MOBILE-ASSISTED PRONUNCIATION TRAINING WITH ADULT ESOL LEARNERS: BACKGROUND, ACCEPTANCE, EFFORT, AND ACCURACY

Kevin Hirschi, Northern Arizona University
Okim Kang, Northern Arizona University
John Hansen, University of Texas at Dallas
Catia Cucchiarini, Radboud University
Helmer Strik, Radboud University

This study investigates the relationships of learner background variables of adult English for Speakers of Other Languages (ESOL) learners and a mobile App designed to promote pronunciation skills targeting features known to contribute to intelligibility. Recruited from free evening classes for English learners, 34 adult ESOL learners of mixed ESOL learning experiences, ages, lengths of residency, and first languages (L1s) completed six phoneme pair lessons on a mobile App along with a background questionnaire and technology acceptance survey (Venkatesh *et al*., 2012). A series of Linear Mixed-Effect Model (LMEM) analyses were performed on learner background variables, technology acceptance, learner effort, and accuracy. The results found a minimal relationship between age, technology acceptance, and effort (7.68%) but a moderate to large relationship between age, technology acceptance and accuracy of consonants (39.70%) and vowels (64.26%). The implications are that learner use of mobile devices for L2 pronunciation training is moderated by various learner-related factors and the findings offer supportive evidence for designing mobile-based applications for a wide variety of backgrounds.

Cite as: Hirschi, K., Kang, O., Hansen, J., Cucchiarini, C., & Strik, H. (2022). Mobile-assisted pronunciation training with adult ESOL learners: background, acceptance, effort, and accuracy. In J. Levis & A. Guskaroska (eds.). *Proceedings of the 12th Pronunciation in Second Language Learning and Teaching Conference*, held June 2021 virtually at Brock University, St. Catharines, ON. https://doi.org/10.31274/psllt.13272.

INTRODUCTION

Kaiser (2018) posits that Mobile-Assisted Pronunciation Training (MAPT) is fundamentally different than other language learning technologies not only because of the mobility afforded by the device, but also because of the cost, ability to use audio and video features, and the communicative purpose of the device. Additionally, modern devices are equipped with increasingly powerful Automatic Speech Recognition (ASR) technology which can provide beneficial immediate feedback on learner speech (Neri, *et al*., 2003). Such feedback serves as not only evidence for the learner to increase target-like production, but if implemented strategically and appropriately, it can also increase motivation and orientation towards pronunciation practice (Suzukida, 2021). However, it is critical that technological considerations not influence intervention design, but that empirically informed pedagogy should be supported by appropriate technology (Neri *et al*., 2002).

With improvements in ASR ability to recognize learner speech (McCrocklin *et al*., 2019), learner use of ASR in low-stakes repetition of important target forms can positively affect learner pronunciation (Chen *et al*., 2020; Park, 2017). Investigations of user activity with ASR-based practice (i.e., the number of task retries) can also serve as a variable of interest as it provides concrete measurements of the learning process (Tejedor-García *et al*., 2016). Such learner differences may play a key role in understanding learner willingness to practice and sustain engagement with ASR. To this end, the present study investigates the role of learner background variables on learner effort and accuracy when completing listening and ASR-based practice tasks. It reports on the creation and use of an intelligibility-based supplemental pronunciation course for Adult ESOL learners using the Novo Play app (Novo Learning BV, 2019), and analyzes the data captured for insights into the relationships of learner background and app interaction.

Learner Background and Pronunciation Training

Despite the growing body of research supporting pronunciation training for many types of learners, few studies have included learner background variables in pronunciation training. We can therefore turn to meta-analytic results for insights into learner background and effects of instruction. For technology-based lab and classroom pronunciation instruction, the institutional setting of high school indicated a higher effect of instruction ($k = 7$, $d = 1.19$) than did those at the university ($k = 46$, $d = 0.77$) (Lee *et al*., 2015). However, the opposite trend was found in studies only concerned with technology-based training. In this case, learners under 17 years old showed lower gains ($k = 6$, $d = 0.46$) than those over 18 ($k = 10$, $d = 0.57$) (Mahdi & Al-Khateeb, 2019). While these findings might confound age with proficiency level (i.e., the level of the high school learners may have been lower than that of the university students), it outlines the possibility that age is a moderator of pronunciation instruction and it may play a differential role in technology-based pronunciation interventions.

Attitudinal factors are also likely to impact the effect of technology-based instruction because of the often optional and independent nature of learner-technology interactions. In technology acceptance research within the Unified Theory of Acceptance and Use of Technology (UTAUT) has served as a model of technology acceptance that researchers have employed in a variety of domains including language learning (Venkatesh *et al*., 2012). The UTAUT is comprised of several constructs, including *performance expectancy, effort expectancy, social influence, facilitating conditions,* and *behavioral intention* which are moderated by background variables such as age, experience, and gender. Ho *et al*. (2010) used the UTAUT to measure technology acceptance when implementing a podcasting task with language learners, finding that *effort expectancy* and *facilitating conditions* were most predictive of podcast use. However, *social influence* was more important for Dutch university students using the MySpeechTrainer program deployed on the same platform as the present study (Strik, *et al*., 2019).

Numerous qualitative reports further indicate a complex relationship between technology acceptance and MAPT. In one study in Korea, middle school students noted they particularly liked the immediate response of an ASR-enabled MAPT app (Ahn & Lee, 2016). However, university students in Taiwan were less positive with its use as their qualitative responses indicate occasional frustration with a different ASR technology (Chen *et al*., 2020). Together, these findings suggest

that MAPT with ASR feedback may be perceived differently by learners with different technologies, expectations or backgrounds.

The Study

In order to further investigate the relationships of learner background and technology acceptance on use of MAPT amongst diverse learners, the present study is guided by the following research questions.

> RQ1: To what extent does learner accuracy vary across target features with a MAPT app?
> RQ2: To what extent do learner background and technology acceptance relate to learner effort when using a MAPT app?
> RQ3: To what extent do learner background and technology acceptance relate to accuracy when using a MAPT app?

METHODS

Participants

Adult ESOL learners (n = 34) were recruited for the study. Their average age was 43 years (range 18 to 71). Ten identified as male and 24 as female. First languages (L1s) were grouped into the language families of *Romance* (22), *Semitic* (2), *Sino-Tibetan* (3), *Slavic* (4), and *Other* (3). Participants were recruited from local community-based learning sessions that provide language instruction to a variety of learners outside of academic settings in the southwest United States. Their lengths of residency in an English-speaking country were recorded as less than a year (11), one to two years (8), three to six years (3), and six or more years (12). Their proficiency was not measured.

Instruments

The participants completed a background questionnaire with items related to their age, length of residency in an English-speaking environment, number of years spent studying English and learning preferences. The post-treatment survey included 19 items adapted from the UTAUT (Venkatesh *et al.*, 2012) using a mobile-friendly, four-point Likert scale.

The Communication Tutor App

The Communication Tutor App was developed and administered on the NovoLearning platform (Novo Learning BV, 2019), a mobile app designed for the creation and delivery of pronunciation training that is commercially available for iPhone and Android devices. Six lessons on minimal pair contrasts were devised expressly for the present intervention and were organized into three sessions. The first session focused on the contrasts of /b-p/ and /ɪ-i/. The second session was on /p-f/ and /ɪ-æ/. The third session targeted /d-ʒ/ and /ɑ-oʊ/. All minimal pairs were chosen from feature lists from Kang and Moran's (2014) high functional load segments and pedagogical guides from

Celce-Murcia *et al*. (2010) with an estimation of the features important for the present population as the precise background was unknown prior to data collection.

Each lesson contained two to four multiple choice items for listening discrimination of the target phonemes in minimal pairs followed by a range of seven to twelve ASR-enabled speaking tasks with words both in isolation and in sentences. For listening discrimination tasks, an L1 speaker audio file was played before answering a two-alternative forced choice question that contained a corresponding option to the phonemic contrast targeted by the lesson. Responses were marked as 0 if incorrect or 100 if correct. For ASR-enabled speaking tasks, an audio model of an L1 English speaker was provided before recording. When the learner had completed the response, the automated ASR feedback process immediately indicated if the response was correct or incorrect. If incorrect, the program provided phonemic-level metalinguistic feedback with audio samples marked visually through colorization. The learner was then prompted to repeat the task but allowed the option with a smaller button on the screen to continue to the next task.

The NovoLearning platform (Novo Learning BV, 2019) uses the built-in speaker-independent ASR processor to compare phonemic target forms pre-programmed by the course designer. To verify ASR accuracy, 120 speech samples (3.1% of all ASR attempts) were independently evaluated by two pronunciation researchers. An agreement rate of 77.2% was lower than a previous ASR L2 study that used a similar technology (86% in Cucchiarini *et al*., 2007).

Procedure

Using their own mobile phones, participants accessed the background surveys via a web browser and installed the mobile application by NovoLearning (Novo Learning BV, 2019). The participants then completed the lessons outlined above in the period of 2-3 weeks, resulting in data stored in the cloud for each task that included a count of the number of attempts and the accuracy. The app presented the six lessons in order and allowed participants to repeat lessons if they desired. Many participants completed the lessons in five or fewer sessions. However, four participants repeated previously completed lessons and one participant repeated a lesson seven times. At the end of the intervention, participants completed the modified UTAUT survey and were compensated with a gift card.

Analysis

For technology acceptance, internal consistency was tested for the UTAUT items and the resulting Cronbach alpha was sufficient (α = .91), however the small sample size resulted in insufficient model fit indices when loaded into SEM analysis (CFI = .24, RMSEA = .22, X^2 = 356.92). As such, the technology acceptance was treated as a composite average of all UTAUT items. Micro-averages of *accuracy* and the number of attempts (i.e., *effort*) were computed per participant within each lesson using linear mixed effect modeling in the lme4 package in R. RQ1 was addressed by fitting the model with the dependent variable of *accuracy*, fixed effects of the *lesson*, and a random effect of the *participant* in order to estimate means across target feature types. For RQ2 and RQ3, parallel models fit the dependent variable of *attempts* or *accuracy*, respectively, to the fixed effects of *age* and *technology acceptance* with the random effects of *lesson* and *participant*. Two other

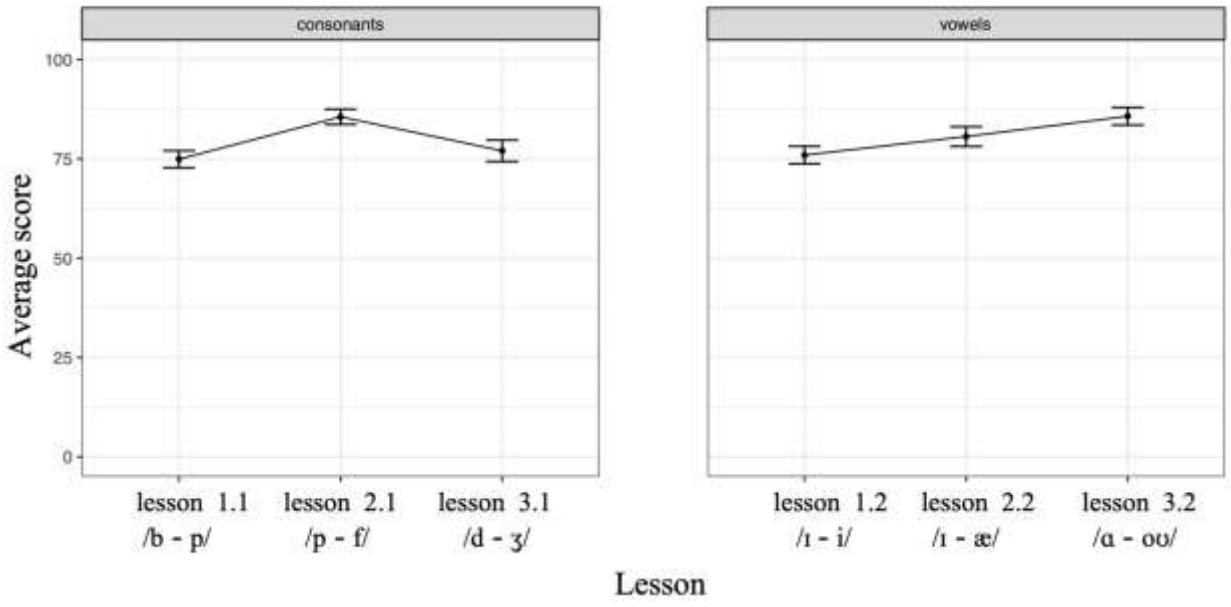
background variables, L1 family and length of residency, were excluded from the model as a process of backwards elimination indicated they did not significantly add to the models. All models met statistical assumptions and R^2 was calculated post-hoc for comparison of effects.

RESULTS

Accuracy scores and the number of attempts (i.e., the operationalization of effort) from 34 participants completing six lessons were compiled. To address the relationship between *lesson* and *accuracy*, the data were visualized using margin mean graphs revealing divergent trends for consonants and vowels (see Figure 1).

Figure 1

Marginal means plot of target segmental accuracy grouped by consonant and vowel lessons



The mixed effect model summary includes estimates of vowels and consonants for each lesson and is presented in Table 1. For consonants, Lesson 1.1 is set as the intercept to which the Lessons 2.1 and 3.1 are compared. Lesson 2.1 was significantly higher than the Lesson 1.1 ($p < .001$), indicating more accurate responses by the ESOL learners in Lesson 2.1. However, Lesson 3.1 was not significantly different ($p = .446$) indicating relatively similar performance between Lesson 1.1 and 3.1. For vowels, Lesson 1.2 was also set as the intercept for comparison. ESOL learner accuracy in Lesson 2.2 ($p = .029$) and Lessons 3.2 ($p < .001$) were both significantly higher than Lesson 1.2. R^2 calculations indicated that the variable *lesson* accounted for 10.83% of the variance in the consonant model and 8.19% in the vowel model. However, the random effect of *participant* accounted for 23.38% in the consonant model and a much higher 53.9% in the vowel model, revealing individual variation to be more influential in vowel accuracy as compared to consonant accuracy.

Table 1

LMEM summary for consonants and vowels by lesson

	Fixed Effects			Random Effect	
Parameters	Estimate (*Std. Err.*)		*p*	participant	*SD*
Consonants					6.81
1.1 /b-p/ (Intercept)	74.91(2.28)	***	<.001		
2.1 /p-f/	10.68(2.77)	***	<.001		
3.1 /d-ʒ/	2.13(2.77)		.446		
Vowels					10.29
1.2 /ɪ-i/ (Intercept)	75.96(2.30)	***	<.001		
2.2 /ɪ-æ/	4.66(2.09)		.029*		
3.2 /ɑ-oʊ/	9.77(2.09)	***	<.001		

*p<0.05, **p<0.01, ***p<0.001

Background, Technology Acceptance, and Effort

RQ2 includes background variables for 28 participants as six did not complete the UTAUT survey. Preliminary plots of *age* and *technology acceptance* as predictors of *accuracy* and *effort* revealed non-linear relationships. Therefore, both *age* and *technology acceptance* were computed into categorical variables. Participants were grouped by decade: 18-29 (7), 30-39 (5), 40-49 (5), 50-59 (7), and 60-71 (4). The composite UTAUT scores were also converted into a categorical variable to describe those with high (14), medium (8), and low (6) acceptance based on cut points that balanced group sizes and clustering of composite scores.

The mixed effect model fit *attempts* to *technology acceptance* and *age* as fixed effects with the random effect of *lesson.* This creates an intercept that includes the first level of both categorical predictors of the dependent variable: the low *technology acceptance* level and participants aged 18-29 level. Model summary results indicate comparisons of levels resulted in only the age level of 60-71 as statistically significant (p = .008). Post-hoc analyses indicate age as more predictive (5.74%) than either *technology acceptance* (0.58%) or the lesson (1.36%). See Table 2 for model summary.

Table 2

LMEM summary for attempts

Parameters	Fixed Effects		Random Effect
	Estimate (*Std. Err.*)	*p*	Lesson *SD*
(Intercept)	1.63(0.36)	<.001***	.18
Tech_Accept_Med	0.37(0.36)	.311	
Tech_Accept_High	0.17(0.30)	.576	
Age_30to39	0.10(0.39)	.807	
Age_40to49	0.01(0.39)	.971	
Age_50to59	0.48(0.33)	.143	
Age_60to71	1.06(0.39)	.008**	

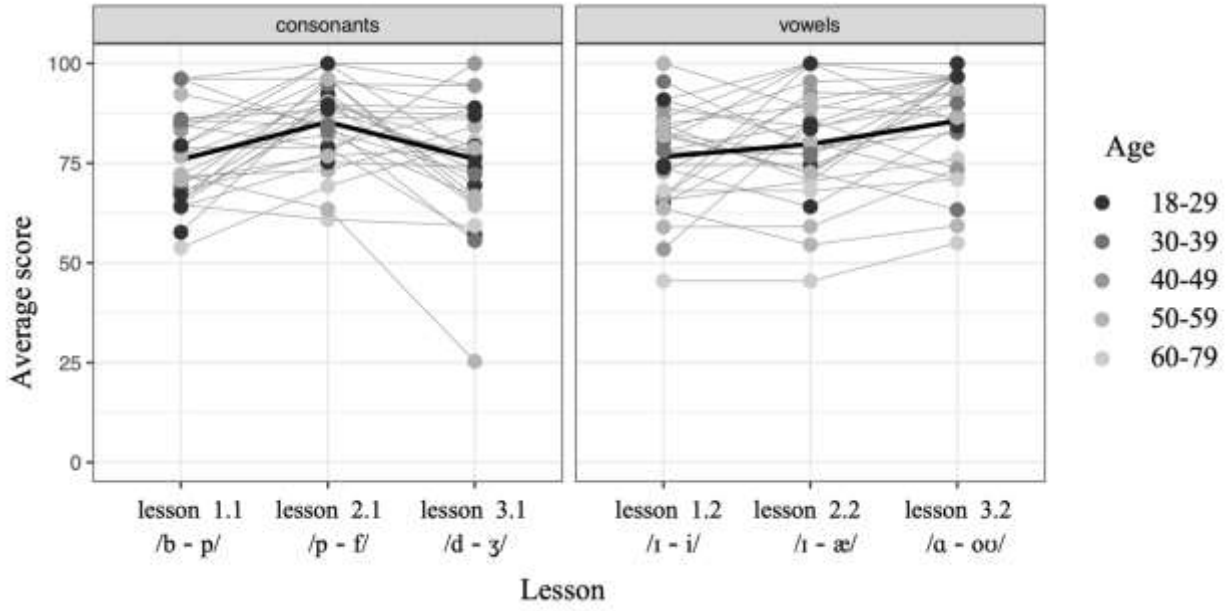
*p<0.05, **p<0.01, ***p<0.001

Background and Attitudes for Pronunciation Accuracy

RQ3 examined the extent to which *age* and *technology acceptance* moderate *accuracy* while using the App and is visualized in Figure 2. The average score plots indicated an upward trend in the accuracy of the vowel-targeted lessons but the same was not true in the consonant lessons similar to analysis for RQ1. The thick black line indicates the mean scores for the lesson.

Figure 2

Accuracy scores per participant with grayscale age bands.



Fixed effects summary results for consonant and vowel models similarly create intercepts that include the first level of both predictor variables. The model reports the *age* category of 60-70 to be significantly different ($p = .029$) to the intercept for the vowel model. In other words, the oldest participants scored on average 16.42 points lower (100-point scale) than the youngest participants. The other age groups did not vary significantly between any other factor level comparisons to the intercept. See Table 3 for summary of models.

Table 3

LMEM summary for consonants and vowel accuracy

	Fixed Effects		Random Effects	
			Participant	Lesson
parameters	Estimate (*Std. Err.*)	*p*	*SD*	*SD*
Consonants				
(Intercept)	80.63(5.59)	<.001***	5.16	4.95
Tech_Accept_Med	-2.44(4.97)	.629		
Tech_Accept_High	-1.78(4.10)	.668		
Age_30to39	3.03(5.39)	.580		

Age_40to49	8.40(5.40)	.135			
Age_50to59	-3.51(4.47)	.442			
Age_60to71	-7.77(5.36)	.162			
Vowels					
(Intercept)	84.22(6.74)	<.001***	9.545		4.2
Tech_Accept_Med	-0.74(6.49)	.910			
Tech_Accept_High	0.97(5.36)	.858			
Age_30to39	-2.01(7.04)	.778			
Age_40to49	0.69(7.05)	.923			
Age_50to59	-5.01(5.84)	.401			
Age_60to71	-16.42(7.00)	.029*			

*p<0.05, **p<0.01, ***p<0.001

In the *attempts* model, *technology acceptance* played a small role in the variance for consonants (1.01%) and vowels (0.20%). *Age* was a much more explanatory predictor in consonants (11.67%) and even more so in vowels (14.38%), partially reflecting the significant relationship of the oldest learners scoring lower in vowels. When considering the random effects, *lesson* explained relatively higher amounts of variation in the consonant (12.82%) model as compared to the vowel model (8.21%), indicating that the learners were less consistent in the consonant lessons despite the divergent patterns found in RQ1. Perhaps of most interest is the random effect of *participant* signifying more consistency in the consonant model (14.20%) than in the vowel model (41.47%). R^2 computations were tabulated for all analyses in RQ2 and RQ3 for comparison between models and converted to percentages. See Table 4 for summary of variance explained.

Table 4

Variance explained summary for RQ2 and RQ3

Model	Variance explained				
	Technology acceptance	Age	Lesson	Participant	Total
Attempts (Consonants and vowels)	0.58%	5.74%	1.36%	-	7.68%

Accuracy (Consonants)	1.01%	1.67%	2.82%	1 %	14.20	3 9.70%
Accuracy (Vowels)	0.20%	4.38%	8 .21%	1 %	41.47	6 4.26%

DISCUSSION

Focusing on adult ESOL participants in non-academic settings, the study investigated the relationships between learner background and attitudes towards MAPT-based technology applications. RQ1 investigated the differences in accuracy scores by lesson without consideration of background variables, resulting in divergent patterns for consonants and vowels. The consonant lessons' results only flagged a significant gain for the second of the three lessons, perhaps due to the repetition of /p/ in Lessons 1.1 and 1.2. However, this relationship plays out differently for vowels, which resulted in significant increases in accuracy for each lesson despite the presence of novel phonemes in lesson 3.2. At first glance, this seems to conflict with previous findings about the non-linear nature of vowel acquisition (Munro & Derwing, 2008), but the differences in the random effect results contribute greatly to this interpretation. The variance explained by participants in the accuracy models (14.20% for consonants and 41.47% for vowels) reveal the extent to which factors not measured in this study relate to inter-learner variability such as learnability and previous exposure.

For the model on learner effort, the relationships found amongst the adult ESOL learner background and attitudes indicate several unexpected findings that have not been previously seen in the literature. Primarily, the technology acceptance results indicated very little explanatory ability in terms of the variance of effort by the participants on each item, a finding somewhat contradictory to the high relationship (61%) with the UTAUT results and the intention to use an App found in Strik *et al*. (2019). However, intention does not appear to equate to realized effort. As the App requested learners to retry tasks they did not do correctly several times, the effort variable represents not only the learner's intention, but also their ability to learn to produce phonemes accurately and requisite conditions for the learner to continue using the app. Additionally, the relationship between age and effort found in the present study is of particular note. The findings of increased use by the 60-71 age group indicates a willingness to engage with MAPT in an otherwise underrepresented group in much of technology-based language learning research. In sum, the limited variance explained by the measured background variables in the effort model (7.68%) warrants further exploration of motivation, aptitude, and other individual differences.

For learner accuracy in the MAPT App, it is of great interest that the parallel analyses on consonant and vowel accuracy resulted in only significant differences for the age range of 60-71 with vowel models as little research has been done on this age group and phonemic acquisition. While the group is small in the present study (n = 5), it is notable that model selection did not indicate that L1 family nor length of residency played a role. Additionally, technology acceptance played a minimal role in the accuracy scores. At first sight, this finding might be counterintuitive as one would expect more tech-oriented users to learn more quickly when using a mobile device.

However, the results generally indicate that linguistic factors are more important than technology acceptance.

This study is limited in several ways in that it did not directly examine the effect of MAPT on the given population in order to describe the relationship between effort and accuracy on individual tasks with increases in competence. In addition, proficiency is not controlled or measured in the study, which may be an important moderator in accuracy in the App. Finally, a lengthier MAPT intervention with more choices and lessons in a longitudinal study may alter the relationships of measured variables.

Implications

The study has numerous implications for L2 pronunciation and MAPT researchers. Primarily, the weak relationships between background variables that are often considered in language teaching (e.g., L1, length of residency) are not corroborated. This is a positive finding for MAPT researchers who wish to apply their findings to populations who are using MAPT applications and intend to work with populations not typically found in L2 pronunciation research. In addition, the results may call for a review and further testing of a technology acceptance model for special populations such as adult ESOL learners. For MAPT-based application developers, the findings are encouraging as they show that performance and effort are not a function of the user technology acceptance. This should encourage developers to focus on the content and relevant linguistic features rather than technological innovations, supporting previous calls by Neri *et al*. (2003), in which pedagogy must drive technology.

ABOUT THE AUTHORS

Kevin Hirschi is a doctoral student in the Applied Linguistics program at Northern Arizona University. His interests include second language pronunciation, corpus linguistics, and mobile-assisted pronunciation technology.

Northern Arizona University
English Department
705 S Beaver St, Flagstaff, AZ 86011
KevinHirschi@nau.edu

Okim Kang is a Professor in the Applied Linguistics Program at Northern Arizona University, Flagstaff, AZ, USA. Her research interests are speech production and perception, L2 pronunciation and intelligibility, L2 oral assessment and testing, automated scoring and speech recognition, World Englishes, and language attitude.

John Hansen serves as Associate Dean for Research, and Professor of Electrical Engineering, Speech & Hearing Sciences at Univ. of Texas at Dallas. He oversees the Center for Robust Speech Systems and serves as ISCA President. He has authored/co-authored 752 papers in the field of speech processing and language technology.

Catia Cucchiarini is Principal Investigator at the Centre for Language and Speech Technology of the Radboud University Nijmegen and Senior Consultant at The Dutch Language Union in the Hague. She conducts research on speech processing, language learning, and speech technology applications in Computer Assisted Language Learning and e-health.

Helmer Strik is Associate Professor in Speech Science and Technology at Radboud University Nijmegen, co-founder and CSO of NovoLearning, and Chair of the ISCA SIG 'Speech and Language Technology in Education' (SLaTE). His research addresses spoken dialogue systems, automatic speech recognition (ASR), and their use in e-Learning and e-Health.

REFERENCES

Ahn, T. Youn, & Lee, S.M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology, 47*(4), 778–786.

Chen, W., Inceoglu, S., & Lim, H. (2020). Using ASR to improve Taiwanese EFL learners' pronunciation: Learning outcomes and learners' perceptions. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching conference*, ISSN 2380-9566, Northern Arizona University, September 2019 (pp. 37–48). Ames, IA: Iowa State University.

Cucchiarini, C., Neri, A., de Wet, F., & Strik, H. (2007). ASR-based pronunciation training: Scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners. *Proceedings of Interspeech 2007* (pp. 2181–2184). Antwerp, Belgium.

Tejedor-García, C., Cardenoso-Payo, V., Cámara-Arenas, E., Gonzalez-Ferreras, C., & Escudero-Mancebo, D. (2016). Measuring pronunciation improvement in users of CAPT tool TipTopTalk! *Proceedings of Interspeech 2016* (pp. 1178–1179). San Francisco, USA.

Ho, C. T. B., Chou, Y. T., & O'Neill, P. (2010). Technology adoption of mobile learning: A study of podcasting. *International Journal of Mobile Communications, 8*(4), 468–485.

Kaiser, D. (2018). Mobile-Assisted Pronunciation Training: The iPhone Pronunciation App Project. *IATEFL Pronunciation Special Interest Group Journal, 58*, 38–52.

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. TESOL Quarterly, 48(1), 176–186.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics, 36*(3), 345–366.

Mahdi, H. S., & Al Khateeb, A. A. (2019). The effectiveness of computer-assisted pronunciation training: A meta-analysis. *Review of Education, 7*(3), 733–753.

McCrocklin, S., Humaidan, A., & Edalatishams, E. (2019). ASR dictation program accuracy: Have current programs improved? In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Ames, IA, September 2018 (pp. 191–200). Ames, IA: Iowa State University.

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning, 58*(3), 479–502.

Neri, A., Cucchiarini, C., & Strik, H. (2003). *Automatic speech recognition for second language learning: How and why it actually works*. Proceedings of 15th International Conference of Phonetic Sciences (pp. 1157–1160). Barcelona, Spain.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*(5), 441–467.

Novo Play. (2019). *Novo Learning BV* (Version 4.0.1) [Mobile app]. App Store. https://apps.apple.com/us/app/novo-play/id1260504406

Park, A. Y. (2017). The Study on Automatic Speech Recognizer Utilizing Mobile Platform on Korean EFL Learners' Pronunciation Development. *Journal of Digital Contents Society, 18*(6), 1101–1107.

Strik, H., Ovchinnikova, A., Giannini, C., Pantazi, A., & Cucchiarini, C. (2019). Student's acceptance of MySpeechTrainer to improve spoken academic English. *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education* (pp. 48–52). Graz, Austria.

Suzukida, Y. (2021). The contribution of individual differences to L2 pronunciation learning: Insights from research and pedagogical implications. *RELC Journal, 52*(1), 48–61.

Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36*(1), 157–178.

APPENDIX

The UTAUT questionnaire (Venkatesh et al., 2012) adapted for CT App, overall α =.91.

Statement		*M*	*SD*
Attitude toward using technology (α =.67)			
At1. Using Communication Tutor is a good idea.	.75	3	0.59
At2. Communication Tutor makes language learning more interesting.	.71	3	0.46
At3. I like learning with Communication Tutor.	.68	3	0.55
Effort Expectancy (α =.76)			
EE1. It is clear and understandable to me how to use Communication Tutor.	.64	3	0.56
EE2. I find Communication Tutor easy to use.	.50	3	0.64
EE3. Learning to use Communication Tutor is easy for me.	.61	3	0.57
Facilitating Conditions (α =.67)			
FC2. I have the knowledge necessary to use Communication Tutor.	.61	3	0.57
FC3. Communication Tutor is compatible with the device I use.	.43	3	0.84
FC4.I have the resources necessary to use Communication Tutor (smartphone, internet connection, microphone etc.)	.64	3	0.73
Hedonic Motivation (α =.78)			
HM1. Using Communication Tutor is fun.	.32	3	0.77
HM2. Using Communication Tutor is enjoyable.	.50	3	0.69
HM3. Using Communication Tutor is very entertaining.	.57	3	0.57
Habit Formation (α =.59)			
Ht1. It has become my habit to learn languages with mobile apps	.32	3	0.86
Ht2. I often learn language(s) in mobile or computer applications.	.21	3	0.79
Ht3. I feel that I must use mobile or computer applications to learn languages.	.50	3	0.92
Performance Expectancy (α =.74)			

6	PE1. I find Communication Tutor useful in my studies.	.61	3.63	0
7	PE2. Using Communication Tutor would enable me to speak English better.	.54	3.84	0
8	PE3. Using Communication Tutor would improve my English-speaking skills.	.61	3.57	0
9	PE4. If I practice English with Communication Tutor, I will increase my chances of studying successfully	.50	3.84	0