International Chinese Language Education Communications
国际中文教育通讯
ISSN 3078-3348

International Chinese Language Education Communications
Volume 3, Issue 1, 78-103
https://doi.org/10.46451/iclec.260301

Received: 25 December 2025
Accepted: 12 February 2026
Published: 2 March 2026

# Teaching Mandarin Tone Listening with a Pitch-Line Display: Materials and Classroom Reflections

Xinsheng Cao
University College Dublin, Ireland
(Email: morantimeday@gmail.com/xinsheng.cao@ucdconnect.ie)

## Abstract

Mandarin lexical tones are difficult for beginners from non-tonal language backgrounds because learners do not automatically treat pitch movement as a stable cue for word meaning. In Arabic-speaking classrooms, the key risk may surface in question framing (here, Mandarin ma yes/no questions). Interrogative prosody may increase cue competition for lexical pitch cues, so this is treated as a classroom stress test rather than a clean sentence-final rise manipulation. This teaching-focused article presents a classroom routine that uses a simple pitch-line display (a Praat-based pitch-contour display; hereafter, pitch-line display) to guide learners' attention to contour shape and turning points. The visual support is based on Praat-derived F0 contours, with light display-level enhancement for readability (e.g., smoothing, redrawing, and adding labels and directional arrows). It functions as an instructional scaffold rather than automated feedback or an analytic tool. The article provides (a) design principles for building instructionally usable contour visuals, (b) a brief teacher-led micro-teaching script that links gesture and contour cues, and (c) a task progression that moves from isolated syllables to statement and yes/no question sentences. The final section offers practical reflections on common confusions, especially those involving rising cues in questions, and describes teacher moves that help learners re-focus on lexical tone cues. The goal is to offer classroom-ready materials that can be implemented within normal teaching time.

## Keyword

Mandarin lexical tones, pitch-line display, visual scaffolding, Arabic-speaking learners, classroom listening tasks

## 基于音高线显示的普通话声调听辨教学：材料设计与课堂反思

曹新胜
都柏林大学，爱尔兰

### 摘要

普通话声调对非声调语言背景的初学者常常很难。学习者往往不会把音高变化当作稳定的“词义线索”。在阿语课堂里，这个问题在疑问句语境中可能更突出。以普通话“吗”是非问句为例，句末上扬的疑问语调容易和词汇声调的音高线索发生竞争，学习者会更依赖“像不像问句”的感觉，而忽视词汇层面的声调信息。基于这一点，本文把疑问句语境当作课堂中的“压力测试”，而不是简单的句末上扬操控。

本文提出一套可直接嵌入课堂的例行教学流程，核心工具是一种简易的音高线显示（pitch-line display），用于引导学习者关注轮廓形状与转折点。该显示基于 Praat 提取的 F0 轮廓，并做了轻量的课堂化处理（如平滑、重绘，以及添加标签与方向箭头），目的在于支持教学，而不是提供自动反馈或技术分析。文章提供三部分可用材料：（a）轮廓可视化材料的设计原则；（b）一段简短的教师主导微型教学脚本，用于把手势线索与轮廓线索连起来；（c）一个从单音节到陈述句与是非问句的任务递进序列。最后，本文结合课堂观察与反思，总结学习者常见混淆点，尤其是疑问句上扬线索导致的误判，并归纳教师在课堂中帮助学习者重新聚焦词汇声调线索的具体做法。本文的目标是提供一套低负担、可操作的声调听辨教学材料，便于在常规课时中使用。

### 关键词

普通话词汇声调，音高线显示，视觉支架，阿语背景学习者，课堂听力任务

### Introduction

Mandarin lexical tones are a persistent bottleneck for beginners from non-tonal language backgrounds. Early on, many learners do not treat pitch movement as a stable cue for word meaning. They often rely on loudness, duration, or overall pitch height first. Only over time do they shift attention toward contour shape and turning points. A useful way to frame this is cue-based. Learners do not simply memorise “Tone 1–Tone 4”. Instead, they gradually learn which acoustic cues to trust, and they shift their attention as tonal categories take shape (Schertz & Clare, 2020; Wang et al., 2020). In classrooms, this re-weighting is shaped by cue competition and limited input (Chandrasekaran et al., 2016; Hao, 2012).

For Arabic-speaking learners, the classroom challenge can be stronger because of familiar L1 prosodic routines. Arabic does not use lexical tone to mark word meaning. However, many varieties use intonation—often with salient rises—to signal yes/no questions and other pragmatic meanings. These routines can pull attention toward sentence-level meaning cues rather than lexical pitch categories, especially when learners first meet Mandarin tones through brief classroom exposure (Almalki & Morrill, 2016; Chahal & Hellmuth, 2014; Hellmuth, 2022). In practical terms, learners may need to build new tone categories while managing well-established intonation habits. This combination can reshape confusion patterns and produce a learning profile that does not match a simple “difficulty ranking” (Hao, 2012).

This reality creates a teaching problem. Teachers need usable ways to see which tonal categories stabilise first, which remain fragile, and which listening contexts trigger breakdown. Early on, tone listening is uneven. One or two tones may settle sooner, while others stay shaky.

Some categories may become internal reference points, while others remain unstable. Short instructional episodes can also produce visible shifts, but those shifts may not last. Classroom evidence therefore needs to be read as patterns of system organisation, not as a fixed tone-by-tone hierarchy (Wang et al., 2020).

A second teaching issue is tone–intonation interaction in real listening. Interrogative rise does not affect all lexical tones in the same way. Under a cue-competition view, interference should be selective, especially when an intonational rise overlaps with a fragile lexical rising category. In class, a rise can sound like “this is a question,” or it can be part of a lexical tone. Which one learners hear depends on what they are used to and what the task asks them to do (Braun & Johnson, 2011; Gussenhoven, 2004; Xu, 2005). In this routine, yes/no questions were realised with ma, so question framing also changes sentence structure and target position. For that reason, the question condition is treated as a classroom “stress test” context, not as a clean sentence-final rise manipulation. In addition, because ma is a sentence-final neutral-tone particle, it may also shape the phonetic realisation of the preceding target syllable (e.g., positional reduction or coarticulation), which is another reason to treat statement–question differences as context-sensitive rather than as a pure intonation effect.

A third need is practical: what does a pitch-line visual actually do while students listen? Teachers often use contour visualisation to make pitch movement visible and to focus attention (Chun & Jiang, 2025; Hardison, 2004). In many schools, this kind of support is easily misread as “AI feedback.” Here I use display-level enhancement to mean cleaning up Praat-derived contour images for classroom readability (e.g., smoothing and adding simple labels/arrows). It does not score learners, generate personalised feedback, or change the audio. The evidence here comes from learners’ task responses within the classroom sequence, not from the visual clean-up itself (Li et al., 2024). What remains under-described is what learners report noticing when contour visualisation is used in real classroom listening, and whether this noticing aligns with the cue re-weighting expected during early system formation (Schertz & Clare, 2020; Chandrasekaran et al., 2016).

Beyond this local setting, the Saudi school context is useful because it makes a common classroom problem easy to see: learners may treat a sentence-level rise as a meaning cue and let it compete with lexical tone cues. In many new or still-developing Mandarin programmes, teachers need tools that work inside normal lessons, without lab equipment or heavy testing. A short routine that combines quick diagnosis, a brief cue reminder, and immediate follow-up tasks can therefore support curriculum building in other early-stage programmes as well. This teaching note uses the Saudi classroom to show a simple “check–teach–check” routine that other early-stage programmes can run for tone listening.

Guiding questions for teachers in similar settings:
(1)Which tones become reliable first across common classroom formats?
(2)Which contexts predictably trigger breakdown (especially question framing)?
(3)When a brief cue reminder is used, what changes immediately—and what does not?

Against this background, this teaching note reports a materials-based classroom inquiry conducted within normal instruction in a newly established Mandarin programme in Saudi Arabia. It offers a replicable routine that combines a quick diagnostic, a brief teacher-led cue reminder (gesture + pitch-line display), and immediate follow-up listening tasks. It also describes how statements and yes/no questions were used as a practical “stress test” for context-sensitive vulnerability. Finally, it reflects on what the classroom evidence suggests—and what it cannot support—about attention guidance under cue competition in early tone learning. The aim is not to claim a controlled causal "visualisation effect," but to provide practice-facing, evidence-informed guidance for teachers working in similar contexts (Almalki & Morrill, 2016; Braun & Johnson, 2011; Wang, 2012; Hellmuth, 2018).

## Teaching Context and Lesson-Package Design

### Teaching setting and learners

This teaching note is based on a Grade 7 beginner Mandarin programme in a Saudi public middle school where Mandarin was newly introduced as a school subject. Learners were Arabic L1 speakers and reported no prior experience with a lexical tone language. At the time of the lesson-package implementation, they had studied Mandarin for about 2–3 months through regular instruction (three lessons per week). A total of 176 students were enrolled in the programme. For the item-level reconstruction, 126 students produced complete responses across the three classroom listening blocks (16 items per block; see Table 1). Incomplete response sets mainly reflected absence or partially missed items during normal class time.

Table 1
*Class Profile and Teaching Context*

| Category | Description |
|---|---|
| Total students in programme | 176 |
| Valid complete task responses (Blocks A–C) | 126 |
| Age range | 12–13 years |
| L1 background | Arabic |
| Experience with tonal L2s | None reported |
| Mandarin learning experience | Approx. 2–3 months (beginner) |
| Instructional hours | 3 lessons per week |
| Learning context | Public middle school (Eastern Saudi Arabia) |

### Design Rationale: A Short Diagnostic-to-scaffold Sequence

The package was designed as a routine classroom sequence, not as a separable lab-style intervention. It combined (a) a baseline listening check, (b) a brief teacher-led micro-teaching episode supported by pitch-line display and gesture, and (c) follow-up listening under two common classroom pressures: clearer pitch input and sentence-level contexts. The purpose was practical: to make learners' early tone-category organisation visible, and to test whether a short, explicit attention cue (contour + gesture) could momentarily re-orient listening toward pitch movement (Schertz & Clare, 2020; Chandrasekaran et al., 2016; Chun & Jiang, 2025; Hardison, 2004).

The reason to select the Tone 2–Tone 3 contrast is because this is a frequent early listening confusion and it is easy to test with a two-choice check. For Arabic-speaking beginners, question rise can mask a lexical rise (Tone 2), so contrasting Tone 2 with Tone 3 supports a compact cue ("rise onset vs. turning point") that fits brief in-class teaching. This narrow focus keeps the micro-teaching feasible, while Blocks A–C still cover T1–T4 for broader diagnosis.

### Materials package

Tone targets and syllables. All listening items used a small set of short syllables (e.g., ma, mi, na, ba, ni) combined with T1–T4. This restricted syllable set reduced vocabulary demands and kept the focus on tone perception rather than word knowledge. Blocks A and B used isolated-syllable tokens, while Block C embedded the same tone targets in short sentences to introduce prosodic context. Full item scripts and answer keys are provided in Appendix 1. The initials (b/m/n) and simple vowel finals (a/i) were chosen to keep segmental complexity low and reduce consonant/vowel confounds, so learners' attention could stay on pitch movement in a beginner classroom context.

Exaggerated-F0 input (Block B). Block B used re-recorded stimuli with an intentionally expanded pitch range and clearer turning points. It did not use algorithmic F0 resynthesis. Block B was included as a classroom diagnostic (a robustness check): it tests whether clearer turning points alone can raise identification accuracy without changing task format or response demands. In this teaching note, it is treated as a classroom-oriented stimulus variation within the same flow, rather than a tightly controlled acoustic manipulation.

Answer sheet and classroom script. Learners marked answers on a four-option listening sheet (T1–T4). The standard instruction was "listen twice" for each item before responding. The blank answer sheet is provided in Appendix 2.

### Audio Preparation and a "minimum control" Approach for Statements vs. Questions

All audio stimuli were recorded by the author for classroom playback. The recordings contain only my voice; no student audio or video was collected. For the statement–question pairs, I used separately recorded statement and yes/no question versions rather than post-hoc pitch manipulation. I exported the files as standard classroom audio and did not apply loudness normalisation or additional audio processing beyond default recording/export settings.

To reduce obvious non-tonal confounds in the statement–question comparison, each statement–question pair was checked on three non-F0 dimensions: sentence duration, speech rate (s/syllable), and mean intensity (Praat intensity, dB). Before screening, narrow acceptability ranges for these measures were set and then applied consistently across all pairs. If a pair fell outside these ranges, the item was re-recorded or removed from the comparison set. This step reduces (but does not eliminate) residual variability, so statement–question differences are treated as context-sensitive classroom patterns rather than a clean estimate of an "intonation effect."

### In-class Implementation Sequence

I ran the routine during normal class time. Each main block took about 10 minutes. I inserted a short micro-teaching segment after Block A and administered Micro-test M immediately afterwards.

Block A (baseline). Learners completed 16 isolated-syllable identification items. I played each item twice. Learners selected one tone label (T1–T4) on the answer sheet.

Micro-teaching (about 8 minutes): pitch-line display + gesture as an attention cue. After Block A, I ran a short, scripted mini-lesson (about 8 minutes). I projected a Praat pitch-line display (a Praat-based pitch-contour display) with light display-level enhancement for readability, and used simple hand gestures to make pitch movement easy to see (Boersma & Weenink, 2025; Li et al., 2024). I kept the message narrow: Tone 2 = rise onset; Tone 3 = low dip/turning point. I selected the Tone 2–Tone 3 contrast in advance because it is a common classroom confusion point where a "rising" impression can mask category-relevant turning points, making it suitable for a short, contrastive attention cue (Schertz & Clare, 2020; Chandrasekaran et al., 2016). The goal was to pull attention toward contour shape and turning points rather than loudness or duration. I treated this as a classroom scaffold, not an experimental “treatment.” Appendix 4 includes the visual and the 8-minute script for direct reuse.

Micro-test M (immediate post-teaching check). Immediately after the micro-teaching, learners answered four fast contrast items targeting Tone 2 vs. Tone 3 (two-alternative). Because it is short and contrast-specific, I treat Micro-test M as a time-local attention check rather than as evidence of general tone learning.

Block B (exaggerated-F0 input). Learners then completed 16 isolated-syllable items using the exaggerated-F0 recordings. During this phase, I did not add extra explanation; the purpose was to provide clearer contour exemplars within the same task format, so the follow-up sentence block could be interpreted against a “clearer-input” reference point rather than as a separate treatment condition. I played each item twice.

Block C (sentence-level items: statements vs. yes/no questions). Finally, learners completed 16 sentence-level items, including statements (C_Q1–C_Q8) and yes/no questions (C_Q9–C_Q16). In Block C, learners identified the lexical tone of the target syllable inside the sentence frame (e.g., 这是 má/mǎ …), while yes/no questions ended with the particle ma, so statement–question differences are treated as context-sensitive classroom vulnerability under question framing rather than a clean intonation manipulation.

### Learner check-in notes

After the routine, I conducted ten brief check-in chats (about 1–2 minutes each). I used seven prompts (grouped into three parts) to capture what learners noticed, what confused them, and how they used the pitch-line visual and gestures. The full prompt is provided in Appendix 3.

### Ethics and permissions

This teaching note reports aggregated, de-identified classroom task responses collected within normal instruction. School-level permission was obtained before any data were used for reporting. Participation in the evidence component was voluntary and had no graded consequences. Because learners were minors, participation followed a school-authorised consent procedure implemented via school channels: guardians were informed and could opt out, and students could decline or stop at any time. No student audio or video was recorded; only the teacher’s stimuli were used for classroom playback. All records were anonymised at source, stored securely, and reported only in group form.

## Classroom Snapshot: What the Routine Showed Me

### From baseline to follow-ups: A quick read of the overall pattern

Across Blocks A–C, overall accuracy stayed around 0.50 (0.49–0.51), so learners were correct on about half of the items (see Table 2 and Figure 1). Sentence items did not lower accuracy compared with isolated syllables. By contrast, the immediate post-teaching Micro-test M was higher (about 0.68; see Table 2). Because Micro-test M was a two-choice T2–T3 check (chance = 50%), it is not directly comparable to the four-choice blocks (chance = 25%) and is reported here as an immediate attention check rather than a general learning outcome.

Because Micro-test M included only four items and restricted the response set to Tone 2 vs. Tone 3, the higher score is interpreted as a targeted, format-specific check administered immediately after the micro-teaching episode, rather than as evidence that general tone identification improved across tasks.

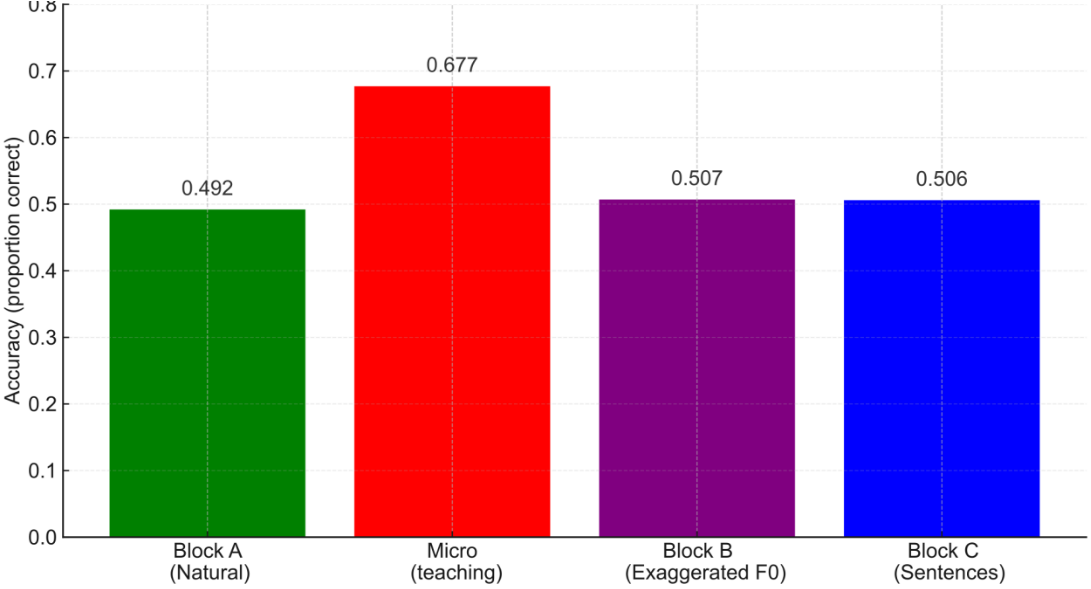
Table 2

*Overall Correct Identification across the Routine Stages (Blocks A–C) Plus the Immediate Contrast Check (Micro-test M).*

| Block | Accuracy (proportion) |
|---|---|
| Block A (natural isolated syllables) | 0.492 |
| Micro-test M (4 items) | 0.677 |
| Block B (exaggerated F0 isolated syllables) | 0.507 |
| Block C (sentence tone identification) | 0.506 |

Figure 1

*Overall Correct Identification across the Routine Stages (Blocks A–C) Plus the Immediate Contrast Check (Micro-test M).*



Practice-facing reading. For teachers, this pattern suggests that a short micro-teaching episode may be followed by a visible “lift” on a narrowly defined contrast check, while broader listening performance can remain flat. This is useful for lesson planning: immediate checks can confirm attention orientation, but they do not substitute for delayed or cross-format evidence.

### Tone-level reliability: What stayed steady, what did not

Although overall accuracy was stable, tone-level behaviour was clearly uneven across tasks. A simple look across Blocks A–C suggests Tone 3 stayed highest and changed the least when formats shifted (see Table 3 and Figure 2). Across blocks, Tone 3 stayed in the mid-0.60s, and its trajectory remained relatively steady.

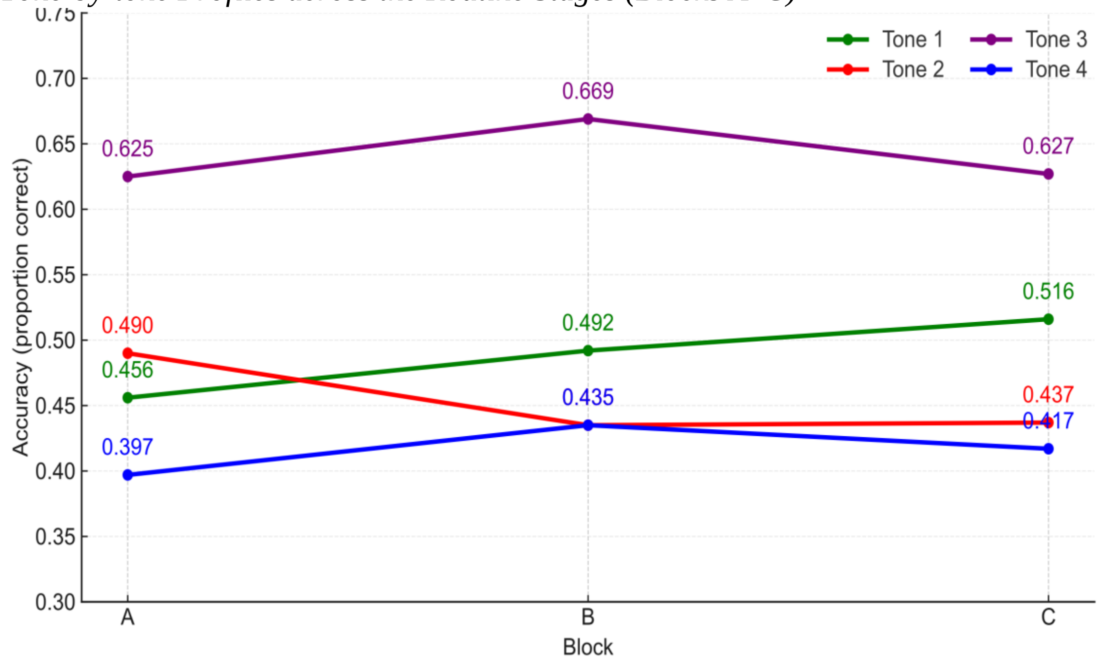
Table 3

*Tone-by-tone Correct Identification across the Routine Stages (Blocks A–C).*

| Tone | Block A | Block B | Block C |
|---|---|---|---|
| Tone 1 | 0.456 | 0.492 | 0.516 |
| Tone 2 | 0.490 | 0.435 | 0.437 |
| Tone 3 | 0.625 | 0.669 | 0.627 |
| Tone 4 | 0.397 | 0.435 | 0.417 |

Figure 2

*Tone-by-tone Profiles across the Routine Stages (Blocks A–C)*



By contrast, Tone 2 showed the most unstable profile across formats. Its accuracy stayed in the mid-0.40s, and it varied more when the task context changed (Table 3; Figure 2). Tone 1 and Tone 4 fell between these extremes, showing lower overall accuracy than Tone 3 but less volatility than Tone 2.

Practice-facing reading. For classroom diagnosis, it is more informative to ask "which tone stays reliable when the format changes?" than to rely on a single "difficulty ranking." In this dataset, Tone 3 behaved like a relatively stable reference point, whereas Tone 2 showed the clearest instability across formats.

**Questions as a classroom stress test (where accuracy dropped most)**

Within Block C, statements were more accurate than yes/no questions at the global level (52.2% vs. 48.0% correct; see Table 4). At the tone level, the statement–question gap differed across tones. Tone 2 showed the largest disruption, dropping by roughly 0.09 from statements to questions. Tone 3 also dropped, but by a smaller margin (roughly 0.04), and it remained the highest tone in both sentence types. Tone 1 and Tone 4 showed smaller gaps (Table 4; Figure 3).

Table 4

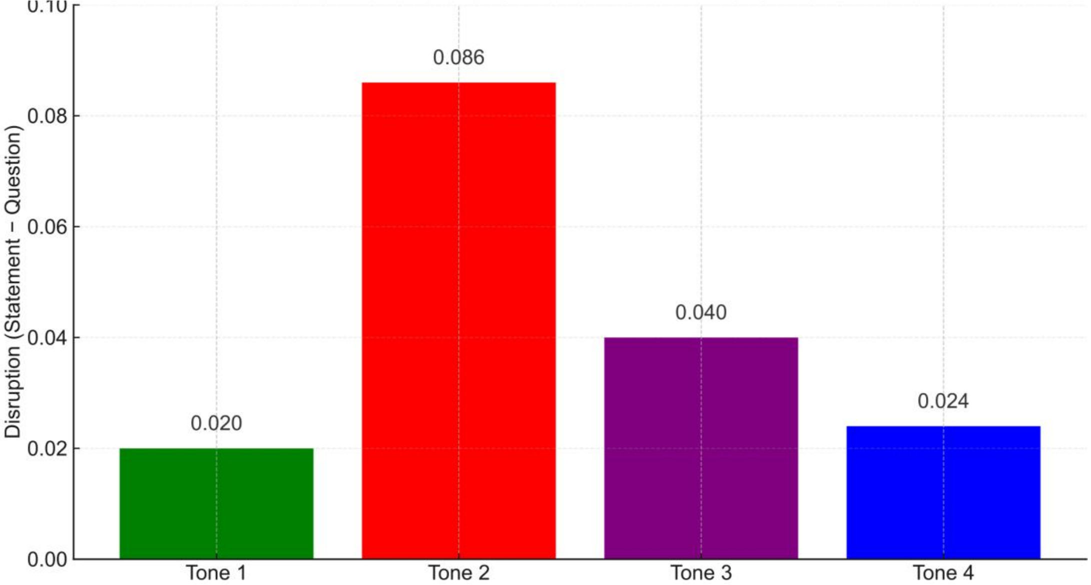
*Sentence Block (Block C): Statements vs yes/no Questions and the Statement–Question Gap (Δ = statement − question).*

| Tone | Statement accuracy | Question accuracy | Disruption (statement − question) |
|---|---|---|---|
| Tone 1 | 0.524 | 0.504 | 0.020 |
| Tone 2 | 0.488 | 0.402 | 0.086 |
| Tone 3 | 0.647 | 0.607 | 0.040 |
| Tone 4 | 0.429 | 0.405 | 0.024 |
| Overall (mean across tones) | 0.522 | 0.480 | 0.042 |

*Note.* Yes/no questions were formed with the particle ma, so this statement–question contrast is treated as a classroom “stress test” under question framing rather than a clean manipulation of sentence-final rise.

Figure 3

*Statement–question Gap by Tone in the Sentence Block (Block C) (Δ = statement − question).*



Importantly, this statement–question contrast should not be treated as a clean intonation manipulation. The yes/no questions were formed with the particle ma, which changes sentence structure and shifts the target syllable away from the utterance-final position; these framing and positional differences may interact with interrogative prosody. Although statement–question pairs were screened to be comparable on basic non-F0 dimensions, F0 contours were not experimentally manipulated. Because ma can affect the preceding syllable (especially reduced Tone 3), I interpret the statement–question contrast as a question-framing stress test rather than a clean prosodic effect. For these reasons, the disruption pattern is interpreted as context-sensitive vulnerability under question framing (consistent with selective cue competition), rather than as a definitive causal effect of sentence-final rise alone.

Practice-facing reading. For teachers working with Arabic-speaking beginners, question contexts may be a predictable “stress test” for the lexical rising category (Tone 2). A practical implication is that sentence practice should not be reserved for “later.” Instead, teachers may need to introduce sentence framing early as a controlled challenge, with explicit cue reminders to keep lexical tone in focus.

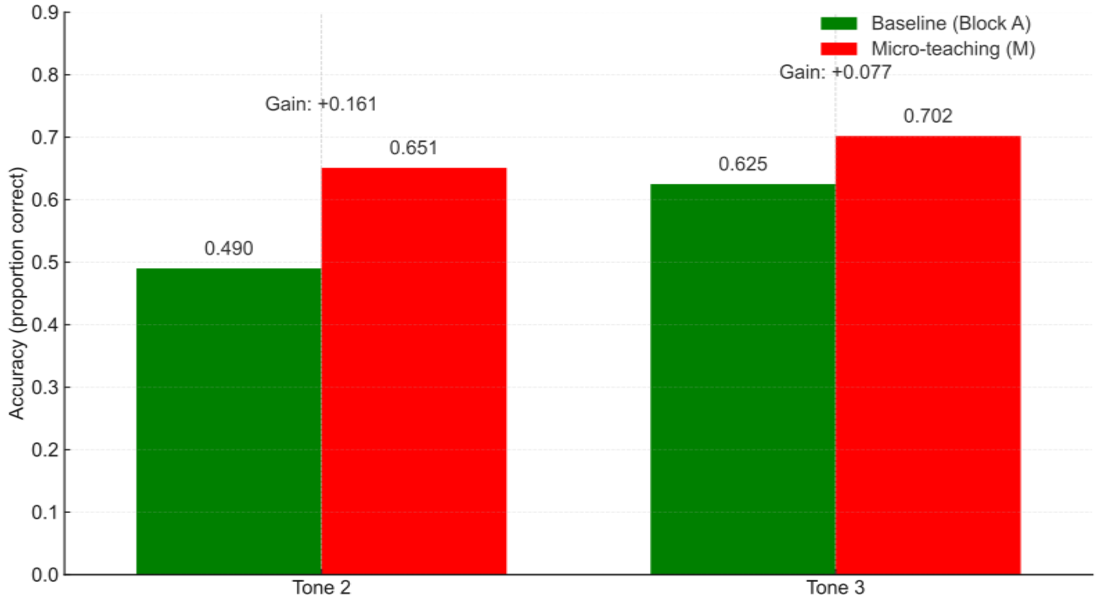
**A short post-teaching contrast check (what it can and cannot tell us)**

Immediately after the teacher-led micro-teaching episode, learners completed a four-item Micro-test M targeting the Tone 2–Tone 3 contrast. This check produced higher accuracy than the baseline pattern for the same contrast. For Tone 2, the immediate shift was about +0.16, and for Tone 3 it was about +0.08 (see Table 5 and Figure 4).

Table 5
*Immediate Contrast Check after the Cue Reminder: Tone 2 vs Tone 3 (Baseline Block A vs Micro-test M).*

| Tone | Baseline (A) | Micro test (M) | Absolute gain (M−A) | Relative gain |
|---|---|---|---|---|
| Tone 2 | 0.490 | 0.651 | +0.161 | 32.9% |
| Tone 3 | 0.625 | 0.702 | +0.077 | 12.3% |

Figure 4
*Immediate Contrast Check after the Cue Reminder: Tone 2 vs Tone 3 (Baseline Block A vs Micro-test M).*



However, because the Micro-test M (a) contained only four items, (b) restricted the response set to two categories, and (c) was administered immediately after instruction, it is interpreted as a time-local, contrast-specific indicator rather than as evidence of broad or durable improvement in tone identification.

Practice-facing reading. For classroom use, this type of rapid contrast check can confirm that learners noticed the intended cues (e.g., rise onset for Tone 2 and the turning point for Tone 3) at that moment. It should not be treated as a substitute for later recycling across formats (isolated syllables, statements, and yes/no questions) or for delayed checks.

**What students said in check-ins (common points)**

Post-task check-ins (n = 10) suggested four classroom-relevant points. First, learners commonly reported relying on non-pitch cues (e.g., “long/short” and “loud/soft”) more than pitch height or contour. Second, learners most frequently selected teacher hand gestures as the most immediately helpful scaffold, while describing the on-screen pitch-line display as easy to see and useful for noticing small differences. Third, most learners (8/10) reported that question sentences were harder than statements and linked this difficulty to rising pitch (e.g., “it makes the tone harder to hear”). Fourth, learners described the pitch-line display mainly as helping them notice differences and construct a contour image, rather than as a motivational “fun” feature; several learners also reported that they would listen differently after using the pitch-line display. Only the third point was asked as a fixed comparative question and tallied as a count; the other points are thematic summaries from brief check-in notes.

Table 6

*Quick Check-in Notes (n = 10): What Learners Focused on, How Often It Came up, and Example Quotes.*

| Theme | Prevalence | Learner-reported pattern (summary) | Related quantitative pattern |
|---|---|---|---|
| Theme 1: Cue preference (duration/intensity > F0) | Theme summary (notes; not tallied) | Learners report attending to "long/short" and "loud/soft" more than "high/down". | Learner notes suggested attention to duration/intensity cues; overall accuracy remained near 0.50 across Blocks A–C, with tone-level differences persisting across tasks. |
| Theme 2: Scaffolds (gesture > example word > pitch-line display; pitch-line display = high precision) | Theme summary (notes; not tallied) | Teacher hand gestures are most frequently chosen; pitch-line display is described as "easy to see/straightforward". | Micro-teaching check accuracy was higher than Blocks A–C; learners described gesture as the most immediately helpful support and the pitch-line display as useful for noticing small differences. |
| Theme 3: Sentence-level difficulty (questions harder) | 8/10 learners | 8/10 learners reported that question sentences were harder; the rising ending made the target tone harder to hear. | In Block C, questions were less accurate than statements overall (48.0% vs. 52.2%); Tone 2 showed the largest disruption (Δ = 0.086). Learners also reported questions were harder. |
| Theme 4: Function of pitch-line display (attention re-orientation) | Theme summary (notes; not tallied) | Learners say it "helps me recognize differences", "see more clearly", and "construct an image". | Learners described the pitch-line display mainly as helping them notice differences ("see more clearly") rather than as "fun"; the micro-teaching check showed higher accuracy (0.677) on a short contrast-focused set. |

Table 6 above summarises these common points and aligns them with the quantitative patterns observed in the routine. Taken together, the check-ins provide descriptive alignment with the classroom snapshot: Tone 3 showed the highest and most stable accuracy across tasks, while Tone 2 showed the clearest cross-format and question-related vulnerability. Learner reports suggest that the pitch-line display functioned primarily as an attention re-orientation tool, with gesture serving as the most immediate scaffold in real time.

### Interpretation boundaries for a practice-based report

Because the routine was embedded in instruction and implemented as a fixed classroom sequence, the evidence in this section is reported as descriptive classroom patterns. Immediate post-teaching shifts reflect time-local performance and are not treated as durable learning effects. Accordingly, interpretation is framed with cautious language (e.g., was followed by, is consistent with, may indicate). Stronger causal claims are reserved for future designs that include delayed checks and systematic comparisons of scaffolding conditions. The statement–question comparison is also structurally constrained by the use of ma (and the resulting positional context), so question-related drops are interpreted as classroom vulnerability under question framing rather than as a controlled prosody contrast.

## Reflection and Teaching Takeaways

### Tone 3 as a usable “reference point” in an emerging system

In this routine, Tone 3 showed the highest and most stable identification across Blocks A–C and remained the most reliable tone when formats changed. This suggests that Tone 3 may settle earlier as a usable reference point in an emerging classroom listening system, rather than reflecting a general “difficulty ranking.”

(1) What the pattern suggests (cautious mechanism).

Tone 3 stayed the most stable category across formats in this routine. This is consistent with a structured-perception view, where early tone learning does not develop as a uniform “difficulty ranking,” but as an uneven system with one or two categories settling earlier than others (Hao, 2012; Schertz & Clare, 2020). In this dataset, Tone 3 may function as a practical reference point that learners can rely on when task demands change and cue competition increases (Chandrasekaran et al., 2016; Qin et al., 2024). This interpretation describes a classroom pattern rather than a fixed rule.

(2) What the teacher can do next week (actionable moves).

Use Tone 3 as the lesson “calibration key.” Start each practice cycle with a short Tone 3 reset (2–3 minutes). Keep the teacher cue language stable: “Find the low dip, then the turn.” Then run contrast work that keeps Tone 3 in the pair (first T2–T3, later T3–T1 or T3–T4). When learners revert to “long/short” or “loud/soft,” interrupt early and redirect attention back to contour shape. The goal is not more explanation. The goal is consistent cue selection.

(3) What material to reuse (Appendix/slide/task pointers).

Reuse the Tone 3 contour slide in Appendix 4 as the fixed visual reference. Keep the same layout each week to reduce cognitive load. For practice, reuse the Block A/B syllable items and Block C sentence frames in Appendix 1, with the same answer-sheet format in Appendix 2. This makes weekly comparisons interpretable and keeps changes attributable to learning rather than new materials.

**Tone 2 as the "fragile" category under cue competition**

(1) What the pattern suggests (cautious mechanism).

Tone 2 showed the clearest instability, especially under sentence framing and in yes/no questions. This is compatible with cue competition: learners may not prioritise pitch early, and a rising category can be especially vulnerable when other rises in the signal carry strong pragmatic meaning (Braun & Johnson, 2011; Chandrasekaran et al., 2016; Schertz & Clare, 2020). In an Arabic-speaking environment, familiar question intonation patterns may further pull attention toward sentence meaning rather than lexical pitch cues (Almalki & Morrill, 2016; Chahal & Hellmuth, 2014; Hellmuth, 2016; Rifaat, 2021; Xu, 2013). This should be read as context-sensitive vulnerability, not as "Tone 2 is always the hardest."

(2) What the teacher can do next week (actionable moves).

Treat Tone 2 as a high-risk target and teach it with "protected practice" before "controlled stress tests." First, do short T2–T3 contrasts with a single fixed cue reminder ("Listen for the rise start"). Next, move into short statement sentences where the target syllable is clearly identifiable. Only then introduce a small set of question items as a planned challenge. Before the question set, give a one-line warning cue: "Question meaning is not your answer. Tone is your answer." If the class collapses, do not extend the question set. Reset with one contrast item, then return to two question items and stop.

(3) What material to reuse (Appendix/slide/task pointers).

Reuse the statement and yes/no question frames in Appendix 1 as your built-in "stress test" set. Keep the same response format from Appendix 2 so learners focus on listening rather than task rules. If you want fast diagnostic feedback, reuse two short prompts from Appendix 3 (what cue they used; whether questions felt harder) as a one-minute exit check with a small sample.

**Micro-teaching as short-term cue focusing, not "general improvement"**

(1) What the pattern suggests (cautious mechanism).

The post-episode Micro-test M showed a contrast-specific spike, while overall accuracy across Blocks A–C stayed largely flat. This pattern is consistent with the idea that a brief scaffold can redirect attention in the short term, especially when the task is narrow and immediately follows instruction (Hardison, 2004; Baills et al., 2019; Chun & Jiang, 2025). It does not, by itself, support a claim of durable or general improvement. In this report, the Micro-test M is treated as a time-local indicator of attention orientation rather than evidence of durable learning.

(2) What the teacher can do next week (actionable moves).

Keep the micro-teaching short and scripted. Limit it to two cues only: Tone 2 rise onset, and Tone 3 dip/turning point. Use the same gesture and the same visual each time. Then use a two-check routine: an immediate contrast check (like Micro-test M) and a delayed mini-check (4–6 items) later that day or in the next lesson. The delayed check is the smallest upgrade that prevents over-interpreting immediate spikes. Do not add more talk. Add one delayed checkpoint.

(3) What material to reuse (Appendix/slide/task pointers).

Reuse the contour visual in Appendix 4 as the core teaching scaffold and keep its design stable. Reuse the Micro-test M format from Appendix 2 for the immediate check. If you add a delayed mini-check, keep the same format and note the timing briefly in a small implementation note (main text or appendix), so readers do not confuse timing with a controlled classroom sequence.

**Why a flat overall score can still be instructionally meaningful**

(1) What the pattern suggests (cautious mechanism).

Overall accuracy staying near the same level does not mean the routine "did nothing." In classroom settings, a single global score can hide uneven category stability and selective breakdown points under different contexts. A structured view predicts that what matters is not only "how accurate," but "which categories remain stable when cues compete" (Chandrasekaran et al., 2016; Schertz & Clare, 2020). Sentence framing can introduce additional prosodic cues, and learners may prioritise these cues differently depending on experience and task demands (Gussenhoven, 2004; Xu, 2013). The practical value of the routine is therefore diagnostic: it identifies where the system is robust and where it is vulnerable.

(2) What the teacher can do next week (actionable moves).

Shift the classroom question from "Did students improve overall?" to "Where does performance break?" Use a simple planning rule: maintain Tone 3 with short reinforcement, but allocate most practice time to Tone 2 across multiple contexts (isolated → statements → questions). Keep sentence practice early, not late, but keep it small and controlled. Plan for one predictable breakdown point (questions) and one recovery step (contrast reset). This gives you a repeatable weekly cycle rather than a one-off activity.

(3) What material to reuse (Appendix/slide/task pointers).

Keep Table 2/Figure 1 as background context, but let tone-level and context-level materials drive decisions. Reuse the same item families in Appendix 1 and the same answer sheet in Appendix 2 to preserve comparability across weeks. This allows you to track "stability vs vulnerability" without constantly redesigning tasks.

**A dual scaffold principle: gesture for immediacy, pitch-line display for precision**

(1) What the pattern suggests (cautious mechanism).

Learner reports suggest a consistent division of labour between scaffolds. Gesture was often described as the most immediate support during listening, while the pitch-line display was described as helpful for seeing small differences and building a clearer contour image. This aligns with pronunciation pedagogy work showing that visual support can externalise pitch movement and make cue focus more explicit, while embodied cues can guide attention in real time (Hardison, 2004; Baills et al., 2019; Chun & Jiang, 2025). The evidence supports an alignment between scaffold roles and observed vulnerability patterns, not a claim that one scaffold "caused" gains.

(2) What the teacher can do next week (actionable moves).

Make scaffold roles explicit and consistent. Use gesture as a three-second pre-listening cue before each set ("Now listen for the rise start / the turning point"). Use the pitch-line display only in short explanation windows (30–45 seconds) after a set, and explain one cue difference only. Avoid turning the visual into a decorative feature. If students treat the line as "the answer," correct it immediately: "The line helps you listen. The sound is the answer." Keep the routine tight and repeatable.

(3) What material to reuse (Appendix/slide/task pointers).

Reuse the same contour visual in Appendix 4 with identical labels and arrows each time. Do not redesign it weekly. Use Appendix 3 prompts periodically (not every time) to check whether learners understood what the scaffold was for (what they listened for; which support helped and why). This keeps your reflective claims grounded.

**Responsible classroom claims and a practical "evidence upgrade" plan**

(1) What the pattern suggests (cautious mechanism).

Because the routine is embedded in normal teaching time and follows a fixed sequence, the strongest defensible claims are descriptive: stability differences across tones, and context-sensitive vulnerability in questions. Immediate post-episode changes are best treated as short-term attention effects unless supported by delayed checks or counterbalanced sequencing. Your statement–question screening reduces obvious non-F0 confounds, but it does not turn the block into a controlled intonation manipulation. This framing protects the credibility of a practice-oriented report.

(2) What the teacher can do next week (actionable moves).

Choose one minimal upgrade. Either add a delayed mini-check (same contrast, 4–6 items), or balance within-block order for statement vs question items using two versions of the playback sequence. Do not add both at once if classroom time is tight. The goal is not "more data." The goal is one small design choice that strengthens interpretability while keeping the activity teachable.

(3) What material to reuse (Appendix/slide/task pointers).

For delayed checks, reuse the Micro-test format from Appendix 2 and keep the same contrast targets. For order balancing, reuse the same sentences in Appendix 1 and only change the sequence. If you report the upgrade, add a short implementation note (one paragraph or a small appendix note) describing what changed and why, without framing it as a classroom sequence.

**A classroom mechanism chain you can act on**

(1) What the pattern suggests (cautious mechanism).

Taken together, the routine supports a coherent, practice-facing chain: learners tend to over-weight non-pitch cues early; brief explicit scaffolding can redirect attention to pitch movement; Tone 3 remains comparatively stable and can serve as a reference; Tone 2 is more vulnerable, especially in questions where rising cues may compete for attention (Chandrasekaran et al., 2016; Schertz & Clare, 2020; Braun & Johnson, 2011; Xu, 2013; Almalki & Morrill, 2016). This chain is presented as "consistent with" the observed pattern, not as a causal model verified by manipulation.

(2) What the teacher can do next week (actionable moves).

Convert the chain into a 20-minute weekly cycle that you can repeat:
Anchor (Tone 3 reset, 2–3 minutes).
Contrast (T2–T3, 6 minutes).
Stress test (short question set, 6 minutes).
Recover (one contrast item + two question items, 3–4 minutes).
Exit check (two students say what cue they listened for, 1 minute).
This keeps your practice systematic, and it directly targets the vulnerability point revealed by the routine.

(3) What material to reuse (Appendix/slide/task pointers).

Use Appendix 4 as the fixed scaffold for the cue reminders. Use Appendix 1 for the items and sentence frames, and Appendix 2 for response consistency. Use Appendix 3 sparingly for quick learner-voice checks that support your reflective interpretation. In reporting, keep the Micro-test M evidence (Table 5/Figure 4) clearly labelled as immediate and contrast-specific, and keep the question vulnerability evidence (Table 4/Figure 3) as the main "classroom stress test" indicator. The wider value of the Saudi case is not that it is unique, but that it highlights what teachers often face when a programme is still building its core routines. When resources are limited and class time is tight, teachers need evidence that is "good enough to guide the next lesson." The routine in this paper is designed for that purpose: it turns tone listening into a repeated classroom cycle, so materials, pacing, and emphasis can be adjusted week by week. In this sense, the contribution is a curriculum-friendly workflow rather than a one-off intervention.

### Limitations and Next Steps (practice-oriented)

#### Fixed classroom sequence and short-term carryover

(1) Limitation (what this routine cannot rule out).

This routine followed a fixed teaching sequence, which is practical for classroom use, but it limits causal interpretation. Later performance may reflect carryover from earlier exposure, short-term practice, pacing effects, or fatigue. Accordingly, block-to-block differences are reported as patterns observed within a realistic teaching flow rather than as clean causal effects.

(2) Next steps (what a teacher can do).

Keep the same lesson structure, but add one low-cost control. Use "two versions" rather than redesigning tasks:

Prepare Version A and Version B of the playback order for the later block(s).

Alternate versions across classes or across weeks (A this week, B next week).

If you only have one class, alternate by groups/rows (left side A, right side B) while keeping the items identical. This reduces the chance that a single fixed order produces a misleading peak or drop.

#### Micro-test M as an attention check, not a durable learning outcome

(1) Limitation (what the Micro-test M can and cannot show).

Micro-test M was intentionally small and narrow. It contained only four items and restricted responses to Tone 2 vs. Tone 3. It also occurred immediately after the micro-teaching episode. These features make it a useful classroom attention check, but they limit generalisation. A higher score is best treated as a time-local indicator of attention orientation after cue focusing rather than as durable improvement or broad tone learning (Hardison, 2004; Baills et al., 2019; Chun & Jiang, 2025).

(2) Next steps (what a teacher can do).

Add one delayed checkpoint without increasing workload:

Keep the immediate Micro-test M unchanged (4 items).

Add a delayed mini-check (4–6 items) either at the end of the lesson or at the start of the next lesson.

Keep the contrast and response format identical, so the only "new" factor is time. This single step upgrades your interpretation from "immediate responsiveness" to "persistence across time," while staying classroom-feasible.

#### Statement–question comparison is not a clean intonation manipulation (ma and position effects)

(1) Limitation (risk valve: ma structure + positional shifts).

The statement–question contrast should not be treated as a clean manipulation of intonation. Yes/no questions were formed with the particle ma, which changes sentence structure and can shift the target syllable away from utterance-final position. These structural and positional differences may interact with interrogative prosody. In addition, ma may affect the tonal realization of the preceding syllable through coarticulation and boundary-position effects. For

Tone 3 in particular, this can create variation between a fuller dip and a reduced “half T3” realization, which means the statement–question contrast may reflect tone-phonetic changes as well as sentence-level question framing. Although statement–question pairs were screened to be comparable in duration, speech rate, and mean intensity, F0 contours were not experimentally manipulated. Accordingly, statement–question patterns are interpreted as context-sensitive vulnerability under question framing, compatible with cue competition accounts, rather than as a clean causal effect of intonation (Chandrasekaran et al., 2016; Schertz & Clare, 2020; Xu, 2013; Gussenhoven, 2004; Almalki & Morrill, 2016; Chahal & Hellmuth, 2014; Hellmuth, 2016; Rifaat, 2021).

(2) Next steps (what a teacher can do).
Strengthen comparability while keeping the material teachable:
Keep sentence length constant across statement and question items.
Keep target-syllable position as constant as possible (same slot in the sentence frame).
Consider adding a small alternative question set (e.g., A-not-A) so you can observe question framing without relying only on ma.
Continue to treat questions as a classroom “stress test,” but avoid language implying a clean intonation manipulation. Because ma questions differ structurally and positionally from statements and F0 was not experimentally manipulated, statement–question patterns are interpreted as context-sensitive vulnerability rather than a clean causal effect of intonation.

**Non-manipulated F0 and classroom-realistic stimuli**
(1) Limitation (risk valve: natural recordings, no resynthesis).
This teaching note did not manipulate F0 contours experimentally. Stimuli were classroom-oriented recordings designed for teaching clarity, not acoustically resynthesised materials. As a result, differences across blocks may reflect multiple co-varying properties in natural speech (even when general recording conditions are kept consistent). Accordingly, evidence is interpreted as classroom-aligned patterns consistent with cue competition and cue focusing, not as definitive causal effects of pitch manipulation.

(2)Next steps (what a teacher can do).
Adopt a “matched-token” approach that remains realistic:
Record a small core set twice in the same session (normal vs. clearer pitch), keeping speaking rate steady.
Reuse the exact same tokens across weeks.
Keep the teaching interpretation consistent: treat clearer-pitch tokens as scaffolding for calibration, not as a treatment intended to produce lasting gains.
This improves comparability without requiring lab resynthesis. Because F0 contours were not experimentally manipulated, evidence is interpreted as classroom-aligned patterns consistent with cue competition and cue focusing, not as definitive causal effects of pitch manipulation.

**Order effects inside Block C and within-block mixing of sentence types**
(1) Limitation (risk valve: within-block order and mixing).
Statement and question items were embedded within the same block. Even when items are screened and matched on non-F0 dimensions, embedding two sentence types in a single block can partially confound sentence type with within-block order, attentional drift, or short-term adaptation. Therefore, sentence-type differences are treated as robust classroom signals only when they align with other indicators and learner reports, and they are reported with cautious language. Accordingly, results are interpreted as context-sensitive vulnerability rather than a clean sentence-type effect.

(2)Next steps (what a teacher can do).
Use one of these minimal, realistic fixes:
Split Block C into two short mini-blocks (statements only; questions only) with a one-line reset cue between them.
Counterbalance the order of these mini-blocks across classes/weeks (statements→questions vs. questions→statements).
If you must keep them mixed, rotate the internal order each week and log the order in a brief implementation note.

These steps directly reduce the "sentence type vs. order" ambiguity without changing the teaching aim. Because statement and question items were embedded within the same block, sentence-type patterns may partially reflect within-block order or attentional drift; accordingly, results are interpreted as context-sensitive vulnerability rather than a clean sentence-type effect.

**Limited item set, narrow lexical coverage, and local generalisation**
(1) Limitation (scope boundaries).
This routine relied on a restricted set of syllables and sentence frames in one local teaching context. This makes it usable as a diagnostic package, but it limits generalisation to broader vocabulary, other proficiency levels, or other L1 backgrounds. The most defensible contribution is therefore a classroom diagnostic profile of stability and vulnerability patterns under a specific routine rather than broad claims about tone learning in general (Hao, 2012; Schertz & Clare, 2020).

(2)Next steps (what a teacher can do).
Expand in a controlled way using a "core + extension" plan:
Keep a small core set constant each week (for tracking).
Add one extension set (new syllables or new frames) to probe transfer.
Change only one dimension at a time (syllable variety OR sentence complexity OR question type), not all at once.

This preserves interpretability while increasing classroom relevance. Given the restricted item set and local context, findings are presented as a classroom diagnostic profile of stability and vulnerability patterns rather than as broad generalisations.

**Measurement boundaries in a normal classroom routine**
(1) Limitation (what was not captured).
As a practice-oriented inquiry embedded in normal teaching time, this routine prioritised feasibility over comprehensive measurement. It did not include long delayed post-tests, and it did not attempt to isolate individual scaffold components as separate treatments. Brief interviews provide learner-aligned interpretation support, but they do not substitute for controlled longitudinal evidence. Accordingly, reporting focuses on descriptive alignment across indicators (performance patterns, disruption indices, learner reports) to support plausible classroom interpretations, not strong causal claims or long-term learning effects.

(2)Next steps (what a teacher can do).
Choose one low-burden add-on that strengthens your "next steps" logic without turning the class into a study lab:
Add a delayed mini-check (per 5.2).
Add a short transfer check using the extension set (per 5.6).

Add a quick production spot-check (unrecorded, pair work) with a simple teacher rubric (accurate / unclear / wrong) to link listening patterns to classroom speaking. Any one of these upgrades makes your reflection more actionable and your claims more credible. This practice-oriented routine supports descriptive classroom patterns and plausible mechanism interpretations, but it is not designed to establish strong causal claims or long-term learning effects.

### A minimal upgrade menu for future cycles (choose one per cycle)

To keep the routine teachable while strengthening interpretability, future cycles can adopt one upgrade at a time:
Delayed mini-check (4–6 items, next lesson) to test persistence beyond immediate cue focusing.
Order balancing (two versions of Block C) to reduce within-block order confounds
Question-type tightening (add a small A-not-A set) to reduce ma-related structural/positional differences.
Matched-token core set (same tokens reused across weeks) to increase comparability without resynthesis.
Core + extension sets to track both stability and transfer with minimal extra time.

## Conclusion

### What the routine revealed in a teachable sequence

This teaching note frames early tone listening as a classroom system rather than a fixed tone-by-tone difficulty list. In this dataset, Tone 3 stayed the most stable across formats, while Tone 2 showed the clearest vulnerability, especially under yes/no question framing. The Micro-test M spike is best read as a time-local shift in attention after cueing, not as evidence of durable improvement.

For teaching, the main value of the routine is diagnostic. It helps a teacher locate where listening breaks, not only whether a global score goes up or down. In this dataset, the predictable break point is Tone 2 in question contexts, and the predictable stabiliser is Tone 3 across formats. Learner reports aligned with this pattern by showing early reliance on non-pitch cues, strong perceived value of gesture, and perceived difficulty in questions.

Because the routine was embedded in normal instruction and followed a fixed sequence, the claims remain practice-facing and cautious. The evidence is presented as descriptive classroom patterns that are consistent with cue competition and cue focusing accounts, not as controlled causal effects.

### A practical teaching takeaway: “anchor–contrast–stress test–recover”

In practical terms, this routine can run as a simple weekly cycle without redesigning the course:
Anchor: a short Tone 3 reset to stabilise cue selection.
Contrast: a narrow T2–T3 contrast set with one fixed cue reminder.
Stress test: a small question set used intentionally, not postponed to “later.”
Recover: a quick reset item to prevent drift back to non-pitch cues.

This cycle does not aim to “prove” improvement in one lesson. It aims to reduce predictable breakdown points and to make cue focus repeatable. The Micro-test M functions as an immediate confirmation that learners noticed the intended contrast at that time. The more meaningful teaching evidence comes from whether the same cue focus survives later in sentences and in questions.

In short, the routine is not a replacement for a full assessment. It is a compact classroom tool for attention guidance, diagnosis, and targeted practice planning.

**What to reuse (a ready-to-run classroom package)**

This teaching note includes a small set of materials that can be reused with minimal preparation. Appendix 1 provides the item scripts and keys. Appendix 2 provides the answer sheet and the Micro-test M format. Appendix 3 provides brief check-in prompts for teacher notes. Appendix 4 provides an example pitch-line visual used for short, explicit cue reminders.

For classroom decision-making, the tables and figures serve three practical purposes: (1) to show which tones stayed more reliable across formats, (2) to show where sentence questions increased difficulty, and (3) to show whether a brief cue reminder "landed" immediately on a narrow contrast check. Reusing the same item families and the same visual labels across weeks helps keep the routine comparable without turning it into a research design.

**Next steps that keep the work classroom-feasible**

A practice-oriented report gains credibility when it shows one small improvement path without becoming a classroom sequence. The most feasible next step is to add one minimal upgrade per cycle:

- a delayed mini-check (to test persistence).
- order balancing in Block C (to reduce within-block order effects).
- a small alternative question set (to reduce ma-related structural/positional confounds).
- a matched-token core set reused across weeks (to strengthen comparability).

These upgrades preserve the central advantage of the routine: it remains teachable, quick to run, and directly linked to lesson planning. At the same time, they strengthen how confidently teachers can interpret the patterns they see. While the materials were developed in a Saudi Arabic-speaking classroom, the design logic is broadly applicable. Any setting where sentence-level rises are salient, and where programmes are still building stable teaching routines, can adapt this "check–teach–check" approach to support tone listening within normal lessons.

**Acknowledgments**

The author used ChatGPT (OpenAI) to assist with English language editing during the revision process. The author reviewed and verified all content to ensure accuracy and scholarly integrity.

**Data Availability Statement**

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**Ethical Approval**

This teaching note reports aggregated, de-identified classroom task responses collected within normal instruction. School-level permission was obtained before any data were used for reporting. Participation in the evidence component was voluntary and had no graded consequences. Because learners were minors, consent followed a school-authorised procedure implemented via school channels: guardians were informed and could opt out, and students could decline or stop at any time. No student audio or video was recorded; only the teacher's stimuli were used for classroom playback. All records were anonymised at source, stored securely with access limited to the teacher-author, and reported only in group form.

## Appendices

### Appendix 1. Classroom task materials (item script and answer key)

This appendix provides the full listening item scripts and the corresponding answer key for Blocks A–C. It is included to improve transparency and make replication possible for the classroom-based tone perception tasks.

A1.1. Listening Item Script (Blocks A–C)
Block A: Natural isolated-syllable items (16 items)
na(3)\
mi(1)\ba(4)\ma(2)\na(1)\mi(3)\ba(2)\ma(4)\mi(2)\na(4)\ma(1)\ba(3)\mi(4)\ma(3)\ba(1)\na(2)
Block B: Exaggerated-F0 isolated-syllable items (16 items)
ba(3)\ma(1)\na(4)\mi(2)\mi(4)\ba(1)\ma(4)\na(2)\mi(3)\ma(2)\ba(4)\na(3)\mi(1)\ma(3)\ba(2)\na(1)
Block C: Sentence-level items (16 items; statements and questions)

| statements | questions |
|---|---|
| 这是 mǎ。 | 这是 m á吗？ |
| 这是 m í。 | 这是 b à吗？ |
| 这是 b à。 | 这是 nǐ 吗？ |
| 这是 nā。 | 这是 mǎ 吗？ |
| 这是 m á。 | 这是 m í吗？ |
| 这是 nǐ。 | 这是 bā 吗？ |
| 这是 mā。 | 这是 m á吗？ |
| 这是 bā。 | 这是 nā 吗？ |

A1.2. Classroom Task Flow (Brief)
Block A → Micro-teaching (pitch-line display + gesture) → Micro-test M (4 items) → Block B → Block C

### Appendix 2. Learner answer sheet (blank version)

This appendix includes the blank answer sheet used during the classroom listening tasks. It shows the exact response format provided to learners across Blocks A–C and the Micro-test items.
B1. Answer Sheet for Blocks A–C and Micro-test M (4 items)
Block A Answer Sheet
请听音频读音节，每题听两遍，在你认为正确的声调（T1/T2/T3/T4）上打 ✓。
التي تعتقد أنها (T1/T2/T3/T4) يُرجى الاستماع إلى المقاطع الصوتية، وسيُشغّل كل سؤال مرتين. ضع علامة (✓) على الدرجة الصوتية
الصحيحة.

| No | T1 | T2 | T3 | T4 |
|---|---|---|---|---|
| 1 | □ | □ | □ | □ |
| 2 | □ | □ | □ | □ |
| 3 | □ | □ | □ | □ |
| 4 | □ | □ | □ | □ |
| 5 | □ | □ | □ | □ |
| 6 | □ | □ | □ | □ |
| 7 | □ | □ | □ | □ |
| 8 | □ | □ | □ | □ |

| | | | | |
|---|---|---|---|---|
| 9 | □ | □ | □ | □ |
| 10 | □ | □ | □ | □ |
| 11 | □ | □ | □ | □ |
| 12 | □ | □ | □ | □ |
| 13 | □ | □ | □ | □ |
| 14 | □ | □ | □ | □ |
| 15 | □ | □ | □ | □ |
| 16 | □ | □ | □ | □ |

Micro-test M Answer Sheet (4 items)

请听录音读音节，在你认为正确的声调上打✓。（4题，仅选T2/T3）

يرجى الاستماع إلى المقاطع الصوتية المسجلة، ووضع علامة (✓) على النغمة التي تعتقد أنها صحيحة.

| No | T2 | T3 |
|---|---|---|
| 1 | □ | □ |
| 2 | □ | □ |
| 3 | □ | □ |
| 4 | □ | □ |

Block B Answer Sheet

请听录音读音节（T2上扬更明显；T3低谷更低），听两遍，打✓。

يرجى الاستماع إلى المقاطع الصوتية المسجلة (سيتم تشغيل كل مقطع مرتين)، ثم وضع علامة (✓) على النغمة الصحيحة.

| No | T1 | T2 | T3 | T4 |
|---|---|---|---|---|
| 1 | □ | □ | □ | □ |
| 2 | □ | □ | □ | □ |
| 3 | □ | □ | □ | □ |
| 4 | □ | □ | □ | □ |
| 5 | □ | □ | □ | □ |
| 6 | □ | □ | □ | □ |
| 7 | □ | □ | □ | □ |
| 8 | □ | □ | □ | □ |
| 9 | □ | □ | □ | □ |
| 10 | □ | □ | □ | □ |
| 11 | □ | □ | □ | □ |
| 12 | □ | □ | □ | □ |
| 13 | □ | □ | □ | □ |
| 14 | □ | □ | □ | □ |
| 15 | □ | □ | □ | □ |
| 16 | □ | □ | □ | □ |

Block C Answer Sheet

请听录音读句子（陈述句 / 疑问句），判断这个句子提到的音节的声调。

.يُرجى الاستماع إلى الجُمَل المسجلة (تصريحية / استفهامية)، ثم تحديد الدرجة الصوتية للمقطع المذكور في الجملة

| No | T1 | T2 | T3 | T4 |
|---|---|---|---|---|
| 1 | □ | □ | □ | □ |
| 2 | □ | □ | □ | □ |
| 3 | □ | □ | □ | □ |
| 4 | □ | □ | □ | □ |
| 5 | □ | □ | □ | □ |
| 6 | □ | □ | □ | □ |
| 7 | □ | □ | □ | □ |
| 8 | □ | □ | □ | □ |
| 9 | □ | □ | □ | □ |
| 10 | □ | □ | □ | □ |
| 11 | □ | □ | □ | □ |
| 12 | □ | □ | □ | □ |
| 13 | □ | □ | □ | □ |
| 14 | □ | □ | □ | □ |
| 15 | □ | □ | □ | □ |
| 16 | □ | □ | □ | □ |

**Appendix 3. Post-task learner check-in prompts (teacher notes)**

This appendix provides the interview questions used to elicit learners' self-reported cue focus (e.g., duration, loudness, pitch movement), perceptions of statement vs question difficulty, and perceived usefulness of teacher gestures and the pitch-line display.

Student Interview Questions (English prompts)

Part A. How you listen to tones

Q1. When you listen to Chinese tones, what do you listen to most?
Do you listen to:
high or down sound?
long or short sound?
loud or soft sound?
or other things?

Q2. Which one helps you more?
The teacher's hand gestures?
The example words?
Or the pitch line on the screen?

Q3. Why?
Is it easy to see?
Easy to understand?
Or easy to remember?

Part B. Sentences

Q4. Which is harder for you?

A normal sentence?
Or a question
Q5. When the sound goes high at the end,
is it harder to hear the tone?

Part C. How you try to listen better
Q6. Next time, what will you listen to first?
High or low sound?
Long or short sound?
Loud or soft sound?
Q7. If there is no line on the screen, is it easier or harder for you？

**Appendix 4. Example Visual Support Used in Micro-teaching**

This appendix provides an example of the pitch-line visual support used in the micro-teaching segment to highlight key contrastive cues for Tone 2 and Tone 3. It also includes (4.1) slide design principles and (4.2) a timed 8-minute teaching script for direct classroom reuse.

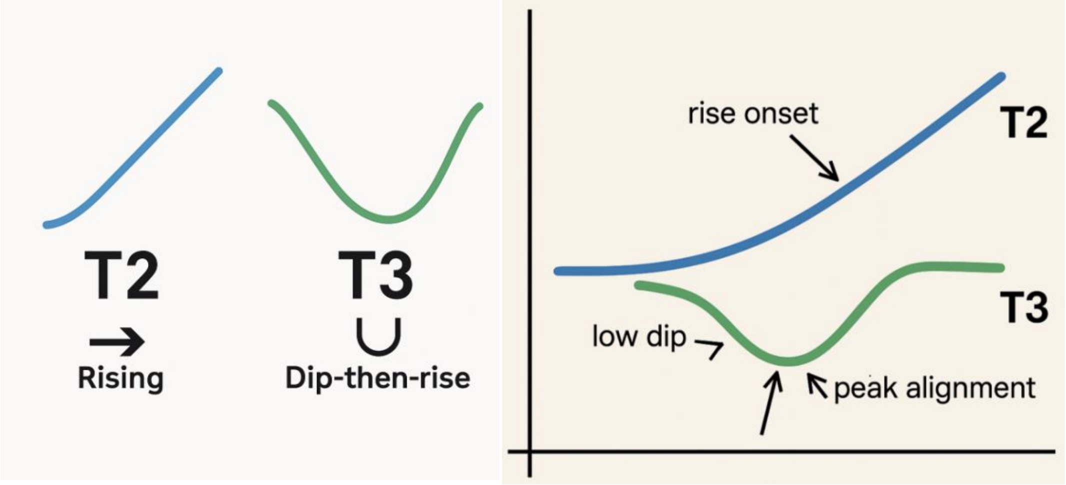
Transparency note on display-level enhancement (visuals only).

In this teaching note, display-level enhancement (sometimes described as "AI-enhanced" visuals) refers only to refinement of Praat-derived pitch-line visuals for classroom readability (e.g., smoothing/clean-up and adding simple labels or arrows). This refinement was applied only to the static contour images shown during teacher explanation. It did not alter the audio stimuli, generate learner-specific feedback, provide automated scoring, or feed into any statistical analysis. Accordingly, the visuals are included as teaching materials for transparency, and all quantitative results are based on learners' task responses within the classroom sequence rather than on the visual refinement itself.

4.1. Pitch-line visual design principles (for classroom projection)
(1) One contrast per slide. Show only Tone 2 vs Tone 3 on one screen.
(2) Fixed layout. Keep the same axis range and slide format across lessons.
(3) One cue label only. Mark the key cue (e.g., rise start for Tone 2; low dip/turn for Tone 3).
(4) Simple direction cue. Use one arrow to indicate movement; avoid extra annotations.
(5) Readability first. Large line, large text, clean background for whole-class viewing.
(6) Sound first, visual second. Keep the pitch-line display hidden during listening; show it briefly only when explaining.

4.2. 8-minute micro-teaching script (condensed: gesture + pitch-line display)
(1) 0:00–0:40 Set the listening target
"Today we listen for pitch movement."
"The sound is the answer. The line only helps you listen."
(2) 0:40–1:30 Gesture cue (Tone 2 vs Tone 3)
Tone 2: "Rise starts early." (rising hand gesture)
Tone 3: "Down, then turn." (dip + turn gesture)
(3) 1:30–5:10 Sound first, then pitch-line display (Tone 2 → Tone 3)
"Listen first. Then we check the line: rise start (Tone 2) vs turning point (Tone 3)."
"Your job is to catch the cue in the sound, not to 'read the picture'."
(4) 5:10–8:00 Fast contrast and transition to the Micro-test M
"Now only two choices: Tone 2 or Tone 3."
"In sentences, still listen for rise start vs turn."

Figure D1. *Example pitch-line visual used in the cue reminder (Tone 2 vs Tone 3).*



*Note.* The visual highlights the rise onset for Tone 2 and the low dip/turning point for Tone 3 to guide attention to contour shape and category-relevant cue differences.

**References**

Almalki, H., & Morrill, T. (2016). Yes/No question intonation in Urban Najdi Arabic. In J. Barnes, N. Veilleux, S. Shattuck-Hufnagel, & A. Brugos (Eds.), *Proceedings of Speech Prosody 2016* (pp. 606–610). International Speech Communication Association.

Baills, F., Su árez-Gonz ález, N., Gonz ález-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, *41*(1), 33–58. https://doi.org/10.1017/S0272263118000074

Boersma, P., & Weenink, D. (2025). Praat: Doing phonetics by computer (Version 6.4.48) [Computer software]. Retrieved December 15, 2025, from https://www.praat.org/

Braun, B., & Johnson, E. K. (2011). How language experience and linguistic function guide pitch processing. *Journal of Phonetics*, *39*(4), 585–594. https://doi.org/10.1016/j.wocn.2011.06.002

Chahal, D., & Hellmuth, S. (2014). The intonation of Lebanese and Egyptian Arabic. In S.-A. Jun (Ed.), *Prosodic Typology II: The Phonology of Intonation and Phrasing* (pp. 365–404). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199567300.003.0013

Chandrasekaran, B., Yi, H.-G., Smayda, K. E., & Maddox, W. T. (2016). Effect of explicit dimensional instruction on speech category learning. *Attention, Perception, & Psychophysics*, *78*(2), 566–582. https://doi.org/10.3758/s13414-015-0999-x

Chun, D. M., & Jiang, Y. (2025). Computer-assisted pronunciation teaching. In C. A. Chapelle, J. Levis, M. Munro, C. Nagle, & A. Huensch (Eds.), *The encyclopedia of applied linguistics*. Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0172.pub3

Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge University Press.

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. https://doi.org/10.1016/j.wocn.2011.11.001

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *8*(1), 34–52. https://doi.org/10.64152/10125/25228

Hellmuth, S. (2016). Exploring the syntax–phonology interface in Arabic. In S. Davies & O. Soltan (Eds.), *Perspectives on Arabic Linguistics XXVII: Papers from the Annual*

*Symposium on Arabic Linguistics, Bloomington, Indiana, 2013* (pp. 75–98). John Benjamins. https://doi.org/10.1075/sal.3.04hel

Hellmuth, S. (2018). Variation in polar interrogative contours within and between Arabic dialects. In *Proceedings of Speech Prosody 2018*. International Speech Communication Association (ISCA).

Hellmuth, S. (2022). Sentence prosody and register variation in Arabic. *Languages, 7*(2), 129. https://doi.org/10.3390/languages7020129

Li, Q., Zhang, J., & Cai, W. (2024). Utilizing ChatGPT to implement differentiated instruction in teaching Chinese as a second language. *International Journal of Chinese Language Teaching, 5*(1), 74–89. https://doi.org/10.46451/ijclt.20240106

Qin, Z., Lee-Kim, S.-I., & Qi, H. (2024). The effect of second-language learning experience on Korean listeners' use of pitch cues in the perception of Cantonese tones. *Second Language Research*. Advance online publication. https://doi.org/10.1177/02676583241244604

Rifaat, K. (2021). Intonation in Arabic. In J. Owens (Ed.), *The Cambridge Handbook of Arabic Linguistics* (pp. 330–352). Cambridge University Press.

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *WIREs Cognitive Science, 11*(2), e1521. https://doi.org/10.1002/wcs.1521

Wang, X. (2012). Difficulties with Mandarin tones: Learners' perspectives and speech data analysis. In *Proceedings of the 15th Oriental COCOSDA Conference*.

Wang, T., Potter, C. E., & Saffran, J. R. (2020). Plasticity in second language learning: The case of Mandarin tones. *Language Learning and Development, 16*(3), 231–243. https://doi.org/10.1080/15475441.2020.1737072

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication, 46*(3–4), 220–251. https://doi.org/10.1016/j.specom.2005.02.014

Xu, Y. (2013). ProsodyPro: A tool for large-scale systematic prosody analysis. In *Tools and Resources for the Analysis of Speech Prosody (TRASP 2013).*

***Xinsheng Cao*** is an M.A. candidate in Teaching Chinese Language and Culture at the Irish Institute for Chinese Studies, University College Dublin, Ireland. His research interests include second language acquisition of Chinese and technology-enhanced language teaching.