Tip Tap Tones: Mobile Microtraining of Mandarin Sounds

Darren Edge[1], Kai-Yin Cheng[1,2], Michael Whitney[1,3], Yao Qian[1], Zhijie Yan[1], Frank Soong[1]
[1] Microsoft Research Asia, Beijing, China
[2] National Taiwan University, Taipei, Taiwan
[3] University of North Carolina Charlotte, Charlotte, NC, USA
{firstname.lastname}@microsoft.com, keynes.z@gmail.com, mwhitne6@uncc.edu

ABSTRACT
Learning a second language is hard, especially when the learner's brain must be retrained to identify sounds not present in his or her native language. It also requires regular practice, but many learners struggle to find the time and motivation. Our solution is to break down the challenge of mastering a foreign sound system into minute-long episodes of "microtraining" delivered through mobile gaming. We present the example of Tip Tap Tones – a mobile game with the purpose of helping learners acquire the tonal sound system of Mandarin Chinese. In a 3-week, 12-user study of this system, we found that an average of 71 minutes' gameplay significantly improved tone identification by around 25%, regardless of whether the underlying sounds had been used to train tone perception. Overall, results suggest that mobile microtraining is an efficient, effective, and enjoyable way to master the sounds of Mandarin Chinese, with applications to other languages and domains.

Author Keywords
Mobile Gaming; Microtraining; Language Learning

ACM Classification Keywords
H.5 Information interfaces and presentation: User Interfaces

General Terms
Design; Experimentation; Human Factors

INTRODUCTION & RELATED WORK
For a second language learner, the primary difficulty in understanding ongoing speech is the perception of words. This difficulty is especially pronounced for learners of tonal languages such as Mandarin Chinese, in which the meaning of a sound is inferred from both its changing pitch and its context of use. Since most people who learn Mandarin do so from non-tonal native languages and lack sufficient vocabulary to establish linguistic context, they often find the four tones to be the hardest part of language acquisition [8]. Within the four Mandarin tones (Figure 1), new learners often cannot differentiate between the *first* and *fourth* tones because of the unfamiliar pitch range, and the *second* and *third* tones because both end in a rising pitch.

MobileHCI'12, September 21–24, 2012, San Francisco, CA, USA.

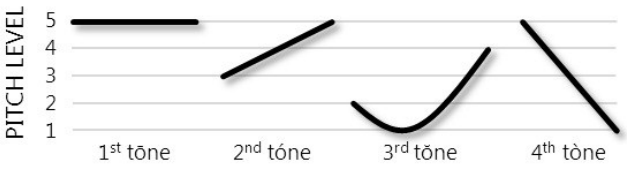

Figure 1. Mandarin Chinese tones on a pitch scale of 1–5.

Tones are not the only difficulty, however. Each syllable is segmented into an *initial* consonant sound followed by a *final* sound which itself can be one or two vowel sounds ending in an optional consonant. There are 23 initials and 34 finals, forming around 1600 tonal syllables in Mandarin Chinese. The importance of understanding this tonal sound system is highlighted in the example of "ma", which when spoken in the first thorough fourth tones "mā", "má", "mǎ", and "mà" (as indicated by the diacritic marks), can refer to the words of mother, hemp, horse, and scold. To complicate further, the Romanized *pinyin* script for pronunciation is idiosyncratic and inconsistent; e.g., "di" and "ci" are pronounced "dee" and "tser" respectively.

As a result of such challenges, the first 2–6 weeks of traditional university-level courses in Chinese are almost exclusively dedicated to the sound system [8]. Since there is no way in which casual language learners can make such commitments, there is a need to support more convenient and accessible learning of the Mandarin sound system.

The standard approach to Mandarin sound instruction is to teach tones, initials, and finals through aural and oral differentiation of minimal pairs (e.g., lā vs. là, zī vs. cī, guǎn vs. guǎng), triplets (e.g., jī, qī, xī), and quadruplets (e.g., mā, má, mǎ, mà). This is grounded in the Speech Learning Model [2], which claims a new phonetic category can only be established if the learner discerns a difference between it and the closest category in their native language. Moreover, adults can only learn novel phonetic contrasts if they re-allocate attention to previously unattended acoustic-phonetic dimensions (i.e., pitch, amplitude, or duration) [3].

Aural perceptual training is effective however, especially when learners are forced to develop and use phonetic codes in short-term memory [3], and the effects have been shown to generalize to novel words and speakers as well as from aural perception to oral production [7]. Recent studies have confirmed that aural training significantly alters brain activation during Mandarin tone perception, showing promising neural plasticity in adult language learners [6].

Figure 2. Tip Tap Tones example: (a) Set speed; (b) *zhen* screen (followed by *zhan*, *chen*, and *chan* screens); (ci). *chan* vs. *zhan* screen with "correct" green flash; (cii) *zhen* vs. *chen* screen with "incorrect" red flash; (d) *chen* vs *chan* vs *zhan* vs *zhen* screen.

MOBILE GAME DESIGN

Inspired by casual mobile games, mobile games for learning Chinese characters [5], and mobile microlearning of location-based language [1], we decided to create a mobile "microtraining" game to help learners master the sound system of Mandarin Chinese in short fragments of free time spread throughout the day.

In crafting the game experience, we drew inspiration from the rhythmic tapping of Tap Tap Revolution and the steadily increasing challenge of Tetris[1]. Both use time pressure and repeated actions to facilitate the kind of concentrated attention that leads to skill development.

TIP TAP TONES GAME

The design of our "Tip Tap Tones" game employs a "press your luck" game mechanic: accurate responses are required to complete a level, but more points are awarded for fast level completion. The game screens are shown in Figure 2:

(a) The learner selects a Mandarin sound playback speed of 1, 2, or 3, with sounds of 1s, 0.5s and 0.25s respectively.
(b) In Level 1, the learner hears a Mandarin sound and must press the button corresponding to the correct tone. The learner must give three correct answers for each sound to progress to a similar sound on the next screen. This repeats for four screens before moving to Level 2.
(c) In Level 2, the learner hears a Mandarin sound and must press the button corresponding to both the correct tone and the correct syllable. The choice is between a minimal pair of syllables from the first level; the learner must give three correct answers for each pair of syllables to progress to the next screen. This repeats for four screens of similar pairs before moving to Level 3.
(d) Level 3 is the same as Level 2, but with a choice between four tones and two minimal pairs of syllables. Six correct answers complete the level and the game.

[1] See www.tapulous.com and www.tetris.com respectively.

Game Mechanic

At a minimum, the learner must give 12 correct answers (from 4 options) over 4 screens to complete Level 1, 12 correct answers (from 8 options) over a further 4 screens to complete Level 2, and 6 correct answers (from 16 options) on the final screen to complete Level 3. A total of 30 correct answers over 9 screens in 60 seconds are thus necessary to complete the game. However, this is not always sufficient. If the learner gives an incorrect response, they have one further chance to give the correct answer and continue their progression. If the learner ever gives two incorrect responses in succession, their *correct count* for the current screen is reset to zero and a new sound is selected at random to start the screen again. Learning is facilitated by replaying sounds using the *repeat* button and by feedback, with a correct response followed by a green flash of the screen and an incorrect response followed by a red flash and "punishing" tactile vibration (Figure 2ci–ii).

Points awarded for correct responses in Levels 1, 2, and 3 are 1, 2, and 3 respectively. In Levels 1 and 2, an additional 5 and 10 points respectively are awarded for completion of each screen. A time bonus is also awarded on completion of each level, calculated as *speed factor* × *time remaining* × *level bonus*. The playback speed factors are 1, 1.5, and 2 respectively for speeds 1, 2, and 3, and the level bonuses are 1, 4, and 16 respectively for Levels 1, 2, and 3.

Rationale

Our design of the core game mechanic draws on established guidelines for educational mobile games [4]. It incorporates *adaptation* by varying the game experience according to performance. Learners who consistently give successive incorrect answers will remain in the screens of Level 1, which provide fewer options and thus support initial tone acquisition. Learners who reach Level 2 must demonstrate an additional awareness of the differences between syllables and their spellings in pinyin. Completing Level 3 requires even finer discrimination of sounds while maintaining both accuracy and speed. Since points increase dramatically with the speed of game completion, we

provide integrated game and learning *goals* that offer a dynamic *challenge* curve leading towards *mastery* [4].

The decision to limit the duration of each game to a short time period was jointly motivated by the mobile context of use and the natural structure of the minimal-difference identification task. In early experimentation, a game duration of one-minute was found to provide novices with sufficient time to complete at least one game screen successfully (unlike a 30 second version), while allowing sustained concentration from more advanced learners without noticeable fatigue (unlike a 2 minute version).

IMPLEMENTATION

We developed the Tip Tap Tones game in Silverlight for Windows Phone 7 devices, in particular the large-screen HTC HD7 shown in Figure 2. Offline processing was used both to generate the sound files played in the game and to create sets of similar syllables requiring differentiation.

Sound Generation

Synthesized speech is easy to produce but has noticeable artifacts. We therefore recorded sounds from a professional, "golden" speaker reading simplified Chinese characters representing all legal tonal syllables in Mandarin. To create both slower and faster versions of these sounds, we performed Pitch Synchronous Overlap Add (PSOLA) timescale modification after tracking the fundamental frequency and marking the pitch epochs of voiced speech. A panel of native Mandarin speakers determined the time-scaled sounds to be natural and free from artifacts, as would be expected from the limited extent of the transformation.

Syllable Selection

Sounds are perceived as similar if they share phonetic components that are the same or similar in their production. For initial sounds in Mandarin, similarity can arise from articulation (position of the tongue) or aspiration (passage of air) or both. Mandarin finals can also share medial semivowels, nucleus vowels, or coda consonants or vowels.

To create pairs of minimal syllable pairs as required by the game mechanic, for each initial and final we crafted the set of other initials and finals that differed in only one phonetic dimension. For each pairwise product of these initial and final sets, we generated lists of all lexical (i.e., meaning carrying) syllable combinations. From these variable length lists, we first removed the syllables that can only take one or two tones. We then extracted all unique "lattices" of four syllables such that two pairs differed only in their initial sound, and two pairs only in their final sound. This resulted in 100 chains covering 1009 tonal syllables, each used to create the four minimal pairs of syllables used in Level 2 of the game. Each game uses a randomly selected lattice. During gameplay, syllable selection is also random as is the selection of tones from those yet to be used on that screen.

EVALUATION

The evaluation of Tip Tap Tones was based on 3 questions:

1. Can we create a *test* of tone and syllable differentiation on which native speakers achieve near-perfect scores?
2. Does the *game* encourage repeat play and facilitate steady improvement for a range of learner abilities?
3. Does *gameplay* improve *test scores* on the identification of tonal syllables that (a) were trained in the game, and (b) were not trained, indicating phonetic generalization?

To answer question 1, we created a test of Mandarin aural perception using the recordings of the golden speaker. This 10 minute test played a series of non-repeatable Mandarin sounds, with the learner identifying either the correct tone or syllable from four similar options. To address research questions 3(a) and 3(b), both sections included not just all syllables trained in the game, but also the special set of syllables that lack an initial sound (but whose pinyin uses *w* and *y* as prefixes for finals beginning with *u* and *i* respectively). Improvements on these syllables would indicate that training through gameplay results in not just the creation of new phonetic categories, but also their generalization to novel contexts. No feedback was given. Validation with native speakers gave near-perfect scores.

To answer questions 2 and 3, we recruited 12 learners of Mandarin Chinese, drawn from both our lab and the local expatriate community (mean age 30, 2 female). All had lived in China for at least several months prior to commencing our study, but none had reached the level of holding basic conversations. Over the course of a 3-week user study, these participants were free to play Tip Tap Tones on a HTC HD7 Windows Phone 7 as much or as little as they desired. We recommended 3–5 minutes of play per weekday, corresponding to 45–75 minutes in total, but compensation (a small gift) did not depend on this.

Participants completed our test at the beginning and end of the study and did not engage in any additional aural training for its duration (aside from occasional pronunciation feedback in weekly Chinese lessons). We validated our test with Mandarin learners of similar background to the study participants, observing stable scores over the course of a week. We concluded that a control group was therefore unnecessary for the study design, which follows that of [3].

Results

Table 1 shows a breakdown of participant performance. Since tests only measure accuracy and not speed, differences can be observed between advanced learners who focus on accuracy with speed (e.g., *P2*, *P9*) and novice learners who focus on accuracy only (e.g., *P1*, *P7*).

Planned two-tailed paired-sample t-tests comparing pre-test and post-test results indicated that tone accuracy increased significantly by 25% for tones trained in the game, with $t_{11} = 4.7$, $p < 0.01$ from means of 57% (*sd* 24%) and 82% (*sd* 13%). A similar 24% improvement in tone accuracy was also observed for syllables that were not trained in the game, with $t_{11} = 4.6$, $p < 0.01$ from means of 58% (*sd* 23%) and 82% (*sd* 17%). Syllable identification also improved by

a significant 12% for syllables trained in the game, with $t_{11} = 3.9$, $p < 0.01$ from means of 68% (*sd* 19%) and 80% (*sd* 17%). Identification of initial-less syllables not trained in the game did not improve, potentially due to their irregular spelling (e.g., the final sound *-ui* becomes the syllable *wei*).

User ID	Game Count	Max Score	Mean Score	CAS Linear Regression		Tone %				Syllable %			
						pre-test		post-test		pre-test		post-test	
				b	R^2	in	not	in	not	in	not	in	not
P1	169	458	157	0.7	0.99	25	29	80	54	52	65	77	75
P2	134	1430	715	4.3	0.88	76	87	82	79	84	90	97	85
P3	104	494	203	1.2	0.92	68	66	86	75	54	80	86	80
P4	78	970	335	3.6	0.98	58	45	95	100	40	60	63	70
P5	66	513	190	1.9	0.93	65	54	88	83	93	75	100	75
P6	60	1042	343	3.8	0.92	89	91	98	95	84	75	100	80
P7	51	152	71	0.7	0.94	17	33	59	62	52	65	56	65
P8	49	1156	350	5.8	0.98	43	70	91	95	84	90	90	85
P9	43	1114	535	10.3	0.85	88	91	95	100	79	80	93	85
P10	39	374	128	1.7	0.86	49	54	77	100	88	80	81	85
P11	33	114	79	1.9	0.95	35	33	59	62	43	80	56	65
P12	22	264	108	1.8	0.43	72	42	70	75	60	66	61	85
mean (*sd*)	71 (44)	673 (444)	268 (198)	3.1 (2.7)	0.89 (0.15)	57 (24)	58 (23)	82 (13)	82 (17)	68 (19)	76 (10)	80 (17)	80 (8)

Table 1. User study game use, cumulative average score (CAS) linear regression (coefficient *b* and adjusted R^2), and pre/post test result for sounds trained "in" the game and "not"

Figure 3 shows a rising *cumulative average score* (CAS) for all learners, indicating improvements in accuracy and speed across sessions. Linear regression gave a strong fit (adjusted $R^2 > 0.85$) for 11/12 learners (see Table 1).

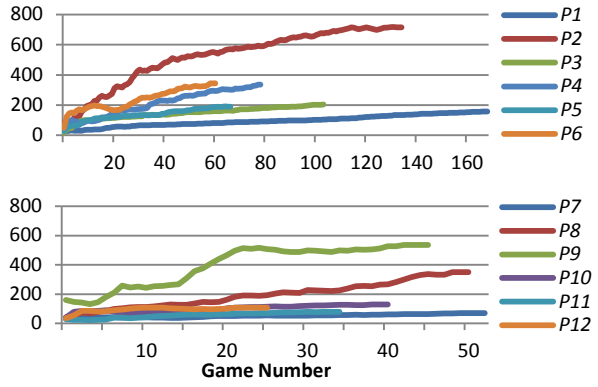

Figure 3 Cumulative average scores over all games played

Qualitative analysis of post-study interviews confirmed tones as the "hardest part" [*P3*] of learning Chinese. All participants thought they had improved, being able to "hear the difference" [*P4*] and "listen quicker" [*P8*]. The minute duration was "great for casual learners" [*P6*], keeping them focused [*P3*] and filling free time [*P1*]. Combined with the feeling of being rewarded [*P6*] by rising scores [*P4*], this time pressure made the game addictive [*P2*]. The increasing "levels of complexity" [*P11*] also made learners pay attention to details they would otherwise miss [*P3*].

Beyond the game, learners reported being better able to listen to tones by "thinking in terms of the Tip Tap Tones screen" [*P8*], read pinyin by "mapping to sounds" [*P10*], and speak by picking out native speaker tones and reproducing them "right away, in the right way" [*P4*].

DISCUSSION & CONCLUSION

Overall, results show that Tip Tap Tones is effective at training learners to correctly identify Mandarin sounds. Large gains in test accuracies and game scores from small time investments suggest mobile microtraining can support efficient and convenient skill development on the move.

Future directions for second-language training include expanding to a "high variability paradigm" with voices from many native speakers (as in [3] and [7]), employing phonetic analysis to mine minimal sound sets in all languages (tonal and otherwise), and using speech analysis for oral pronunciation training. Similar game mechanics could also be useful in domains beyond aural language training (e.g., aural identification of notes for musicians, visual identification of colors for designers, or haptic identification of Braille codes for the visually impaired).

The broadest contribution of this paper is as a case study of how mobile HCI can transform traditional activities by transporting them to new contexts. In particular, we have shown how our design of mobile microtraining has transformed a slow-paced, low-feedback drill into a fast-paced, high-feedback, learner-driven game playable anywhere. Tip Tap Tones is available as a free Microsoft Research application on the Windows Phone marketplace.

ACKNOWLEDGMENTS

We thank all of our testers, participants, and reviewers.

REFERENCES

1. Edge, D., Searle, E., Chiu, K., Zhao, J. & Landay, J. (2010). MicroMandarin: Mobile Language Learning in Context. CHI 2011, 3169–3178
2. Flege, J.E. (2007). Language contact in bilingualism: phonetic system interactions. Laboratory phonology, 9, 353–382
3. Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. Journal of the Acoustical Society of America, 89(2), 874–886
4. Thomas, S., Schott, G. and Kambouri, M. (2003). Designing for Learning or Designing for Fun? Setting Usability Guidelines for Mobile Educational Games. MLEARN 2003
5. Tian, F., Lv, F., Wang, J., Wang, H., Luo, W., Kam, M., Setlur, V., Dai, G. & Canny, J. (2010). Let's play chinese characters: mobile learning approaches via culturally inspired group games. CHI 2010, 1603–1612
6. Wang, Y., Sereno, J., Jongman, A. & Hirsch, J. (2000). Cortical reorganization associated with the acquisition of Mandarin tones by American learners: An fMRI study. Proc. 6th Int. Conf. on Spoken Language Processing, II, 511–514
7. Wang, Y., Spence, M.M., Jongman, A., & Sereno, J.A. (1999). Training American listeners to perceive Mandarin tones. J. of The Acoustical Society of America, 106(6), 3649–3658
8. Xing, J.Z. (2006). Teaching and Learning Chinese as a Foreign Languge. Hong Kong University Press