INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Mobile-Assisted Prosody Training for Limited English Proficiency: Learner Background and Speech Learning Pattern

Kevin Hirschi[1], *Okim Kang*[1], *Catia Cucchiarini*[2], *John Hansen*[3], *Keelan Evanini*[4], *Helmer Strik*[2,5]

[1]Northern Arizona University, USA
[2]Radboud University, Nijmegen, Netherlands
[3]University of Texas at Dallas, USA
[4]Educational Testing Service, USA
[5]NovoLearning B.V., Nijmegen, Netherlands

kevinhirschi@nau.edu, okim.kang@nau.edu, c.cucchiarini@let.ru.nl,
john.hansen@utdallas.edu, kevanini@ets.org, w.strik@let.ru.nl

Abstract

The use of Mobile-Assisted Pronunciation Training (MAPT) has been increasing drastically due to the personal and interactive nature of mobile devices. However, MAPT applications lack support from empirical evidence as research on MAPT-based acquisition, particularly related to prosody, has been rare. The present study employs a MAPT application with lessons on lexical stress and prominence with Limited English Proficiency (LEP) users (n = 31) of mixed ages and first languages. Then, 16 experienced raters conducted discourse-based prosodic analysis on unconstrained speech collected at the beginning and the end of the intervention. A series of mixed-effect model analyses were conducted on learner effort, improvement and learner background to investigate their relationship with accentedness and comprehensibility. The results indicated that present MAPT prosody interventions were effective for comprehensibility but not accentedness, however, learner effort on lexical stress and prominence exhibit differing patterns. Similar to previous findings, learner age impacts production more than the length of residency or history of language study. Implications include a prosody-based MAPT application; support for the treatment of accentedness and comprehensibility as separate, but related constructs; and a further understanding of the role of learner-related factors in prosody intervention.

Index Terms: Mobile-Assisted Pronunciation Training (MAPT), Limited English Proficiency (LEP), lexical stress, prominence

1. Introduction

Prosody, referring to variation in pitch, loudness, tempo, and rhythm for the purposes of this study [1], is known to influence listener perception of second language (L2) comprehensibility and accentedness [2]. Accordingly, prosodic features are commonly targeted in pronunciation interventions, some of which use technology. However, in the past few decades, researchers and teachers have incorporated mobile devices in Mobile Assisted Pronunciation Training (MAPT) as it allows for engagement different from personal computers through (a) embedded audio playback, (b) simplified voice recording, and (c) the possibility of Automatic Speech Recognition (ASR) to provide feedback [3]. While numerous studies have shown positive effects of technology in pronunciation acquisition with English for Academic Purposes (EAP) learners [4], little is known about the effect of MAPT and adult learners outside of higher education. Designated as Limited English Proficiency (LEP) by the US Federal Government, these individuals have limited English ability and primarily speaks a language besides English. LEP individuals come from diverse background including both immigrants and native born, but are overall less likely to be educated, more likely to live in poverty, and have limited access to language skill improvement in the US [5].

2. Prosody in Speech Perception

2.1. Lexical Stress

Stress is one of the most important speech properties that determines listeners' judgments of accented speech [6], [7] and affects L2 speakers' oral production [2]. In the pronunciation-based hierarchical structure proposed in Kang [8], stress (both lexical and prominence below) was first ranked, followed by fluency measures, segmental errors, and tone choices. Despite variation in lexical stress in global varieties of English, strong evidence exists in its relationship with comprehensibility in addition to accentedness [9], and, may be particularly relevant for recently immigrated LEP participants from locales with differing stress patterns.

2.2. Prominence

Prominence (sentence-level or tonic stress) is an important feature in the pronunciation syllabus, including proposed standards for English as an International Language [10]. Historically, research has shown that misplaced prominence causes confusion in the listener, resulting in communication breakdown [11], [12]. The proper use of sentence stress could reduce the cognitive load of the listeners while processing the content of L2 speech and previous studies have indicated that prominence makes a significant contribution to L2 comprehensibility ratings [13]. Approaches to instruction on prominence typically center around communicative language teaching but also support scaffolded training through perceptive and controlled tasks [14].

2.3. The Study

The present study seeks to understand the relationships between MAPT on prosody through two research questions:

http://dx.doi.org/10.21437/Interspeech.2020-2901

1. To what extent does MAPT on prosodic features effect learner speech?
2. How do background variables contribute to MAPT acquisition of prosody?

3. Method

3.1. Participants

LEP participants were recruited from community organizations that offer free English courses in the Southwest United States. 31 LEP learners (22 female, 9 male, median age = 42 years, range = 18 to 71) of mixed proficiency and ten different L1s completed lessons on a variety of pronunciation and speaking issues using a MAPT application. Unconstrained speech files from these participants were then rated by 16 native or highly proficient English speakers (10 female, 6 male, 7 NS, 9 NNS) who had experience teaching or researching English prosody.

3.2. Instruments

3.2.1. Background Questionnaire

LEP participants first completed a pre-intervention adapted language contact questionnaire with items related to their age, L1, Length of Residency (LOR), and experience in studying English. Participants were assisted in cases when translation was needed. All surveys were optimized for mobile devices.

3.2.2. The Novo Play app

The lessons on pronunciation were delivered through the Novo Play app (https://novo-learning.com). Each lesson took 10-15 minutes and consisted of 8-15 tasks completed over a period of 1-2 weeks. Within each lesson, 2-4 listening tasks presented the target feature with enhanced visualization (e.g., lib**rar**ian for lexical stress) and an audio sample of a native speaker saying the word.

Speaking tasks also presented the target form in written (without visual enhancement in this case) and audio form. The audio model could be played by tapping a button or the learner's response could be given with the microphone button. When the participant tapped the microphone button, the ASR feature was engaged and upon completion of the utterance, it immediately analyzed the learners' speech. If the utterance was correct according to the forms programmed by the researchers, a green check mark appeared, and the learner was prompted to continue to the next task. If the ASR detected an incorrect form, a red box appeared that the learner could tap for feedback, presented through phonemic transcription with marked stressed and unstressed forms. Guidance in interpreting these forms was provided textually throughout the lessons.

If a response was marked incorrect, learners were encouraged, but not required, to repeat the task until their responses were correct. At the end of the lesson, learners were also told they could repeat the lesson. For previous validation and user perception of the NovoLearning platform see [15], [16].

3.2.3. Lexical Stress and Prominence Lessons

The lexical stress lesson was adapted from a classroom pedagogical text [17] with sentence-length targets collected from spoken corpora. Through modeling the correct stress, three high-frequency words (*librarian*, *technician*, and *politician*) were then presented and elicited from the learner first in a multiple-choice perception task, then a single-word speaking task, and finally a sentence-length speaking task. Each speaking task received immediate ASR feedback as outlined above and native speaker audio models were available.

A similar lesson for prominence included three statements with contrastive prominence based on contextual differences adapted from a pedagogical resource [18]. For example, one task presented a misunderstanding of a telephone number segment (e.g., *925*) accompanied with the task of stressing the nine over the other numbers as a form of correction. Each statement was presented with two contrastive target prominence forms. The resulting speaking tasks also included ASR feedback and audio models by native speakers.

3.2.4. Pragmatic and Other Lessons

Pragmatic lessons captured speech through a contextualized Discourse Completion Task (DCT). The task was presented using a photograph and a description of the situation in text. Unconstrained speech was elicited during the DCT which was followed by suggestions of pragma-linguistic forms appropriate to the task. Participants often chose to repeat these tasks. Other lessons included segmental target features and rhythm. However, because of pedagogical concerns in ordering the lessons, the present study reports only on the effect of two prosodic lessons (lexical stress and prominence) with speech from two high-imposition DCTs sequenced to allow a pre / post measurement.

3.2.5. DCT Rating

The DCT speech files were collected and reviewed for completeness. The last attempt with a complete response was selected and embedded into an online survey platform for the 16 experienced raters to evaluate. Each evaluator rated all samples in randomized order for accentedness, comprehensibility, lexical stress accuracy, and prominence accuracy on a 100-point slider scale in which 100 was the highest score (i.e., not accented at all, very comprehensible, very accurate in lexical stress, and very accurate in prominence), and zero was the lowest score (i.e., very accented, not at all comprehensible, not at all accurate in lexical stress, and not at all accurate in prominence) [19].

3.3. Analysis

The results of the expert raters were examined for reliability using ICC(3,*k*) and subsequently joined with participant background variables of LOR, years spent studying English, and participant age. Data on participant use of the Novo Play app was compiled and the effort in learning, operationalized by the number of attempts within a lesson, was used as a predictor variable.

A series of mixed effect models was performed with four dependent variables: (a) accentedness, (b) comprehensibility, (c) lexical stress accuracy, and (d) prominence accuracy. The models included the fixed effects of time (Time 1 is the DCT unconstrained speech sample before the prosody intervention and Time 2 is after the intervention), effort on the lexical stress MAPT lesson, and the prominence MAPT lesson. Random effects included LOR (cut into three categories: <1 year, 1-4 years, 5+ years), age (cut into five categories per decade), years spent studying English, and rater (required for the present design). Residuals for each model were plotted and examined for violations of the assumptions of normally distributed residuals and homoscedasticity. None were found.

4. Results

4.1. Impact of MAPT

Results were tabulated for the unconstrained speech at *Time 1* (before intervention) and *Time 2* (after intervention). The mean scores indicated gains for all four target constructs from the *Time 1* to *Time 2*. Additionally, the rater reliability was above 0.94 for all constructs in each group (see Table 1).

Table 1: *Descriptive statistics and rater reliability.*

	M	SD	ICC(3,*k*)
Accentedness			
Time 1	26.19	23.00	0.97
Time 2	28.14	23.15	0.96
Comprehensibility			
Time 1	72.91	24.01	0.95
Time 2	79.67	21.03	0.94
Lexical Stress			
Time 1	63.92	25.47	0.97
Time 2	67.62	23.57	0.94
Prominence			
Time 1	58.41	25.47	0.97
Time 2	61.30	24.81	0.95

Descriptive results were plotted with standard error bars to visualize differences from *Time 1* to *Time 2*. The resulting visual indicates gains in each of the four measured constructs (see Figure 1).

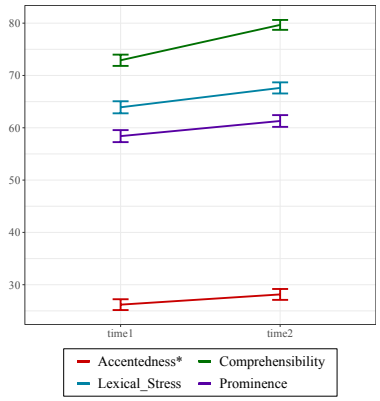

Figure 1: *Marginal means by Time 1 and Time 2. Note: * indicates a reversed scale.*

4.1.1. Learning Gains and Speech Pattern

The fixed effects results indicate a number of significant relationships between the learning gains, prosodic accuracy, and speech perception ratings (see Table 2). Primarily, the factors of time (intercept and *Time 2*) was flagged as significant in all models except *Accentedness*, indicating a gain in proficiency during the intervention. Reviewing the estimated means reveals the extent to which this varies as in the *Accentedness* model, the second measurement increased the score minimally (2.01) with larger differences found in *Lexical Stress* and *Prominence* models (3.81, 2.90, respectively), and the largest difference in C*omprehensibility* (6.73).

The effort in the lexical stress lesson was significant in only the C*omprehensibility* ($p < .001$) and the L*exical Stress* ($p = 0.01$) models, indicating that more effort in the lesson was related to a higher *Time 2* score for both. Effort in the *Prominence* lesson was significantly but negatively related to the rating score for all four constructs ($p < .001$).

Table 2: *Fixed effects of the four models.*

	Est.	SE	t	p
Accentedness				
(intercept)	32.25	4.34	7.42	< .001***
Time 2	2.01	1.12	1.79	0.07
Effort L.S. lesson	0.00	0.13	0.04	0.97
Effort Prom. lesson	-0.17	0.02	-7.01	< .001***
Comprehensibility				
(intercept)	68.20	3.63	18.79	< .001***
Time 2	6.72	1.20	5.56	< .001***
Effort L.S. lesson	0.89	0.14	6.59	< .001***
Effort Prom. Lesson	-0.17	0.03	-6.69	< .001***
Lexical Stress Rating				
(intercept)	65.84	4.14	15.92	< .001***
Time 2	3.81	1.26	3.03	< 0.01**
Effort L.S. lesson	0.38	0.14	2.73	< 0.01**
Effort Prom. Lesson	-0.19	0.03	-7.13	< .001***
Prominence Rating				
(intercept)	63.28	4.37	14.49	< .001***
Time 2	2.91	1.31	2.22	0.03*
Effort L.S. lesson	0.17	0.15	1.14	0.26
Effort Prom. Lesson	-0.20	0.03	-7.08	< .001***

4.1.2. Effort and Outcome Variables

In order to better understand the relationship between effort on the MAPT lessons and the dependent variables, marginal mean plots with loess smoothing were generated. In these plots, the x-axis serves as the number of attempts within the MAPT lesson and the y-axis is the perception rating of the experienced evaluators on the *Time 2* speech file. All participants completed at least six attempts (i.e., one attempt per task), however, several participants repeated tasks in the prominence lesson in excess of 50 times. As outlined above, participants were prompted to repeat a task if their attempt was inaccurate and were encouraged to repeat lessons if they found them beneficial.

Figure 2 reveals a small but positive trend between effort on the lexical stress lesson and *Time 2* ratings of all constructs except *accentedness*, indicating that quickly mastered items were related to unaccented speech. However, participants who repeated the lexical stress items numerous times also resulted in unaccented speech. Those who repeated the items only once or twice had more accented speech in the *Time 2* ratings.

A parallel plot was constructed to visualize the relationship between effort in the prominence lesson and the ratings across the four constructs (see Figure 3). The results diverge from the prior figure in that they reveal an increase in rating until approximately 20 attempts are made (~ three repetitions of each item). However, after this point, the ratings trend flat until approximately 60 attempts (~ ten attempts on each item) and then downward with additional attempts.

4.2. Participant Background

Random effect summaries were generated for each dependent variable (see Table 3). The variables of prior years of English study was dropped from the models as precursory analyses indicated this variable explained little variance in the outcome

variable. In the resulting models, the effect of rater explained much variance across dependent variables. However, a divergent pattern emerges for the role of age and LOR: age explains more variance in *Accentedness* and *Comprehensibility* while LOR explains slightly more variance in *Lexical Stress* and *Prominence* ratings.

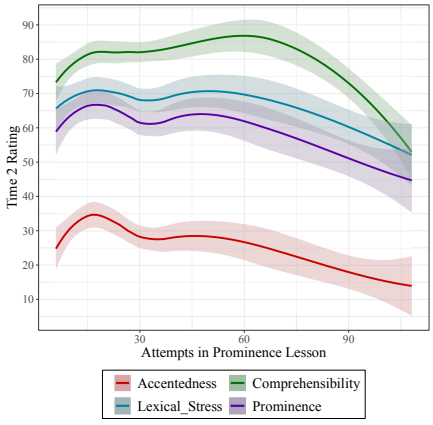

Figure 2: *Ratings and effort in the lexical stress lesson with SE regions.*

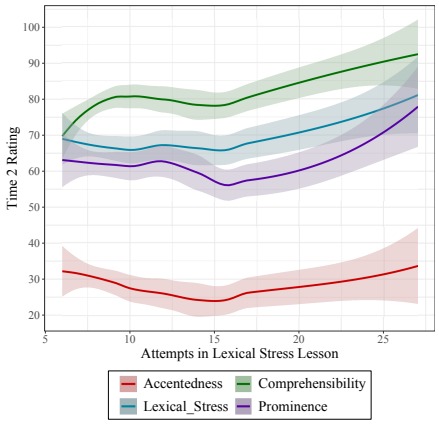

Figure 3: *Ratings and effort in the prominence lesson with SE regions.*

5. Discussion

Overall, the results provide evidence for the effectiveness of the MAPT app in the acquisition of lexical stress and prominence as detected in ratings of comprehensibility and accentedness. The MAPT intervention resulted in significant gains in comprehensibility (6.72), lexical stress (3.81), and prominence (2.91) despite the nature of unconstrained speech. However, it did not result in a significant improvement in accentedness, supporting claims that comprehensibility and accentedness are separate but related constructs [13].

Additionally, the plots reveal much about the complexities between effort in the suprasegmental lessons and the four constructs at hand. For lexical stress, highly proficient learners may not gain much through a single attempt in a MAPT lesson. Lower proficiency learners may need three or more attempts to see gains. For prominence, a different pattern is detected. The repeated attempts to complete a task resulted in measurably lower outcomes after an estimated 10 attempts per item. This may have been due to a lack of saliency of target prominence forms in the items as they may have been too difficult for some participant but also supports claims of the complexity of training accurate prominence [14].

Table 3: *Random effects of the four models.*

	Variance	SD
Accentedness		
Rater	211.50	14.54
Age	7.96	2.82
LOR	4.27	2.07
Comprehensibility		
Rater	125.24	11.19
Age	7.43	2.73
LOR	2.34	1.53
Lexical Stress		
Rater	207.10	14.40
Age	1.21	1.10
LOR	1.96	1.40
Prominence		
Rater	195.19	13.97
Age	6.23	2.50
LOR	6.14	2.48

In terms of learner background, the results illustrate the complexity of acquisition as moderated by learner background variables. Age plays a larger role in acquisition for the speech perception constructs of accentedness and comprehensibility. However, LOR and age explain similar amounts of variance in the prosody ratings, a finding supported by prior research in the effect of age on acquisition [20]. The exclusion of years of English study is also of interest as it is contrary to L2 pronunciation studies with university students and may reveal the complex backgrounds of the LEP population and their daily use of English. The findings support the consideration of LEP learners differently from EAP learners.

The study is limited in several ways. First, the small sample size resulted in some factor levels with few participants, limiting generalizability of the findings related to background variables. Second, the very short duration of the intervention targeting prosody may not have been enough to impact accentedness despite gains in other measures. Third, a delayed posttest must be employed in order to detect sustained proficiency gains from an intervention.

6. Conclusions

In a population that tends to be underserved with economic and educational opportunities [5], an individualized learning program is essential to promote the efficiency of proficiency for civic and employment success. The findings of the present study are promising because they revealed the MAPT-based interventions on selected speech features could be effective for such participants. In particular, gains not only in the target features of the intervention were detected, but an increase in comprehensibility was also found.

7. Acknowledgements

The authors would like to thank NovoLearning (https://novo-learning.com) for the use of their platform and the LEP participants, their teachers, and the raters for their time.

8. References

[1] D. Crystal, *A dictionary of linguistics and phonetics*, 6th ed. Malden, MA; Oxford: Blackwell Pub, 2008.

[2] O. Kang, "Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness," *System*, vol. 38, no. 2, pp. 301–315, 2010.

[3] D. Kaiser, "Mobile-Assisted Pronunciation Training: The iPhone Pronunciation App Project," *IATEFL Pronunciation Special Interest Group Journal*, vol. 58, pp. 38–52, 2018.

[4] J. Lee, J. Jang, and L. Plonsky, "The effectiveness of second language pronunciation instruction: A meta-analysis," *Applied Linguistics*, vol. 36, no. 3, pp. 345–366, 2015.

[5] J. Batalova and M. Fix, "A profile of limited English proficient adult immigrants," *Peabody Journal of Education*, vol. 85, no. 4, pp. 511–534, 2010.

[6] J. Field, "Intelligibility and the listener: The role of lexical stress," *TESOL quarterly*, vol. 39, no. 3, pp. 399–423, 2005.

[7] B. W. Zielinski, "The listener: No longer the silent partner in reduced intelligibility," *System*, vol. 36, no. 1, pp. 69–84, 2008.

[8] O. Kang, "Relative impact of pronunciation features on non-native speakers' oral proficiency," in *Proceedings of the Pronunciation in Second Language Learning and Teaching*, J. L. & K. LeVelle, Ed. 2013.

[9] T. Isaacs and P. Trofimovich, "Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings," *Studies in Second Language Acquisition*, vol. 34, no. 3, pp. 475–505, 2012.

[10] J. Jenkins, "A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language," *Applied linguistics*, vol. 23, no. 1, pp. 83–103, 2002.

[11] A. Cutler, "Beyond parsing and lexical look-up," in *New approaches to language mechanisms: A collection of psycholinguistic studies*, R. J. Wales, E.C.T. Walkers., North-Holland, 1979, pp. 133–149.

[12] L. D. Hahn, "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals," *TESOL quarterly*, vol. 38, no. 2, pp. 201–223, 2004.

[13] O. Kang, D. Rubin, and L. Pickering, "Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English," *The Modern Language Journal*, vol. 94, no. 4, pp. 554–566, 2010.

[14] J. M. Levis and A. O. Silpachai, "Prominence and information structure in pronunciation teaching materials," in *Proceedings of the 9th Pronunciation in Second Language Learning and Teaching conference*, 2018, pp. 2380–9566.

[15] H. Strik, A. Ovchinnikova, C. Giannini, A. Pantazi, and C. Cucchiarini, "Student's acceptance of MySpeechTrainer to improve spoken academic English," 2019.

[16] C. Giannini, "Evaluating the effectiveness of a tool to improve vocabulary skills in an academic context: MySpeechTrainer," Master's Thesis, 2019.

[17] M. Celce-Murcia, D. Brinton, and J. Goodwin, *Teaching pronunciation: A course book and reference guide*. Cambridge University Press, 2010.

[18] L. Grant, *Well said: Pronunciation for clear communication*. Heinle & Heinle, 2001.

[19] C. A. Roster, L. Lucianetti, and G. Albaum, "Exploring slider vs. categorical response formats in web-based surveys," *Journal of Research Practice*, vol. 11, no. 1, pp. D1–D1, 2015.

[20] T. Piske, I. Mackay, and J. Flege, "Factors affecting degree of foreign accent in an L2: A review," *Journal of phonetics*, vol. 29, no. 2, pp. 191–215, 2001.