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 583–588.
- [47] Kun Li, Jing Li, Yufang Song, and Hewei Fu. 2015. Rating Algorithm for Pronunciation of English Based on Audio Feature Pattern Matching. In *MATEC Web of Conferences*, Vol. 22. EDP Sciences, 01032.
- [48] Kun Li, Xiaojun Qian, Shiyin Kang, Pengfei Liu, and Helen Meng. 2015. Integrating acoustic and state-transition models for free phone recognition in L2 English speech using multi-distribution deep neural networks.. In *SLaTE*. 119–124.
- [49] Kun Li, Xiaojun Qian, and Helen Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 1 (2016), 193–207.
- [50] Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. 1957. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology* 54, 5 (1957), 358.
- [51] Patsy M Lightbrown and Nina Spada. 2000. Do they know what they're doing? L2 learners' awareness of L1 influence. *Language Awareness* 9, 4 (2000), 198–217.
- [52] Guanhong Liu, Xianghua Ding, Chun Yu, Lan Gao, Xingyu Chi, and Yuanchun Shi. 2019. "I Bought This for Me to Look More Ordinary" A Study of Blind People Doing Online Shopping. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [53] Pengfei Liu, Ka-Wa Yuen, Wai-Kim Leung, and Helen Meng. 2012. menunciate: Development of a computer-aided pronunciation training system on a cross-platform framework for mobile, speech-enabled application development. In *2012 8th International Symposium on Chinese Spoken Language Processing*. IEEE, 170–173.
- [54] Jingli Lu, Ruili Wang, and Liyanage C De Silva. 2012. Automatic stress exaggeration by prosody modification to assist language learners perceive sentence stress. *International journal of speech technology* 15, 2 (2012), 87–98.
- [55] Jingli Lu, Ruili Wang, Liyanage C De Silva, Yang Gao, and Jia Liu. 2010. CASTLE: a computer-assisted stress teaching and learning environment for learners of English as second language. In *Eleventh Annual Conference of the International Speech Communication Association*.
- [56] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.. In *Interspeech*, Vol. 2017. 498–502.
- [57] Fanbo Meng, Helen Meng, Zhiyong Wu, and Lianhong Cai. 2010. Synthesizing expressive speech to convey focus using a perturbation model for computer-aided pronunciation training. In *Second Language Studies: Acquisition, Learning, Education and Technology*.
- [58] Fanbo Meng, Zhiyong Wu, Jia Jia, Helen Meng, and Lianhong Cai. 2014. Synthesizing English emphatic speech for multimodal corrective feedback in computer-aided pronunciation training. *Multimedia tools and applications* 73, 1 (2014), 463–489.
- [59] Fanbo Meng, Zhiyong Wu, Helen Meng, Jia Jia, and Lianhong Cai. 2012. Hierarchical English emphatic speech synthesis based on HMM with limited training data. In *Thirteenth Annual Conference of the International Speech Communication Association*.
- [60] Helen Meng, Yuen Yee Lo, Lan Wang, and Wing Yiu Lau. 2007. Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*. IEEE, 437–442.
- [61] Richard I Miller. 1990. *Major American Higher Education Issues and Challenges in the 1990s*. Higher Education Policy Series 9. ERIC.
- [62] Joan C Mora and Isabelle Darcy. 2017. The relationship between cognitive control and pronunciation in a second language. *Second language pronunciation assessment* (2017), 95.
- [63] Murray J Munro, Tracey M Derwing, and James E Flege. 1999. Canadians in Alabama: A perceptual study of dialect acquisition in adults. *Journal of Phonetics* 27, 4 (1999), 385–403.
- [64] Ambra Neri, Catia Cucchiarini, and Helmer Strik. 2006. ASR corrective feedback on pronunciation: Does it really work? (2006).
- [65] Ambra Neri, Catia Cucchiarini, Helmer Strik, and Lou Boves. 2002. The pedagogy-technology interface in computer assisted pronunciation training. *Computer assisted language learning* 15, 5 (2002), 441–467.
- [66] Ambra Neri, Ornella Mich, Matteo Gerosa, and Diego Giuliani. 2008. The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning* 21, 5 (2008), 393–408.
- [67] Yishuang Ning, Zhiyong Wu, Jia Jia, Fanbo Meng, Helen Meng, and Lianhong Cai. 2015. HMM-based emphatic speech synthesis for corrective feedback in computer-aided pronunciation training. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4934–4938.
- [68] Richard Ogden. 2017. *Introduction to English Phonetics*. Edinburgh university press.
- [69] Mirian Oliveira, Claudia Bitencourt, Eduardo Teixeira, and Ana Clarissa Santos. 2013. Thematic content analysis: Is there a difference between the support provided by the MAXQDA® and NVivo® software packages. In *Proceedings of the 12th European Conference on Research Methods for Business and Management Studies*. 304–314.
- [70] Marta Ortega and Valerie Hazan. 1999. Enhancing acoustic cues to aid L2 speech perception. In *Proceedings of the International Congress of Phonetics Sciences*. 117–120.
- [71] Martha C Pennington. 1999. Computer-aided pronunciation pedagogy: Promise, limitations, directions. *Computer Assisted Language Learning* 12, 5 (1999), 427–440.
- [72] Janet Breckenridge Pierrehumbert. 1980. *The phonology and phonetics of English intonation*. Ph.D. Dissertation. Massachusetts Institute of Technology.
- [73] Linda Polka and Janet F Werker. 1994. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human perception and performance* 20, 2 (1994), 421.
- [74] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*. 3171–3180.
- [75] Tiago Ribeiro and Ana Paiva. 2012. The illusion of robotic life: principles and practices of animation for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 383–390.
- [76] Ellen Ricard. 1986. Beyond Fossilization: A Course in Strategies and Techniques in Pronunciation for Advanced Adult Learners. *TESL Canada Journal* (1986), 243–253.
- [77] 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*. 1–13.
- [78] Pamela Rogerson-Revell. 2011. *English phonology and pronunciation teaching*. Bloomsbury Publishing.
- [79] Winifred Strange. 1995. Speech perception and linguistic experience: Theoretical and methodological issues.
- [80] Winifred Strange, Valerie L Shafer, et al. 2008. Speech perception in second language learners: The re-education of selective perception. *Phonology and second language acquisition* 36 (2008), 153–192.
- [81] Frank Thomas, Ollie Johnston, and Frank Thomas. 1995. *The illusion of life: Disney animation*. Hyperion New York.
- [82] Ingo R Titze and Daniel W Martin. 1998. Principles of voice production.
- [83] Nikolai Sergeevich Trubetzkoy. 1969. Principles of phonology. (1969).
- [84] Ganna Veselovska. 2016. Teaching elements of English RP connected speech and CALL: Phonemic assimilation. *Education and Information Technologies* 21, 5 (2016), 1387–1400.
- [85] Amy B Wöhrlert and Vicki L Hammen. 2000. Lip muscle activity related to speech rate and loudness. *Journal of Speech, Language, and Hearing Research* 43, 5 (2000), 1229–1239.
- [86] Ka-Ho Wong, Wai-Kim Leung, Wai-Kit Lo, and Helen Meng. 2010. Development of an articulatory visual-speech synthesizer to support language learning. In *2010 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 139–143.
- [87] Ka-Wa Yuen, Wai-Kim Leung, Peng-fei Liu, Ka-Ho Wong, Xiao-jun Qian, Wai-Kit Lo, and Helen Meng. 2011. Enunciate: An internet-accessible computer-aided pronunciation training system and related user evaluations. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*. IEEE, 85–90.
- [88] Fan-Gang Zeng, Kristina M Martino, Fred H Linthicum, and Sigfrid D Soli. 2000. Auditory perception in vestibular neurectomy subjects. *Hearing research* 142, 1-2 (2000), 102–112.
- [89] Junhong Zhao, Hua Yuan, Wai-Kim Leung, Helen Meng, Jia Liu, and Shanrong Xia. 2013. Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8218–8222.
- [90] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. In *AAAI Conference on Artificial Intelligence (AAAI)*.
- [91] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeItTalk: Speaker-Aware Talking-Head Animation. *ACM Transactions on Graphics* 39, 6 (2020).