JASA REVIEW-ARTICLE

A cross-linguistic review of citation tone production studies: Methodology and recommendations

Chenzi Xu[1,a] and Cong Zhang[2]
[1]Faculty of Linguistics, Philology and Phonetics, University of Oxford, Oxford, OX1 2HG, United Kingdom
[2]Speech and Language Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

ABSTRACT:
The study of citation tones, lexical tones produced in isolation, is one of the first steps towards understanding speech prosody in tone languages. However, methodologies for investigating citation tones vary significantly, often leading to limited comparability of tone inventories, both within and across languages. This paper presents a systematic review of research methods and practices in 136 citation tone studies on 129 tonal language varieties in China, including 99 studies published in Chinese, which are therefore not easily available to an international scientific readership. The review provides an overview of possible analytical decisions along the research pipeline, and unveils considerable variation in data collection, analysis, and reporting conventions, particularly in how f0, the primary acoustic correlate for tone, is operationalised and reported across studies. Key methodological issues are identified, including small sample sizes and inadequate transparency in communicating methodological decisions and procedure. This paper offers a clear road map for citation tone production research and proposes a range of recommendations on speaker sampling, experimental design, acoustic processing techniques, f0 analysis, and result reporting, with the goal of facilitating future tonal research and enhancing resources for underrepresented tonal varieties.
 https://doi.org/10.1121/10.0032356
(Received 18 July 2024; revised 18 September 2024; accepted 20 September 2024; published online 15 October 2024)
[Editor: James F. Lynch] Pages: 2538–2565

I. INTRODUCTION

Lexical tones are relative pitch contrasts that differentiate lexical meanings. For instance, in Standard Mandarin, the syllable *ma* can be produced with four different lexical tones and indicate different meanings, as illustrated in Table I. Tone languages are languages that feature such lexical pitch contrasts. They take up a significant proportion of the world's languages: Yip (2002) suggested a range of 60% to 70% and Maddieson (2023) estimated 47%. The first step to any speech prosody research in a tone language is to investigate its lexical tone system—how many lexical tones there are in the variety and what they are like in citation form. Citation tones in a language serve as the baseline for understanding the tonal system, without the influence of complex linguistic contexts or tonal interactions. They are often analysed as fundamental building blocks for surface melodic patterns of larger linguistic units. Despite the significance, however, there appears to be a lack of consensus regarding the methodologies employed for conducting such investigations.

It has been over a century since the pioneering piece by Chao (1922) on the experimental methods for citation tone research in Chinese languages. This article was published at a time when instruments such as gramophone (Jones, 1909) and Rousselot apparatus (Bradley, 1916) were employed to plot the pitch curves on a musical stave. Although these creative early methods of pitch measurement and visualisation are now outdated, Chao's insights and principles of conducting experiments continue to influence modern tonal research. With the technological advancements over the past 100 years, it is high time that we set up a new guidance on the analysis of citation tones. We conducted a systematic methodological review of citation tone studies of the last five decades on tone languages within China and explored the issues involved in examining citation tones, as well as proposed a comprehensive set of recommendations for future research.

A. Citation tone

Citation tone is a lexical tone in its citation form. The term "citation form" refers to "the spoken form a word has when produced in isolation, such as when cited for purposes of illustration, as distinguished from the form it would have when produced in the normal stream of speech" (Collins Online Dictionary, 2024). Tone can be conceptualised in two approaches: as concrete surface contrasts (Pike, 1948) or an underlying property of morphemes (Hyman and Leben, 2020). The starting point of citation tone studies is usually to determine the surface tonal contrasts, and many studies directly take the citation form as the underlying form in phonological derivations [e.g., Shen (1992) on Mandarin tone sandhi].

[a]Email: chenzi.xu@ling-phil.ox.ac.uk

https://doi.org/10.1121/10.0032356

TABLE I. Examples of the four lexical tones in Standard Mandarin.

Lexical tone	Chinese example	Phonetic transcription	Possible meaning
High-level tone	妈	ma˥	"mother"
Rising tone	麻	ma˧˥	"hemp"
Fall-rise (dipping) tone	马	ma˨˩˧	"horse"
Falling tone	骂	ma˥˩	"scold"

Lexical tones can be divided into contour tones and level tones. Within Chinese varieties, both types of tones are common. Mandarin varieties [e.g., Li *et al.* (2019) on Tianjin Mandarin] tend to be contour-tone based; many southern varieties, such as Cantonese [e.g., Zee (1991) on Hongkong Cantonese] and Hakka [e.g., Lee and Zee (2009) on Meixian Hakka] varieties, have tonal contrasts which differ in tone contours and tone levels; and in some varieties, such as Tibeto-Burman varieties [e.g., Qu and Yang (2019) on Benna Hani], only contrastive level tones exist.

In Chinese and Southeast Asian tone languages, pitch is contrastive on most syllables. Some languages feature specific types of tones, such as checked tones and neutral tones. Checked tones, commonly known as the entering tones in Middle Chinese, are different from the other tones in that it is not a concept purely related to tonal contour or levels; instead, they indicate that the syllable carrying a checked tone ends in an oral (no audible release) or glottal stop. For instance, in Cantonese, the word 閪/hɐp/ "union" ends with a final unreleased [p˺] and is thus considered to have a checked tone. Checked tones are sometimes not included in the total number of tones when they are merged into other tone categories. Neutral tones are also referred to as "light tones," translated from the Chinese term. An example would be the question particle 吗/ma/ in Mandarin. It is the same syllable as the examples in Table I, only with a neutral tone. Neutral tones are often produced with reduced loudness and duration as well as centralised vowels in Mandarin and, therefore, regarded as reduced tones instead of full tones (Xu, 2024). A neutral tone has also been considered as a contextual tone (Yip, 2002): in Mandarin, it mostly occurs enclitically in unstressed positions, where its pitch patterns heavily depend on the preceding syllable [see Chen (2015) for a review of neutral tones across Chinese varieties]. With the perspective of neutral tones as a phenomenon neutralising contrasts among lexical tones in unstressed positions (Chao, 1968), they are usually excluded from the tonal inventory of a language variety.

B. Lexical tone transcription

The description of a tone system often concerns how pitch is transcribed. The International Phonetic Association endorsed two sets of transcriptional conventions of tones considering its prevailing usage by linguists working with tone languages in Africa and Asia (Maddieson, 1990). The Africanist tradition utilises accent marks superscript to primarily vowels (e.g., ê for a falling tone), while the Asianist tradition employs Chao's tone letters (e.g., ˥˩ for a falling tone). Each tone letter consists of a simplified time-pitch bar or shape attached to the left of a vertical reference line that defines the tone stave (Chao, 1930). Its numeric equivalent is also commonly used. In Chao's tone letter system, the tonal range is idealised as a five-level scale, with 1 representing the lowest pitch and 5 the highest. The juxtaposition of numerals from left to right signals a tone contour: its beginning and end, and any inflection point in the middle (e.g., 51 for a falling tone).

Contrary to the highly conventionalised presentation of phonemes, a wide range of qualitative and quantitative methods are available to describe citation tones, ranging from impressionistic text description (e.g., rising, falling), tone letters and numerals (e.g., 55, 213), phonologised notations (e.g., H, L, HL), to coefficients or parameters from various functional models of f0 contours [e.g., Andruski and Costello (2004)]. The choice of a method or a combination of these methods depends on the purposes and analytical framework of the research, whether it is oriented to language documentation and description, theoretical analysis regarding the abstract phonological representation of tone, or testing the validity of models through analysis-by-synthesis.

C. The current study

The chief goals of this study are to (1) have a thorough overview of the current field by reviewing the methodologies used in studying citation tones and (2) highlight and illustrate effective methodologies and practices informed by recent linguistic advances. The guiding question of the systematic methodological review is: How is citation tone production research conducted? Specifically,

(1) How are speech data of citation tones collected?
(2) Which technologies and analytical tools are commonly used in citation tone production studies?
(3) What methods are employed in analysing f0 data of citation tones?
(4) How are citation tones reported and transcribed?

To answer these questions, we reviewed studies from multiple sources. We limited our scope to language varieties in China, most of which are tone languages of a similar type, featuring large tone inventories, unitary contour and level tones, and primarily lexical rather than grammatical tones (Ratliff, 2015). Section II details the relevant search (sampling), screening, and coding procedures, together with the full inclusion and exclusion criteria.

Section III synthesises information from all the selected studies and presents the findings of our review. In this review, we focused on the analysis of the pitch/f0 pattern, since it is the primary acoustic correlate of tone in most tone languages, though tone can be signalled by many other phonetic cues, such as amplitude envelope [e.g., Zhou and Martin (2012)], phonation [e.g., Yu and Lam (2014)], and

duration [e.g., Howie (1976)]. We also left neutral tones, which do not occur in isolation, out of the account.

In Sec. IV, we evaluate the methodologies reported in the literature in the spirit of reproducible and replicable research and propose a set of recommendations aimed at promoting principles and practices of reproducibility in the design, implementation, and communication of citation tone research. The recommendations that benefit the study of tone languages in China are likewise useful for analysing citation tones in many other tone languages in different parts of the world, for examining tone production in both first and second language acquisition, and for investigating other pitch phenomena in prosody research.

Finally, in Sec. V, we conclude the review by addressing unresolved issues and highlighting directions for future studies.

II. METHODS

We consulted a range of peer-reviewed phonetics journals and international conference proceedings in English, as well as assorted journal articles in Chinese, and established a database of citation tone studies. A coding protocol was then developed to synthesise the methodological information from the literature, resulting in a comprehensive review datasheet. The procedure of creating the database and the review datasheet is illustrated in Fig. 1 and introduced in detail in this section.

A. Literature search and retrieval

Given the language expertise of the two authors and the research scope of varieties in China, we drew from academic sources in both English and Chinese. The inclusion of Chinese publications greatly enriched our review with first-hand data and analyses of a broader range of under-represented language varieties, which are not easily accessible to an international scientific readership, and also alleviated the publication bias by incorporating exploratory and descriptive research without significant theoretical findings.

For papers written in English, we started with widely cited journals including *Journal of International Phonetics Alphabet* (JIPA) and *Journal of Phonetics* (JPhon) through searching for the keyword "tone" and going through all results manually. Then we resorted to proceedings of high calibre international conferences including the *International Congress of Phonetic Sciences* (ICPhS) and *International Conference on Tone* and *Intonation* (TAI). All ICPhS proceedings ranging from the first in 1932 to the 20th in 2023 were searched, with the exception of the 8th in 1975, for which no proceedings were published. The first TAI proceeding published in 2021 was included. Then we searched for any remaining articles using *Linguistics* and *Language Behaviour Abstracts* (LLBA), a commonly used database that indexes a variety of articles in linguistics. We employed the "advanced search" function with keywords and combinations of keywords in the fields of title and abstract. For articles written in Chinese, we searched in *China National Knowledge Infrastructure* (CNKI), the largest multidisciplinary electronic database of academic resources in China, using the keyword "单字调" ("mono-character tone" or "citation tone"). All searches were conducted between 28 August and 31 August 2023.

Table II lists the sources, keywords, and search fields in our replicable search procedure, and presents the number of results returned and the number of articles included in the current study. Duplicate articles from multiple searches were included only once.

The keyword "citation tone," unexpectedly, failed to yield any studies of potential interests in the English literature during our search, though its Chinese counterpart "单字调" resulted in a retrieval of 99 studies. For instance, no study contained the string "citation tone" in their title since the very first ICPhS proceedings in 1932. Therefore, keywords with a much broader scope such as "tone" were used as shown in Table II, which introduced challenges in manually screening and identifying pertinent studies.

B. Inclusion and exclusion criteria

Explicit inclusion and exclusion criteria were developed tailored to our research goals. A study was included if it met all of the following conditions:

(1) It was a published article in an academic journal or a peer-reviewed conference.
(2) It contained empirical data featuring the production of citation tones in a tone language spoken in China.
(3) It focused on the production of citation tones by native speakers of the language variety, without reported speech impairments, and mainly reside or grow up in the regions where the language variety was predominantly spoken, rather than by second language (L2) learners or members of the overseas diaspora.
(4) It featured the complete set of full tones in the tonal inventory of a language variety. A study examining a subset of lexical tones was included only when checked tones were omitted.
(5) It included an analysis of the f0 contours of the citation tones produced in isolation.

Studies were excluded for the following reasons:

(1) The study featured a tone language spoken outside China.
(2) The study was a dissertation or thesis.
(3) The study was a literature review or theoretical piece without any empirical data.
(4) The study featured the perception of tones without a speech production component.
(5) The study featured L1 or L2 acquisition, examining the production of citation tones by young children, language learners, members of the overseas diaspora, or individuals with speech and hearing disorders.
(6) The study concentrated on tone sandhi or tones in various tonal contexts without a component of tones produced in isolation.

https://doi.org/10.1121/10.0032356

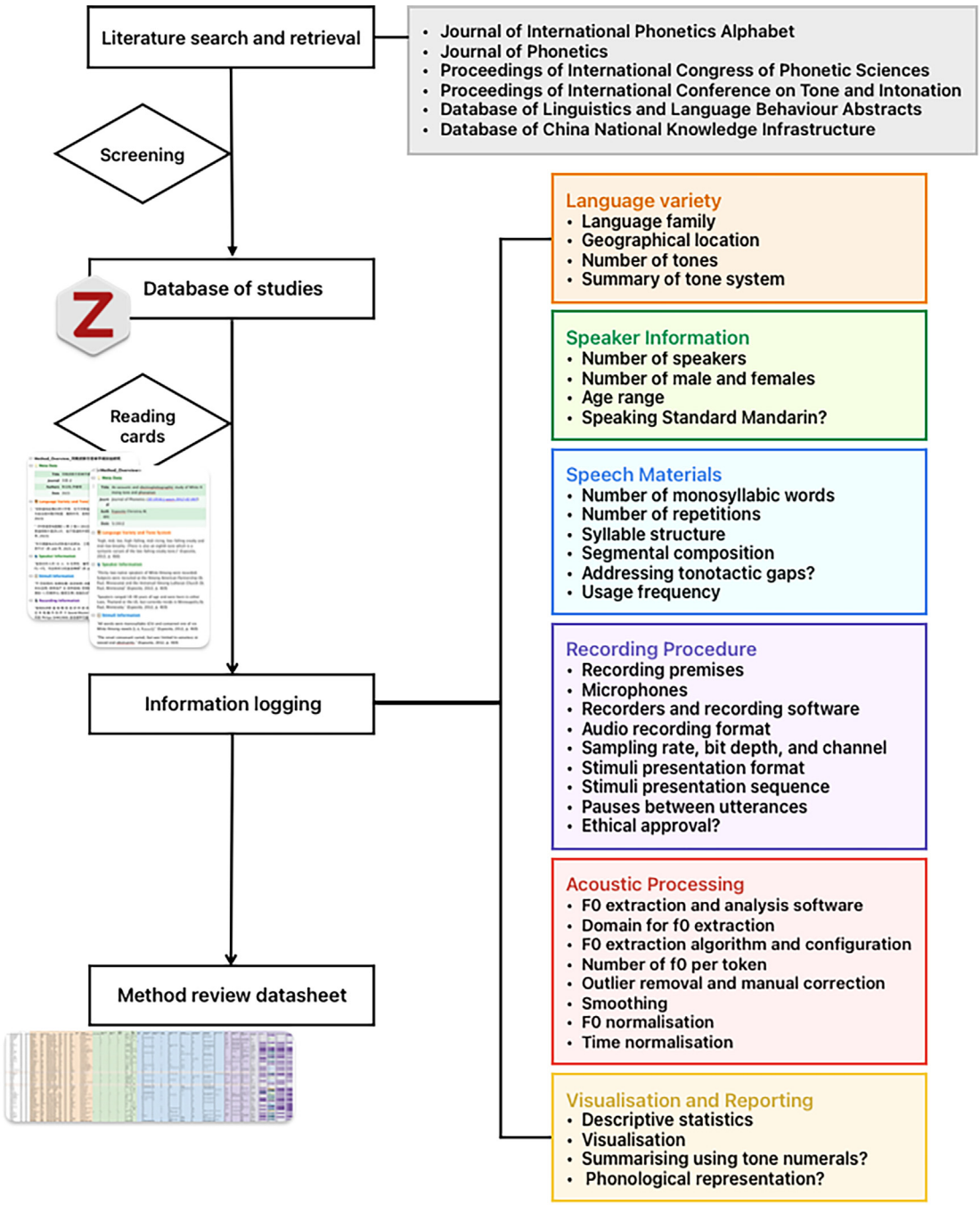

FIG. 1. (Color online) The flow chart of creating the review database and datasheet. Relevant methodological descriptors are shown in six coding categories under information logging, where the question marks indicate Boolean values.

(7) The study did not address some non-checked tones in the tone system of a variety.
(8) The study only targeted specific features of citation tones (e.g. duration, phonation, or laryngeal settings).
(9) The full text of the article was not available or accessible.

Studies that did not align with these requirements were excluded in the present review, despite their valuable contributions to the general field.

The searches returned over a thousand studies related to tone, among which 136 empirical studies were selected and retrieved. There are 99 Chinese articles sourced from a diverse array of journals in the CNKI database. The remaining 37 English articles comprise 9 from various ICPhS proceedings, 20 from the *JIPA*, and a few from other journals, mostly via the *LLBA* database, including the *JPhon*, *Phonetica*, *International Journal of Chinese Linguistics*, *Languages Language* and *Linguistics Lingua* and *Australian Journal of Linguistics*. It is worth noting that while this collection of citation-tone papers may not be exhaustive, the coverage of over 100 studies from well-cited scholarly

sources over the past five decades (1976–2023) offers a reflective overview of the field and provides valuable insights into research methodology.

The database features 129 language varieties, as shown geographically in Fig. 2, where small islands are not shown on the map due to the lack of language data. If a paper contained age-related sub-varieties, they were counted under the same category in the calculation of language variety.

Figure 3 displays the number of research articles in the database by their publication year and by the language group involved. The period from 2011 onwards has seen a boom in citation tone studies within our database. It is worth mentioning that the year 2011 shows a noticeable surge in the number of studies in the figure. A possible cause is that the 17th ICPhS was held in Hong Kong that year, which promoted the representation of scholars with a language background and expertise in tone languages compared to previous years. Therefore, hosting conferences in diverse regions worldwide may enhance the inclusiveness of research samples, perspectives, and practices. Figure 3 also shows that varieties of Mandarin Chinese are more frequently researched than varieties in other groups, likely to be reflective of the larger population of Mandarin speakers in China.

C. Information logging

The database is accessible through a public Zotero library (Xu and Zhang, 2024a). For each selected study, the description of research methods was annotated and distilled into a row in our comprehensive review dataset, with six key categories of information logged: (1) basic information about the language variety, (2) speaker details, (3) design of speech materials, (4) aspects of recording procedures, (5) acoustic measurements and processing techniques, and (6) reporting of findings. To facilitate information logging, a customised "reading-note card," a note-taking text file as shown in Fig. 1, was created using the Zotero plugin ZotCard (2023) for each paper. In the review datasheet, the methodological descriptors were encoded as string (textual), numeric, and boolean (True or False) values. Below we describe each key category in more detail.

Language variety: Basic information about the language variety including the language family and where the variety is spoken were logged. We also noted down the basic information of the tonal system of each variety, such as the number of tones and the description or representation of each tone, if such information was available.

Speaker information: The number of speakers involved and the linguistically relevant demographic information such as age and sex were logged. Considering Standard Mandarin's special status being the official language in mainland China, whether they speak Standard Mandarin was also tracked, which largely reflects a description of their language background.

Speech materials: The number of monosyllabic words for each tone category and their total count, along with the number of repetitions, were logged. We also tracked whether the design of speech materials controlled for syllable structure and segments, addressed any tonotactic gaps, and considered the usage frequency of the words.

Recording procedure: This included experiment setup, such as recording premises; equipment, such as microphones, recorders, and recording software; as well as the way speech materials were presented to participants, including format, sequence, and rate. Additionally, key acoustic parameters of the digital audio recordings including the sampling rate, bit depth, codec, and the number of channels were tracked. We also coded whether information on ethical approval was reported.

Acoustic processing: We recorded detailed information about the specific software for acoustic analysis, as well as the algorithm for f0 extraction and its associated configuration. The method employed for normalising f0 and the procedures for addressing f0 outliers were also tracked.

Visualisation and reporting: How f0 results of citation tones were reported and visualised was logged. A final notes column was added to provide supplementary information pertaining to the methodology or offer additional observations on the findings.

The resulting datasheet with metadata and methodological annotations of the selected articles in the database is publicly available online (Xu and Zhang, 2024b), alongside a script to reproduce the respective counts presented in this paper.[1]

TABLE II. The number of journal articles according to keywords and combinations of keywords in the search procedure. If an article appears in more than one search result, it is only counted once in the "No. of selected" column.

Sources	Keywords	No. of results	No. of selected
JIPA	Chinese tone(full text)	87	20
JIPA	tone(title)	51	0
JPhon	tone(title, abstract, keywords)	145	1
ICPhS	tone(title)	271	9
TAI	tone(title)	14	0
LLAB	citation tone(title)	6	0
	tone system(title)	53	0
	tone(title) AND production(title)	101	0
	tone(title) AND acoustics(title)	72	2
	tone(title) AND production(abstract)	223	2
	tone (title) AND acoustics(abstract)	272	3
CNKI	单字调 (title)	146	99
Total		1441	136

III. CROSS-LINGUISTIC REVIEW ON METHODOLOGY

The database encompasses a broad spectrum of tone languages, with the reported number of citation tones ranging from 2 to 14, as shown in Fig. 4. The most common tone system in the database features a four-way tone contrast, primarily due to the large inclusion of Mandarin varieties in the database—49 Mandarin varieties have four citation tones. The number of varieties decreases as the number of citation tones increases from 5 to 10.

The categorisation of citation tones, though, depends on the analytical framework used in the papers. For instance,

https://doi.org/10.1121/10.0032356

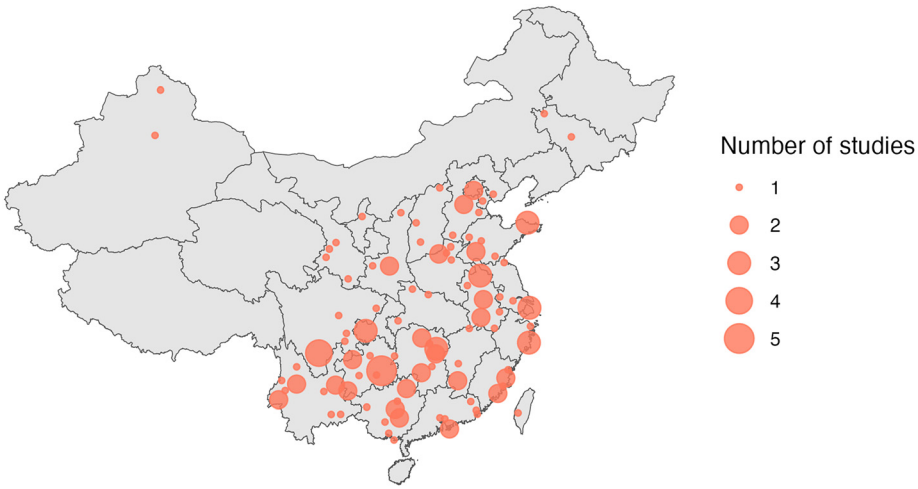

FIG. 2. (Color online) Geographical distribution of language varieties included in the survey of citation tones of Chinese languages. The geographic coordinates of language varieties are aggregated over prefecture-level cities and autonomous regions. The size of the dots indicates the number of academic papers centred on varieties of the given region in the survey. The dataset of administrative boundaries shown is compiled by the United Nations OCHA Regional Office for Asia and the Pacific.

while Cantonese has been considered to have nine tones [e.g., in Gu *et al.* (2007)], many contemporary analyses agree on six tones [e.g., in Fung and Lee (2019)], considering the six distinct pitch patterns. In the latter case, the shorter duration of the checked tones alone did not lend them to being categorised as independent tones. The two varieties reported to have as many as 10 and 14 tones are Jingxi Zhuang (Tai) and Madong Cantonese, both located in Guangxi Province, China. Their categorisation was based on historical categories, with tonal distinctions arising from assorted tone splits that trace back to the Proto-Tai Tones (Gedney, 1972) and Middle Chinese, respectively. Among the 14 reported tones though, the pitch patterns of the checked (*rù*) tones seemed to be similar to some of the *shǎng* and *qù* tones in Madong Cantonese (Liang and Tang, 2017).

The review unveiled considerable variation in how the pitch patterns of citation tones were elicited, measured, and reported, with much of this variation stemming from diverse conventions and practices by different schools of researchers, as well as varying expectations of different publications. *JIPA* is well-known for hosting papers on phonetic

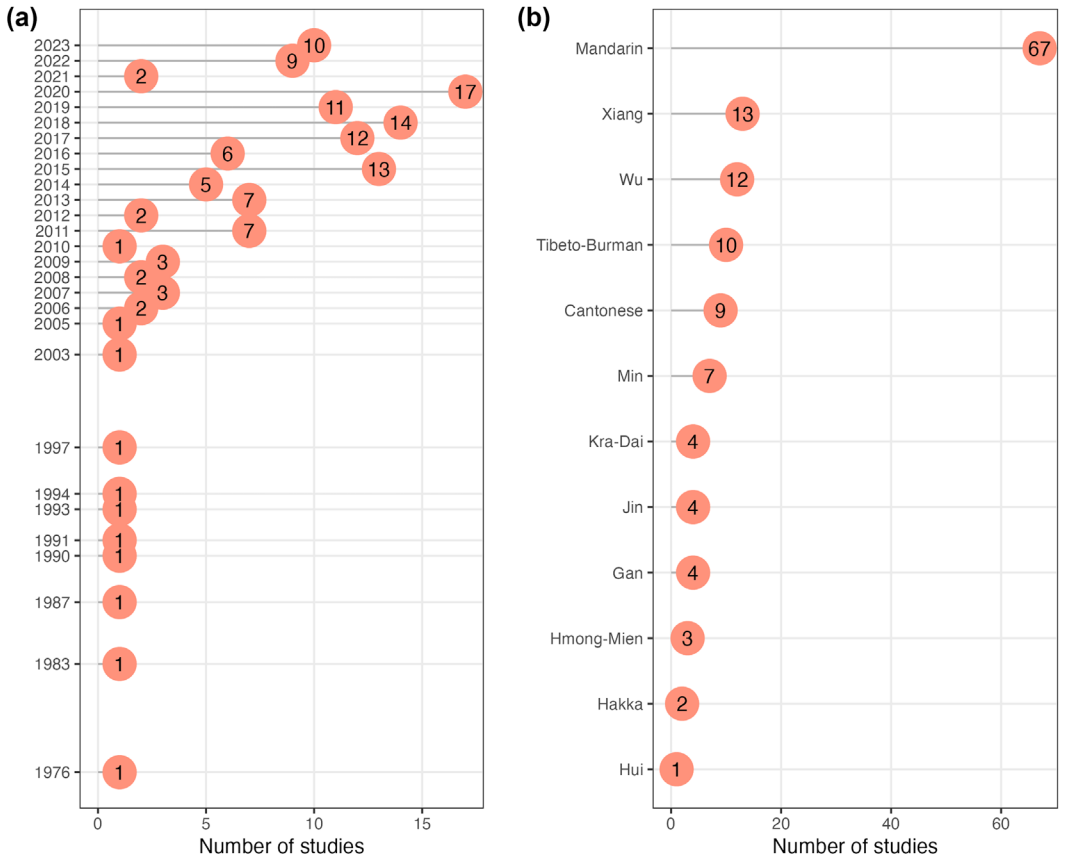

FIG. 3. (Color online) The number of publications by year (a) and by language group (b) in the database.

descriptions of un(der)documented linguistic varieties. The majority of the *JIPA* inclusions here, for instance, were published as *Illustrations of the IPA*. Of the IPA illustrations, 17 out of 19 featured only a single speaker with an extremely limited report on speech materials, recording procedure, and f0 processing details. These studies, though effectively demonstrated the use of the IPA symbols, lack reproducibility and generalisability, since it is difficult to determine whether the observed linguistic phenomena are representative of the language or idiosyncratic to the speaker. Our search revealed that there were only a small number of publications on citation tone research in internationally renowned journals. Many leading peer-reviewed journals in the field gravitate towards in-depth theoretical or confirmatory research, which tends to marginalise the lesser-studied language varieties that lack fundamental knowledge and insights, due to the scarcity of high-quality descriptive and exploratory research.

The collection of monosyllable recordings in fact constitute a specialised corpus for citation tone research. In addition to speech recordings, an integral part of a corpus is metadata, including administrative, editorial, and descriptive information about the speakers, materials, and recordings, so as to provide the essential context for an empirical investigation.

This section presents the findings of the methodological review across all studies in our database, focusing on the corpus size, corpus design, data analysis, and the reporting of citation tones. The results are descriptive, the majority of which report on the proportion of articles that fulfill relevant characteristics relative to the number of articles where each characteristic is applicable. Prior to the presentation of results, the rationale for each characteristic or descriptor examined is also explained.

A. Corpus size

The primary objective of citation tone research is to identify the prototypical tone patterns of a particular language variety. This informs two key aspects of the experimental design: first, the selection of a sufficiently large and diverse group of speakers is crucial to ensure that the results reflect the broader population of the speech community; second, the use of an adequate number of monosyllabic word items is essential so that the findings generalise beyond the specific recorded sample of the language materials (Clark, 1973). Additionally, repetitions of the same item produced by a speaker also enable a more robust estimate of its representative pitch pattern. In other words, the tone patterns of a variety should be generalised over speaker, word, and utterance idiosyncrasies. Figure 5 summarises how studies in the database vary in the way they report (or not report) the number of speakers, items, items per tone category, and repetitions.

The majority of studies (94 out of 136 studies) employed speech data from between one to four speakers, with a single speaker being the most common (52 studies, 17 were *Illustrations of IPA*). The number of studies that involved five and above speakers have significantly dropped. Roughly 26% (36 studies) of the studies involved at least six speakers, which, according to Ladefoged (1997), is the absolute minimum number necessary for meaningful measurements. However, recent technological developments have made it possible to acquire high-quality recordings from more participants with accessible devices, such as a smartphone, as in Hou *et al.* (2023). It is no longer recommended to present data solely based on instrumental records of a single speaker, as Ladefoged (1997) emphasised. The lack of representativeness and generalisability can be mitigated by increasing the number of participants, except in cases of endangered languages where only a few speakers remain.

Studies varied widely in the total number of monosyllabic word items ranging from 12 to 3900 [Fig. 5(b)]. Five studies involved recordings of over a thousand items produced mostly by a single speaker, and three of them were published in JIPA illustrations with only a brief section on tones. Despite that in Chen and Guo (2022) two speakers produced 3900 monosyllabic words, the tone pattern illustration is based on only ten words sampled from each tone category produced by one speaker. When a study aims to document and trace tone changes conditioned on segmental properties [e.g., Li (2019)], or explores an understudied tone system without prior studies on its tone categories, it is

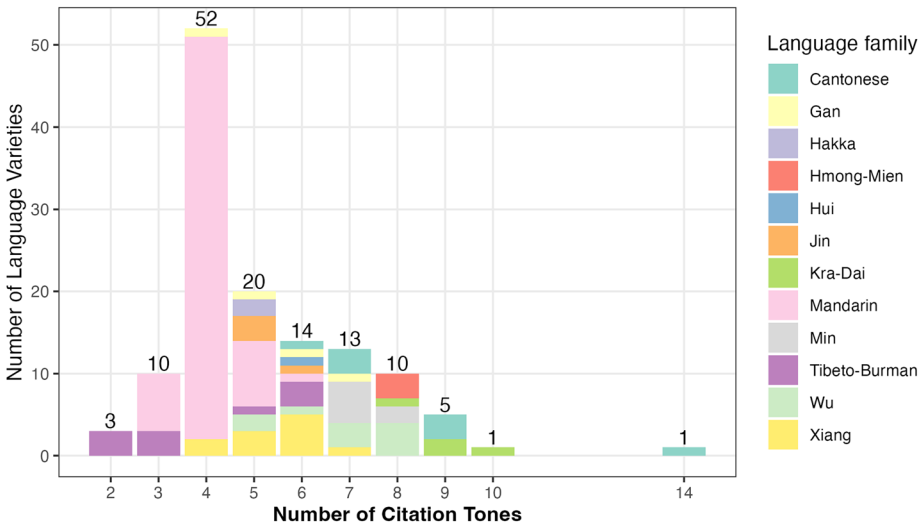

FIG. 4. (Color online) The distribution of the reported number of citation tones across language varieties by language family

https://doi.org/10.1121/10.0032356

reasonable to include a large number of words for a comprehensive analysis, although this approach is time-consuming and can be very demanding for participants.

Approximately 19% (26 studies) of the studies in the database omitted details regarding the number of items in each tone category, among which 14% (19 studies) did not disclose the total number of items nor any other information about the monosyllabic speech materials. Among the studies that reported on the number of items per tone category, 22 had varying numbers of items across tone categories. The remaining studies reported item counts per tone category ranging from 3 to 120, with 10 being the mode (23 studies).

Only about 60% of the studies reported whether their production experiments included any repetitions. Most of these studies used between one and three repetitions, with two repetitions being the most common (33 studies), as shown in Fig. 5(d). The remaining 40% did not report on the number of repetitions in their study, but were likely to have defaulted to a single-repetition design that involves only one instance of each monosyllabic word item.

B. Corpus design

Ideally, measures should be taken to safeguard the long-term preservation of primary data, with appropriate

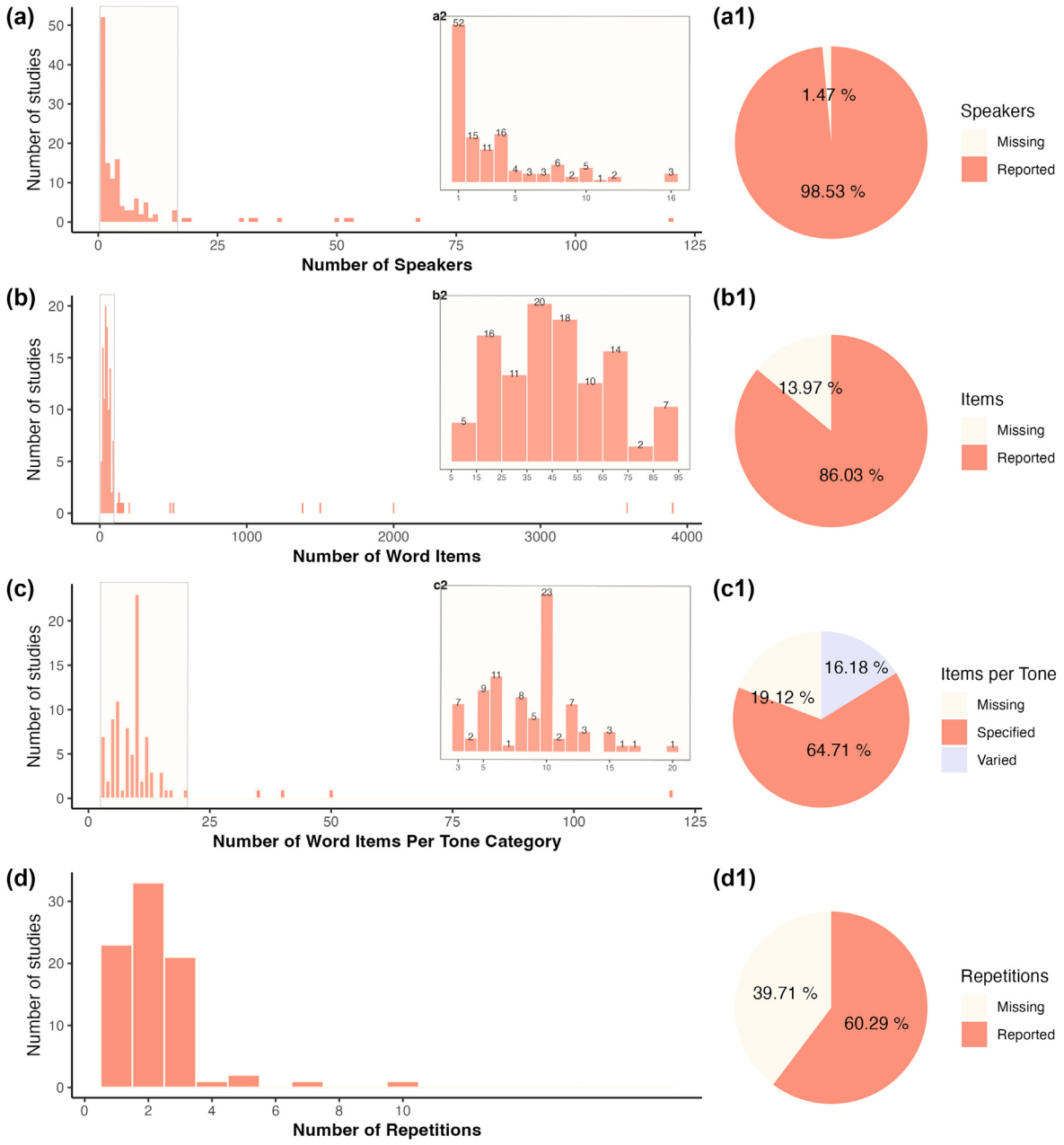

FIG. 5. (Color online) Distribution of speakers [(a) bin width = 1], monosyllabic word items [(b) bin width = 10], word items per tone category [(c) bin width = 1], and repetitions per word item [(d) bin width = 1] reported across studies in the database. The pie charts on the right [(a1), (b1), (c1), (d1)] illustrate the percentage of studies that provided the corresponding information, with "missing" indicating a lack of reporting. The embedded bar charts (a2), (b2), (c2) provide a zoomed-in view of the denser parts of the distributions, highlighted by the boxed areas in (a), (b), and (c), for better visibility.

consent, so that it can be employed in new theoretical or comparative ventures as well as in re-evaluating and validating established theories. In this context, metadata comprising annotations on the demographics of the speakers, their language backgrounds, the design of speech materials, and strategies for digital archiving and storage are essential for the organisation, reproducibility, integrity, and interpretability of the data in a study. Our survey revealed that the majority of papers in the database did not provide sufficient metadata.

1. Speaker demographics

Basic speaker demographics relevant to the pitch analysis of citation tones include sex, age, geographical location, language background, and language use. Additional demographic information can be valuable when the study is oriented towards sociolinguistic phenomena. The descriptions of speakers reflect how representative they are of the speech community for the language variety being studied.

Figures 6(a), 6(b), and 6(c) present the distribution of speaker selection based on sex as reported by the studies in the database. Only about half of the studies (68 studies) reported involving both male and female speakers. Surprisingly, of the eight studies that involved solely speakers of a single sex, all exclusively chose male speakers without justifications.

While most studies reported the sex (94%) and age (85%) of the speakers, only about 30% of the studies described whether they speak Standard Mandarin (Putonghua), as shown in Fig. 6(e). The use of Standard Mandarin is an indicator of their di-/multi-glossic or bi-/multi-lingual status, considering the language policy and increasingly widespread use of Standard Mandarin in China. In other words, it is important to know whether the speakers' speech in the target variety is potentially influenced by Standard Mandarin. Studies that did not report on Standard Mandarin use generally failed to provide any information on the language background of the speakers.

2. Speech materials

Many factors are shown to influence f0 such as vowel quality [e.g., Shi and Zhang (1986) and Whalen and Levitt (1995)], consonant voicing and aspiration [e.g., House and Fairbanks (1953)], the location of a syllable in speech [e.g., Selting (2007)], word frequency and ambient noise [e.g., Zhao and Jurafsky (2009)], and so on. These factors are potential sources of confound when we investigate the pitch patterns of lexical tones. Many confounding issues, however, can be addressed during the design phase of speech materials and in the setup of a production experiment.

To tease apart lexical tones from pitch variation attributed to other sources, the speech materials are ideally well-controlled minimal sets of monosyllables, in which their segmental compositions are the same with the tone being the only variable. There are, often though, tonotactic gaps (i.e., unattested syllable-tone combinations) in such minimal sets. In these cases, near-minimal sets in which the rhymes or vocalic parts remain consistent are considered. When designing speech materials, it is, therefore, crucial to consider a number of key aspects, including syllable structure, the selection of consonants and vowels, the presence of tonotatic gaps, and the usage frequency of the chosen monosyllabic words.

As is evident in Fig. 7, the rationale behind the choice of speech materials is often not well-explained in the studies. Only 40% of the studies (55 studies) described their choice of consonants, a mere one-third (46 studies) detailed their choice of vowels or rhymes, and slightly less than one-third (45 studies) of the studies controlled for the syllable structure. While 25 studies indicated the use of common monosyllabic words, only a handful addressed the tonotactic gaps: four studies left these gaps empty and seven studies filled them with syllables that had minimal differences. Without presenting the details and explaining the materials used in production experiments, the pitch patterns of citation tones remain difficult to be distinguished from other tonal effects.

3. Recording specifications

The technical specifications of digital audio recordings are important for any speech corpus, both for preservation and reproducibility purposes and for efficient audio processing and analysis. Figure 8 shows that the majority of the studies in the database did not specify the basic digital audio specifications including audio format, bit depth, sampling rate, and channel number. 32% of the studies (44 studies) reported to have used the WAV format for audio recordings. 34% of the studies (46 studies) reported the bit depth and all of them used 16 bits. 42% of the studies (57 studies) adopted the sampling rate of 16 000 Hz or above.

4. Recording procedures

Out of the entire database, only one study mentioned ethical approval. Most studies in the database did not adequately detail how speech materials were presented to participants. As Fig. 9 demonstrates, only 10% of the studies (13 studies) reported the presentation format, which varied from tables and prompt cards of Chinese characters, allowing speakers to anticipate the next morphemic sequences, to individual characters shown on a computer screen. The study of Qu and Yang (2019) on Benna Hani adopted oral prompts given participants' illiteracy in Hani orthography. 21% of the studies (28 studies) specified the sequence for reading out the characters: half of them had participants read from tables organised primarily by tone categories, while the other half arranged the characters in pseudo-randomised order. Additionally, 21% of the studies (29 studies) described the pause length between the production of characters, ranging from one to ten seconds, with most pauses which being two to three seconds.

Figure 9 also illustrates the distribution of studies according to the recording premises and the use of microphone. 17% of the studies (23 studies) recorded speakers in professional sound-attenuated rooms such as phonetics labs

https://doi.org/10.1121/10.0032356

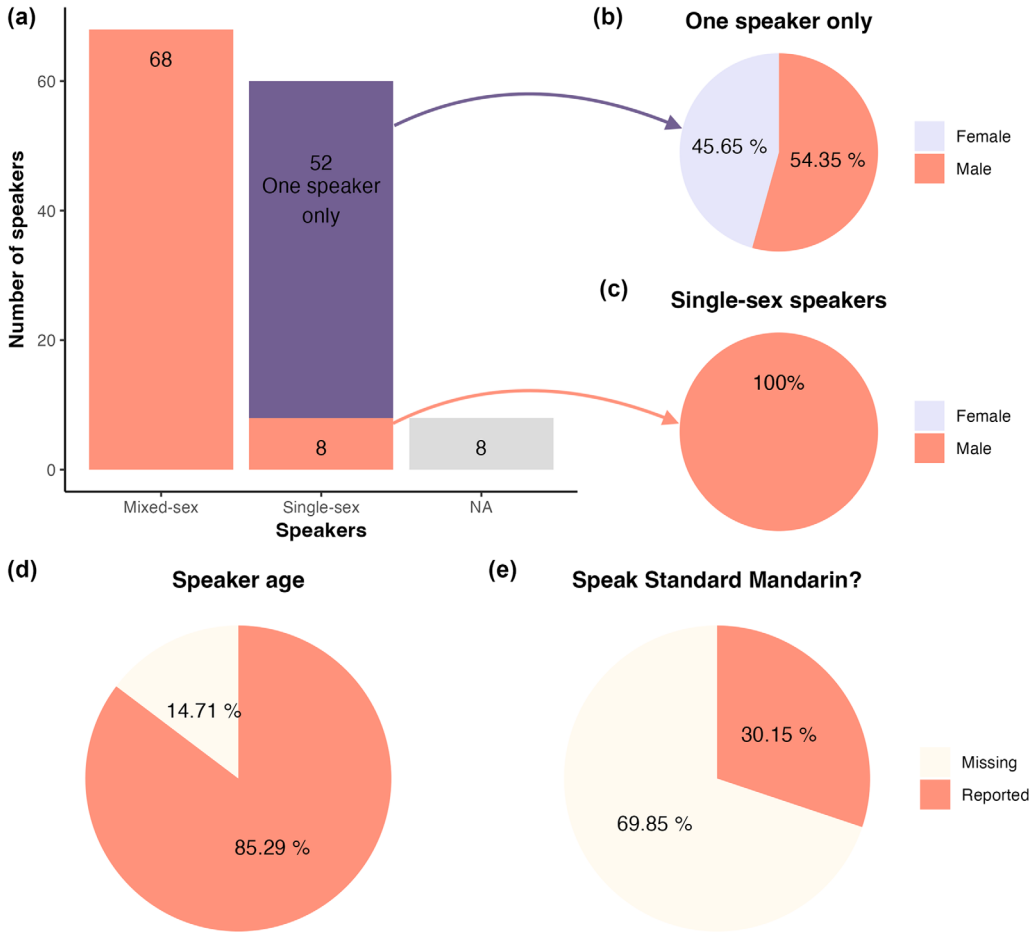

FIG. 6. (Color online) Overview of speaker demographics across studies. (a) Breakdown of studies by speaker sex composition, (b) speaker sex distribution in one-speaker studies, (c) speaker sex distribution in multi-speaker single-sex studies, (d) percentage of studies reporting speaker age, and (e) percentage of studies reporting speakers' Standard Mandarin use.

and anechoic chambers, whereas 56% (76 studies) of the studies did not mention the recording environment. 67% (91 studies) of the studies did not provide information regarding the use or absence of a microphone. Amongst the 45 studies reported to have used a microphone, the most frequently used types were headband microphones (14 studies), free-standing microphones (11 studies), and lavalier microphones (7 studies).

As the Sankey diagram in Fig. 10 illustrates, laptops and computers, allowing for augmentation by additional hardware and software, have predominantly served as the primary interfaces for recording speech in the studies within the database. ADOBE AUDITION, or formerly known as COOL EDIT PRO, emerges as the most frequently used software for recording (50 studies), followed by PRAAT (12 studies). Other recording devices include recorders (12 studies),

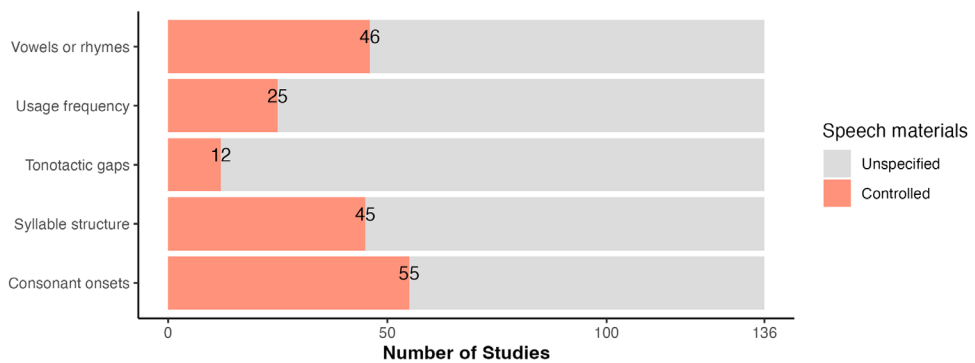

FIG. 7. (Color online) Number of studies reporting aspects of speech material design in the database.

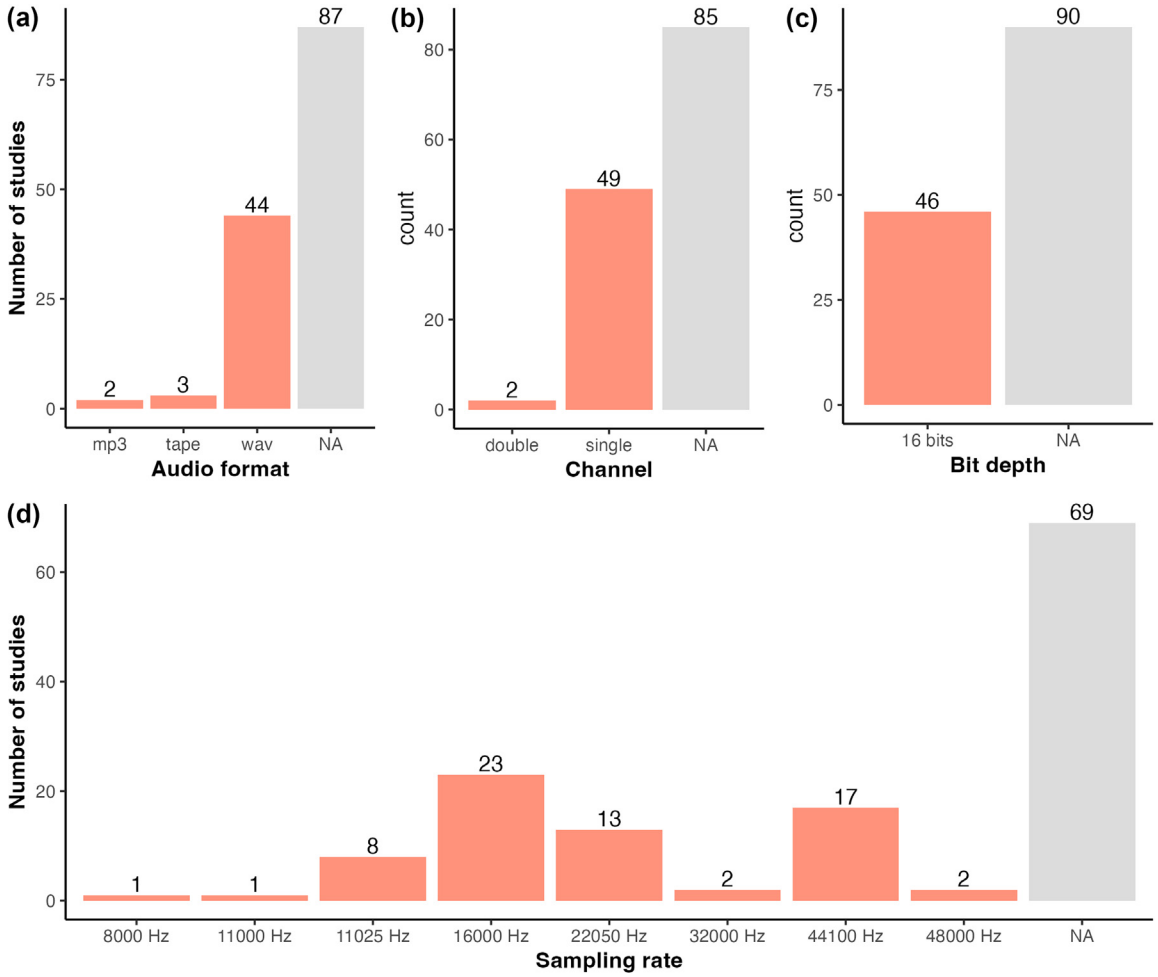

FIG. 8. (Color online) The distribution of studies in the database in terms of audio format (a), number of channels (b), bit depth (c), and sampling rate (d).

cassette tapes (3 studies), and mobile phones (2 studies, published in 2023). 26% of the studies (36 studies) did not introduce their recording devices and another 4% (6 studies) did not specify the recording software used when using computers as an audio interface.

C. Data analysis

The widespread availability of digital recordings, advancement in signal processing tools, and increased access to automated techniques have elevated acoustic analysis of the fundamental frequency as a major approach to tone research. 128 out of 136 studies involved acoustic analyses and the remaining 8 seemed to be purely auditory-based. The most crucial steps for acoustic analysis included the extraction, outlier treatment, and normalisation of the f0 values.

1. F0 extraction

Figures 11 and 12 summarised the software and methods for f0 measurements and normalisation, as well as steps of f0 preprocessing. PRAAT has been the most common software for f0 extraction amongst the studies (72% of the acoustic studies), although no PRAAT-based studies clearly specify which f0 measurement method was used, considering that PRAAT allows autocorrelation and cross correlation methods, with an optional prior low-pass filtering of the waveform. 9% of the acoustic studies (11 studies) did not at all report on the analysis software or algorithm used for f0 measurement. The mainstream approach to f0 extraction involved linear scaling of the time domain, whereby a fixed number of f0 measures, ranging from 3 to 30 [Fig. 11(c)], distributed at equal intervals. Among the 109 studies that adopted the mainstream approach [Fig. 11(d)], 83% (91 studies) extracted 9 to 11 f0 values per monosyllabic token, with 10 extracted f0 values being the mode (36 studies).

2. F0 outlier treatment

About 38% of the acoustic studies (49 in total) reported to have excluded or removed parts of the f0 contour, either before or after f0 extraction. These parts were analytically determined irrelevant to the defining characteristics of a citation tone, or considered as potential measurement artefacts. The strategies for removing such leading or trailing parts seem arbitrary and variable; for instance, Zhang (2022) removed the first 3 points from 23 equidistant f0 points, while Wan and Wang (2023) discarded the first and last two points from 20 equidistant f0 points. The majority of these studies provided only a vague description of the

https://doi.org/10.1121/10.0032356

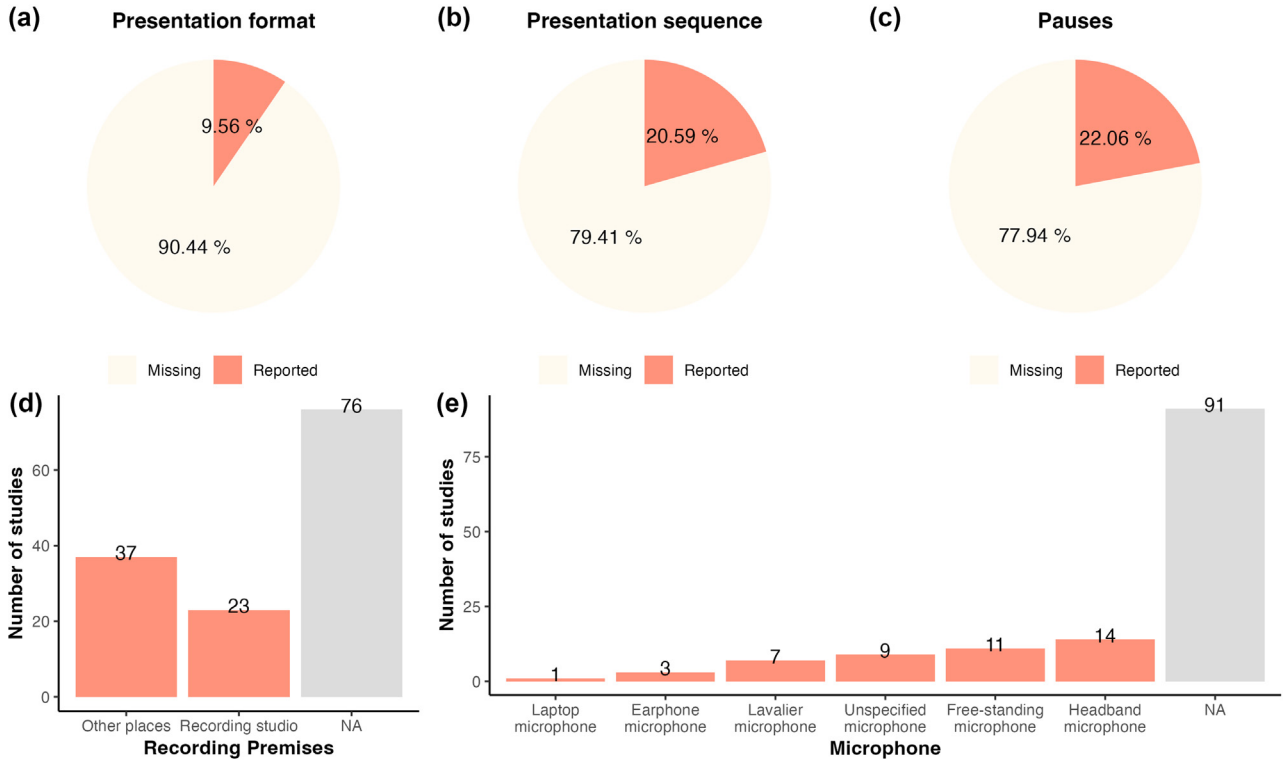

FIG. 9. (Color online) Percent of studies reporting the presentation format (a), presentation sequence (b) and pauses between utterances (c), and the distribution of studies in terms of recording premises (d) and microphone (e).

procedure, e.g., excluding the beginning and ending parts characterised by sharp f0 changes; however, such description does not suffice for reproducibility. In addition, 16% of the acoustic studies (19 studies) reported manual correction of the f0 measurement and 2.5% (3 studies) smoothed the raw f0 measures over time [e.g., Rose (2019) and Wang *et al.* (2020)].

3. F0 normalisation

Most studies normalise f0 measurements in Hz to reduce speaker variability, thereby facilitating the analysis of the tone patterns independent of individual speaker differences. F0 normalisation is mostly based on common normalisation techniques such as semitone distance, Z-score normalisation and Fraction of Range (Min-Max) normalisation (Rose, 1987), or a combination thereof, as shown in Fig. 11. Many studies also scale the normalised f0 to the range of $[0, 5]$ or $[1, 5]$ [e.g., Tian and Hong (2023)], so as to match the arbitrary 5-point scale in describing tone established in Chao (1930). The two predominant normalisation strategies are illustrated in Eq. (1a) or its variant (1b) with 67 studies, and in Eq. (2) with 27 studies, sometimes known as the T score (Shi, 1990) and LZ score (Zhu, 2004),

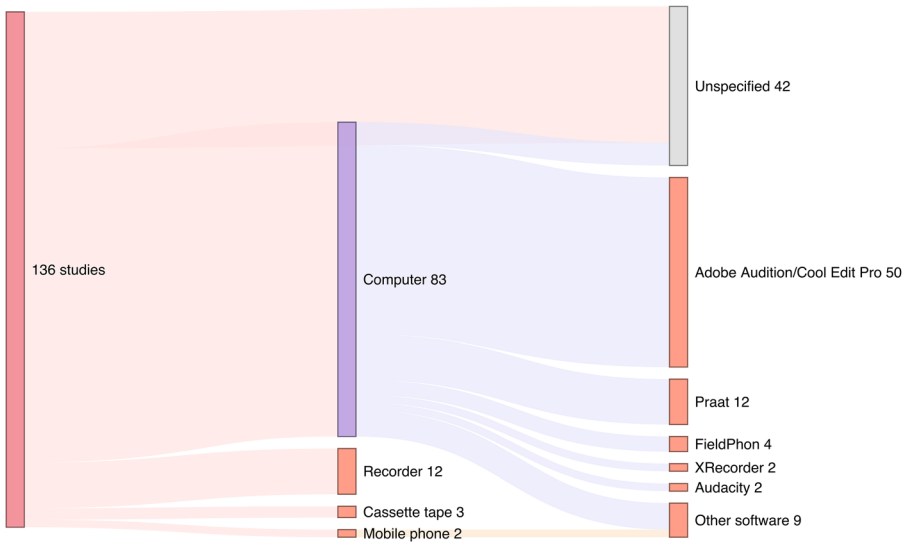

FIG. 10. (Color online) Frequency of usage in recording devices (middle panel) and software (if applicable, right panel) across all studies in the database.

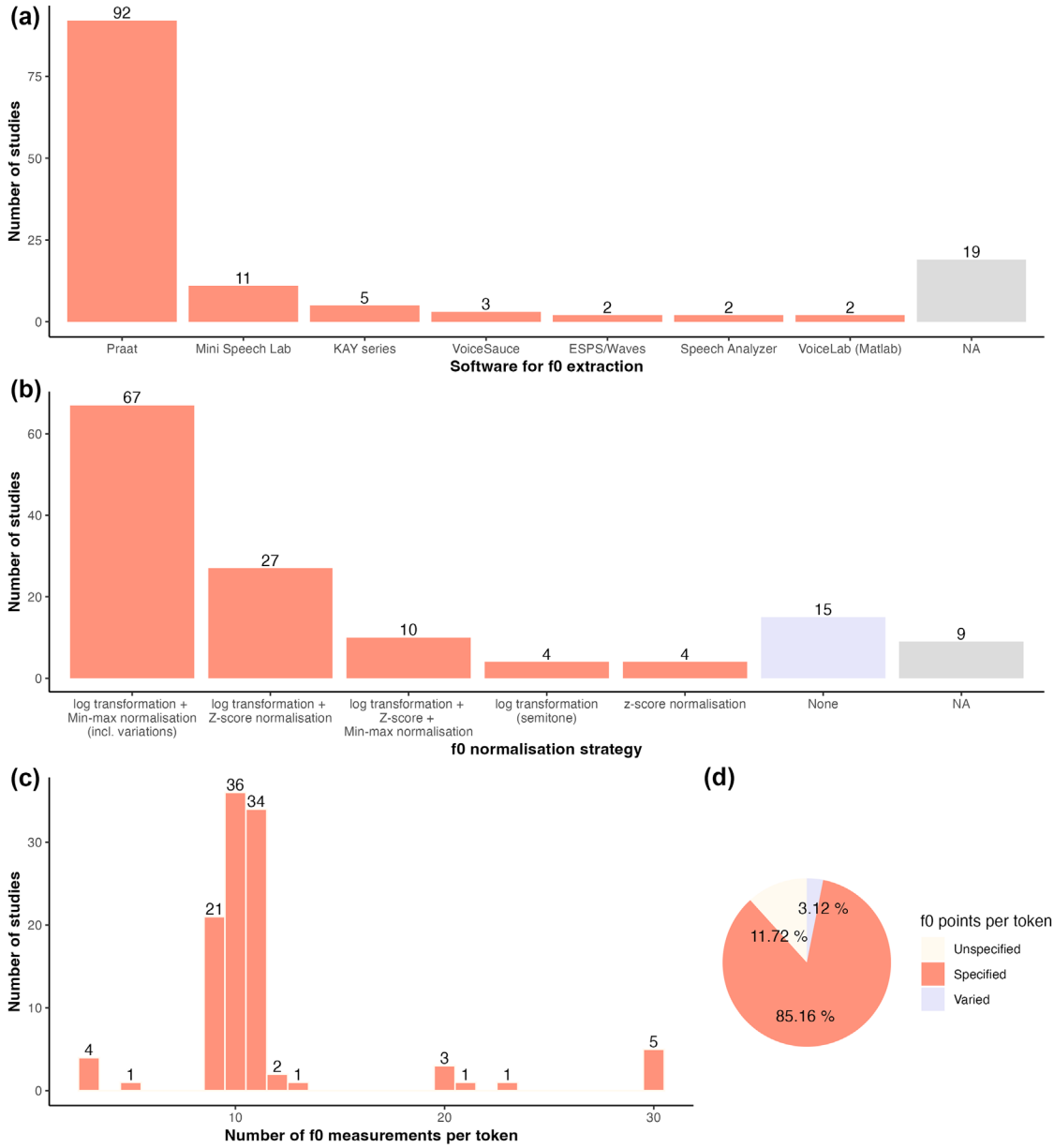

FIG. 11. (Color online) The distribution of studies in the database in terms of software for f0 extraction (a), method of f0 normalisation (b), and number of f0 measurements per token (c). (d) shows the percentage of studies that provided information on the number of f0 measurements per token.

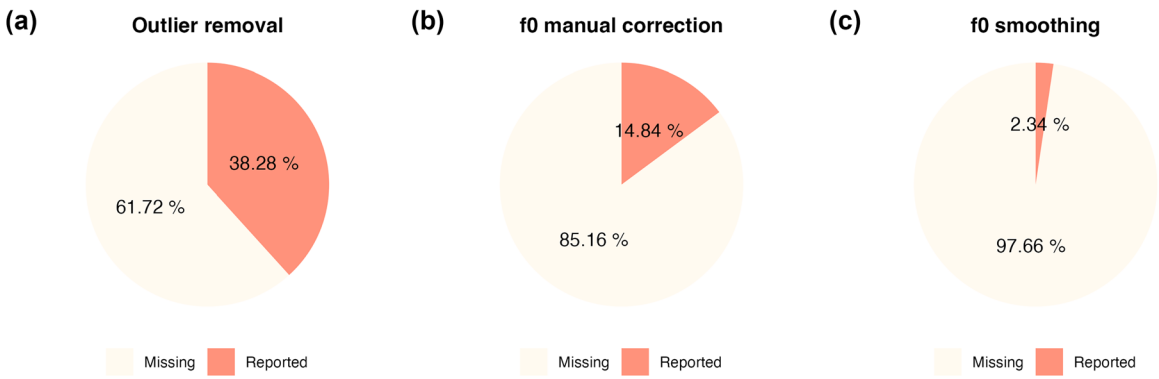

FIG. 12. (Color online) The percentage of studies that reported on f0 preprocessing in the database: (a) outlier removal, (b) manual f0 correction, and (c) f0 smoothing.

https://doi.org/10.1121/10.0032356

respectively. They are essentially the Fraction of Range and Z-score normalisation applied to the logarithmically transformed f0 values,

$$f_0^{'} = \frac{\log(f_0) - \log(f_{0\min})}{\log(f_{0\max}) - \log(f_{0\min})} * 5, \tag{1a}$$

$$f_0^{'} = \frac{\log(f_0) - \log(\mu_{f_{0min}} - \sigma_{f_{0min}})}{\log(\mu_{f_{0max}} + \sigma_{f_{0max}}) - \log(\mu_{f_{0min}} - \sigma_{f_{0min}})} * 5. \tag{1b}$$

In Eq. (1a), an f0 is transformed into a fraction of the difference between two range-defining logarithmic f0 values, usually the speaker's lower and upper limits of the tonal domain. It is, however, often not clear how these range-defining values are specifically defined. These can be, simply, the speaker's highest and lowest f0 values among all f0 data, or the average f0 values of the highest and lowest sampling points across all tones (Zhu, 1999). In the variant Eq. (1b), the range-defining values, for instance, are the average f0 of the highest sampling points plus one standard deviation of all f0 values at this time point and the average f0 of the lowest sampling points minus one standard deviation of all f0 values at this time point,

$$f_0^{'} = \frac{\log(f_0) - \mu_{\log(f_0)}}{\sigma_{\log(f_0)}}. \tag{2}$$

Another 10 studies combined Eqs. (2) and (1a), with logarithmic and Z-score transforms followed by a Fraction of range scaling.

D. Reporting citation tones

The majority of the studies (121 studies, 89%) reported citation tone results using the numerical equivalent of the Chao tone letter system, adopted by the IPA in 1989. The derivation of pitch variations into the five-level tone numerals, though, varies greatly among these studies. The most common strategy is manifested in the Fraction of Range normalisation methods illustrated in Eqs. (1a) and (1b), where $f_0^{'}$ is rescaled to the range of $[0, 5]$. This indicates that there are six reference lines from 0 to 5 (i.e., 0, 1, 2, 3, 4, 5) and the five intervals in-between represent the five pitch height zones, and each zone corresponds to a number in the tone stave. There is also a slightly different rescaling scheme as shown in Eq. (3), where $f_0^{'}$ is in the range of $[1, 5]$. In this case, there are five reference lines from 1 to 5 that directly define the stave,

$$f_0^{'} = \frac{\log(f_0) - \log(f_{0\min})}{\log(f_{0\max}) - \log(f_{0\min})} * 4 + 1. \tag{3}$$

In addition to variations in the rescaling scheme, different strategies exist for the assignment of tone numerals based on rescaled f0 values. Figure 13 shows how the tone numeral 2 was assigned from a rescaled f0 across the 77 studies that adopted the Fraction of Range normalisation methods. The main challenge lies in the handling of borderline $f_0^{'}$ values – those that fall close to a reference line in the interval-based tone stave system or those in the middle of an interval in the reference line-based tone stave system.

29% of these studies (22 studies) used clear-cut integer boundaries in the interval-based tone stave system. For instance, if an $f_0^{'}$ falls between 1 and 2, it is assigned a Chao numeral 2. This means that an $f_0^{'}$ of 1.05 is assigned a 2, despite being close to the reference line of 1, while an $f_0^{'}$ of 0.95 is assigned a 1. To reduce the ambiguity right at the integer boundaries, one study excluded the right edge (e.g., [1, 1.99]) and another excluded the left edge (e.g., [1.01, 2]). In a study that used reference line-based tone stave system, the boundaries were set in the middle of intervals. In this case, tone numerals 1 and 5 corresponded to a smaller range (e.g., $f_0^{'} \in [1, 1.5]$ for 1) than the other mid numbers (e.g., $f_0^{'} \in [1.5, 2.5]$ for 2).

Another popular strategy, used by 26% of these studies, employs an ambiguous boundary zone of ± 0.1 around the integer boundaries. For example, the Chao numeral 1 is assigned to $f_0^{'}$ values in the range of 0 to 1.1 and the Chao numeral 2 to $f_0^{'}$ values in the range of 0.9 to 2.1. In other words, an $f_0^{'}$ in the overlapping range of $[0.9, 1.1]$ is allowed two possible tone numeral assignment, either 1 or 2. Rather than the rigid boundary, this strategy provides a more nuanced and flexible method for assigning tone numerals. 42% of these studies (32 studies) did not include any details regarding the correspondence between the normalised f0 measurements and the iconic tone numerals.

IV. RECOMMENDATIONS FOR BEST PRACTICES

With the methodological variation and the absence of adequate methodological details observed in the 136 studies, this section proposes recommendations for future research on citation tones, from speech data collection to f0 data analysis, and directs future researchers to relevant resources. These recommendations are designed to provide an

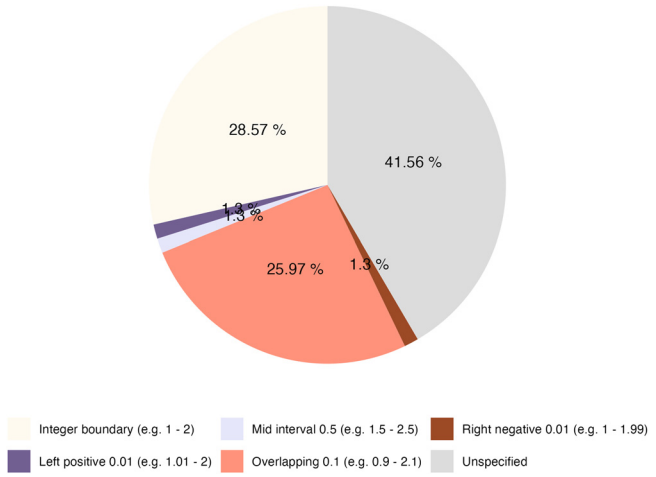

FIG. 13. (Color online) Percentages of studies using different numerical range of normalised f0 for the Chao numeral "2."

objective depiction of citation tones while ensuring the reproducibility of the analysis and the transparency of reporting.

A. Research ethics

Before embarking on linguistic fieldwork and data collection, it is worth carefully considering the ethical aspects. This typically involves the comprehensible communication of research objectives, potential impact, and instructions to participants, the establishment of protocols for seeking permissions for making speech recordings, the responsible and secure storage, dissemination, access, and use of the information and recordings, and the assessment of risks for both the researchers and participants. Additionally, participants should be informed of their right of withdrawal, fairly compensated, and acknowledged for their contributions. These may all be mandatory components of a formal research ethics application. For more guidance on ethical field research, see Bowern (2015).

Ethical approval should be secured when relevant institutional infrastructure, such as an ethics committee, is in place. Informed consent must be obtained from participants prior to any speech elicitation following the ethical approval. In explaining the research goals, you may avoid being too specific about the linguistic features of your investigation (e.g., tone, in our case), so that participants' attention would not be directed to these features in their speech, consciously or subconsciously. If participants request further information about the research, a debriefing can be arranged at the end of the recording.

B. Data collection

Most citation tone studies involve the elicitation of monosyllables via a speech production experiment and thus the development of a specialised speech corpus of monosyllabic words produced in isolation. This section delineates the exemplary practices in setting up speech production experiment and establishing the monosyllabic corpus.

1. Speaker demographics

a. Demographics. The ideal participants are the monolinguals/monodialectal of the language variety being studied, who have stayed in the region or speech community for a long time and use the variety as their primary medium of communication, with no history of speech or hearing disorders and no missing teeth. Such participants are, however, often hard to find, considering the useful and sometimes compulsory addition of a lingua franca such as English and Standard Mandarin to their linguistic repertoire. In this case, it is imperative to prioritise those who have substantial exposure to and predominant use of the targeted variety amongst other varieties.

The recommendation is to create and administer a sociolinguistic questionnaire[2] to participants in order to gather information about their language background, so as to contextualise the findings. The questionnaire includes, but is not limited to, information on their first language, knowledge of other languages, language proficiency, educational qualifications, and frequency of language use in various social contexts, along with basic personal information such as age and sex.

b. Number of participants. It is recommended to include multiple speakers with a balanced sex representation, thereby enabling a robust and inclusive dataset and controlling against individual idiosyncrasies. Ladefoged (1997) suggested having at least half a dozen people of each sex preferably. Ladefoged's recommendations are, however, not based on statistical considerations, as Roettger and Gordon (2017) emphasised. The minimum sample size allowing for statistically robust inferences over a speaker population depends largely on statistical power, which is determined by the true (or expected) effect size, the sample size, and the degree of variability (Vasishth and Nicenboim, 2016), and may be estimated from previous studies (Kirby and Sonderegger, 2018). The statistical power or the effect size, however, can be hard to define in pioneering and descriptive studies of citation tones. While there is no set answer to the required number of participants, which is largely dependent on the size of the speech community and the funding availability, a larger sample size is more desirable and has become ever increasingly feasible in the current era of big data, with the advent of affordable digital storage and automatic signal processing toolkits. However, we acknowledge that resource constraints pose significant challenges for many varieties, which should not preclude reporting findings from such varieties due to sample size issues (Lakens, 2022).

2. Speech materials

Three main factors critically inform and dictate the design of speech materials for citation tone studies: first, the knowledge of the tone categories within the language variety, as evidenced by consistent findings in existing literature, if any; second, the presence of a writing system for the language variety in question; and third, the literacy of the participants.

Speech materials typically consist of orthographic representation of multiple minimal sets where monosyllables only differ in tone. The selection of monosyllables should prioritise free morphemes that commonly occur in speech. For each tone category, it is recommended to incorporate a diverse array of items with varied segmental structures and compositions, thereby generalising over item-specific idiosyncrasies. Zhu (2010) advocated for the inclusion of syllables containing voiceless unaspirated stop consonants including /p/ and /t/ and monophthongs such as /a/, /i/, and /u/, due to the relative ease of isolating the tone-bearing part. Some other studies [e.g., Xu (1997)] preferred fully voiced syllables such as /ma/ with a nasal consonant onset to avoid the potential f0 perturbation brought up by voiceless consonants. Another common choice is vowel-only syllables.

https://doi.org/10.1121/10.0032356

These analytical choices, built on theories of the tone domain of the language variety being studied, may facilitate subsequent f0 analyses. The specific selection of syllables depends on the phoneme inventory and syllable structure of the language variety, as well as the research focus. In research with an emphasis on tonal variation and change, it may be beneficial to consider syllables with onset consonants varying in voicing and aspiration, as these attributes of onset consonants potentially play a significant role in inducing pitch differences. If there are tonotactic gaps, choose close minimal sets wherein the vowel or rhyme parts are kept consistent. An effective report on speech material design should present a full list of the selected monosyllabic words with IPA transcriptions and translations, as well as justifications for the selections. The materials are recommended to be displayed in a tabular form to clearly communicate different conditions, such as tone categories or syllable structures.

It is also advisable to include a number of repetitions, usually three times or more, so that spurious effects such as speakers' temporary lapses in concentration or occasional speech errors do not unduly influence the results.

In setting up production experiments for citation tones, it is important to carefully avoid tone sandhi, tonal coarticulation, and list intonation [e.g., a sequence of step accents in producing the berry list in Liberman and Pierrehumbert (1984)]. Rose (1993) observed an audible pitch difference between the first and last tokens on each prompt card and indicated that the speakers might have treated morphemes on each card as a single intonational discourse unit with its own declination. Xu (2022) instructed speakers to repeat each morpheme six times consecutively with pauses in between in their own pace and found that speakers naturally treat repetitions of the same morpheme as an organised list. This study further modelled the declination in the course of producing these monosyllabic lists.

When working on a minority language with scarce prior studies on its tones, an extensive selection of monosyllabic morphemes should be included for the purpose of encompassing the full spectrum of potential tone categories. If the variety has evolved from a historical prototype, the selection should cover morphemes representative of all historical tone categories.

In circumstances where a spoken variety lacks a fully fledged writing system, or when participants are mostly illiterate, researchers may resort to pictorial materials. An alternative is to employ question-answer interactions instead of a word list, guiding participants with verbal prompts to articulate the target monosyllables. For instance, if a participant mentions a target word in conversations, researchers can then invite them to repeat it in isolation. In these cases, the principles of controlling tonal contrasts and ensuring diverse item selection remain the same. Additionally, the discussion should take into account the potential influence of intonation.

3. Recording specifications

The most important recording specifications to consider are file format, sampling rate, bit-depth, and channel number, all of which determine the quality of the recorded signal and consequently the types and parameters of acoustic analyses. These parameters should be meticulously documented and reported in order for the study to be replicable.

a. File format. The audio recording should be in an uncompressed and lossless format, preferably the pulse-code modulation (PCM) waveform audio file format WAV, instead of a lossy format with a proprietary compression algorithm such as MP3 (Watt, 2013). There are other lossless file formats available, such as FLAC or AIFF, but we strongly recommend WAV, since more analysis tools support WAV.

b. Sampling rate. According to Nyquist–Shannon sampling theorem, to faithfully reproduce sounds, the recording sampling rate must be at least twice the frequency of the original sound. Many speech corpora utilise a sampling rate of 16 kHz [e.g., Huang *et al.* (1998) and Tang *et al.* (2021)], adequate for capturing speech sounds up to 8 kHz, which is sufficient for most speech analyses to date. However, it is recommended to record at a higher sampling rate when possible. Since human ears typically hear between 20 Hz to 20 kHz, recording at 44.1 kHz can cover the entire auditory spectrum.

c. Bit depth. Bit depth determines the number of possible amplitude values that can be recorded in each sample. Similar to sampling rate, higher bit depth creates recordings of higher quality. It is especially relevant to the difference between the low volume and high volume. For speech recordings, 16-bit is sufficient and recommended (Jones and Knight, 2013), but 24-bit can further reduce digital white noise if it is a viable option.

d. Channel number. A mono-channel recording is sufficient for capturing all useful information, especially when only one speaker is speaking at a time. Stereo recordings are usually not required in phonetic research unless the objective is to capture different speakers in different channels.

For citation tone studies, storage space is usually not a problem, allowing for the acquisition of high-quality recordings, for instance, 16-bit or 24-bit mono-channel recordings with a sampling rate of 44.1 kHz. Butcher (2013) suggests recordings at a minimum of 16-bit resolution and a 44.1 kHz sample rate, which is more than sufficient for most phonetic analyses but aligns with certain archival standards in the U.S. (p. 64).

4. Recording procedures

a. Premises and equipment. An ideal premise for speech recording meets two criteria: first, there is minimal background noise; second, there is little reverberation (Maddieson, 2001). An anechoic chamber or a soundproof recording studio is undoubtedly the best option for recording if available, with the caveat that the unfamiliar setting is less

likely to be the context for naturalistic speech of the language variety being studied, especially for a non-standard variety. A suitably quiet small room with soft furnishings in a local environment may be more practical and accessible. If possible, avoid spacious rooms with flat walls to minimise reverberation. Placing soft towels on hard surfaces surrounding the recorder and speaker is also a good way of reducing reverberation. Some noise sources can be eliminated including turning off the fan, air conditioner, or heater, closing the windows and curtains, switching off mobile phones, and preventing other people or animals from entering. These may not be the most comfortable setting and should be explained to the participants.

With the advancement of audio technology, portable devices for high-quality recording are ubiquitous nowadays. Generally, DAT (digital) recorders are more than adequate for linguistic phonetic purposes (Ladefoged, 1997). Directly recording onto laptops and even mobile phones may be considered, though the sound card is unlikely to be of the highest quality. In such cases, the recording software and its version number should be documented. The use of built-in microphones of these devices is, however, often not recommended due to their poor frequency response. For citation tone recordings, we work with one speaker at a time. A unidirectional microphone that primarily picks up sound from directly in front of its head is preferred, to mitigate background noise. An excellent setting would be fastening a unidirectional condenser microphone close to the speaker's mouth to a headset (Maddieson, 2001). We should avoid putting the microphone head right in front of the speaker's mouth or nose (usually slightly to the side or below the mouth), otherwise we capture excessive popping sounds or breathing noises. For setup guidance of lapel and free-standing microphones, see Butcher (2013), p. 65.

Remote data collection is feasible for citation tone studies, particularly when working with known tone systems that do not involve voice quality contrasts. The remote methods have generally found f0 tracking to be reliable [e.g., Sanker *et al.* (2021) and Zhang *et al.* (2020, 2021)], and see Zhang *et al.* (2024)[3] for a more detailed review of the reliability of remote recording methods on different acoustic measurements.

b. Procedure and etiquette. The word list should be pseudo-randomised so that each word occurs in various linguistic contexts to counterbalance potential priming effects. The randomisation also precludes rhythmic patterns arising from the regular repetition of the same word. Depending on the number of the selected monosyllables and their repetitions, which determines the total length of the experiment, it may be worth considering dividing the experiment into blocks, with breaks in between, where each block consists of a randomised set of a single repetition of the monosyllables.

In terms of the stimuli presentation, present one word at a time on a display with pauses randomly varying from 2 to 4 s so that participants cannot predict which word comes next and do not readout these words in a connected manner. This can be achieved in multiple ways such as POWERPOINT and PSYCHOPY (Peirce *et al.*, 2019). The time-varying pauses are conducive to preventing the list intonation often observed when speakers are asked to enumerate a series of items [Liberman and Pierrehumbert (1984), p. 171]. Embedding a target word in a carrier sentence is not recommended in citation tone studies due to the tone sandhi effects and tonal coarticulation. It is necessary to monitor the recording using headphones and check the signal level, i.e., the volume of the recorded speech signal, as speakers vary their loudness to avoid overloading the signal and audio clipping (Ladefoged, 1997). Guidance can usually be found on recording device or software manuals.

Before the reading task, it is beneficial to start with a "dry run" using non-target filler words to familiarise the speaker with the task, and set the appropriate recording level. In this dry run, pay attention to the presence of the 50 or 60 Hz Electric Network Frequency from the power supply, i.e., the subtle buzzing hum from an electrical device such as an old fridge or air-conditioner. If this is an issue, try recording without mains power. Most recorders can also function in battery mode. Make sure that there is a plentiful supply of batteries.

Researchers conducting the production experiments should, if possible, communicate with participants in the target language to maintain an authentic context for using the target language, otherwise recruiting a local assistant fluent in the target language is advisable. The experiment instructions should be standardised and presented in written form to ensure uniformity in information received by each participant. In cases involving illiterate participants, the same instructions can be read aloud.

It is also worth paying attention to the data transfer, storage, and backup methods. It is recommended that the recordings be backed up immediately, upon completion of a recording session. They should adhere to a consistent file naming system that anonymises the speakers, uniquely identifies each recording, and incorporates metadata about the recording, demographic, and technical information. The demographics of the participants and the technical details constitute the metadata for the specialised corpus, and they should be stored in non-proprietary formats such as Unicode plain text and comma-separated values (CSV). If any online transfer tool or cloud storage service is used, it is essential to check their privacy policy, server location, access controls, encryption options, retention and recovery policies. It should also be checked whether a copy will be stored and whether such a service complies with the regional standards and regulations, such as General Data Protection Regulation (GDPR) when dealing with data from the European Union.

C. Data analysis

This section offers insights into f0 data analysis, as f0 is the primary acoustic correlate for tone.

1. Acoustic preprocessing

When the recordings are of high quality, not much acoustic preprocessing is needed. In a citation tone production experiment, it is often convenient to record a speaker throughout a whole block or session. Forced alignment can be helpful in identifying and isolating individual syllables from a long recording and providing phone-level segmentation automatically. Available forced aligners for Standard Mandarin include P2FA (Yuan and Liberman, 2008), Montreal Forced Aligner (McAuliffe *et al.*, 2017), Charsiu (Zhu *et al.*, 2022b), with the capacity of training models for other varieties [e.g., Xu (2023)]. Training models for new varieties sometimes require a grapheme-to-phoneme mapping if the writing system is well-established [e.g., Min Chinese and Yue Chinese grapheme-to-phoneme models can be found in Zhu *et al.* (2022a)]. Note that downsampling may be required since common open-source forced aligners such as Montreal Forced Aligner (McAuliffe *et al.*, 2017) require 16-bit and 16 kHz WAV files. It is advisable to consider generating a downsampled set for the automatic alignment, while using the original high-quality recordings with the generated time alignments (e.g., TextGrids) for acoustic analysis.

For f0 analysis, one option is to perform an f0 tracking on the whole long recording and then query the resulting large pitch profile utilising the timing information. Alternatively, it may be beneficial to extract the speech interval of each monosyllabic token and save it as a separate audio file, to streamline the access, query, and management of individual sound tokens. With the syllable-level segmentation, each monosyllabic interval can be extracted. To prevent boundary-related information loss resulting from windowing in acoustic analysis, the boundaries for extraction should be at least the length of one analysis window beyond the actual syllable boundaries. This can be achieved through zero-padding, whereby a segment of silence, equivalent to at least a window length, can be added to both edges of each extracted sound interval. For instance, when the minimum periodicity frequency is set to be 75 Hz, PRAAT does not compute f0 values in the first or last 20 ms of each sound interval because the analysis requires a 40 ms window for every pitch frame. The analysis window can be 80 ms, twice the effective length, if the option *Very accurate* is set to be *on* in PRAAT. Therefore, to obtain f0 measures for a monosyllabic token using PRAAT in this case, we can zero-pad the edges of the token by 40 ms.

2. F0 extraction

There are a number of f0 extraction techniques including the autocorrelation-based methods, the cepstrum method, the linear predictive coding (LPC) method, the simplified inverse filter tracking (SIFT) method, the average magnitude difference function (AMDF) method, and so on (Oh and Un, 1984). In speech research, the default autocorrelation algorithm in PRAAT, *Sound: To Pitch (filtered ac)* (Boersma, 1993) in versions later than Version 6.4 or *Sound: To Pitch* in versions before 6.4, the RAPT algorithm *getf0* (Talkin, 1995) released in the ESPS (Entropic Signal Processing System) package (Entropic Research Laboratory, 2006) and in the REAPER (Robust Epoch Pitch EstimatoR) program, the STRAIGHT algorithm (Kawahara *et al.*, 2001), and the YIN algorithm (de Cheveigné and Kawahara, 2002) often emerge as the top candidates for f0 tracking. Despite their prevalence, there is no conclusive evidence on which is superior. Reichardt *et al.* (2016) showed that the ESPS (RAPT) algorithm outperformed the AMDF and PRAAT algorithms overall, although the differences among these algorithms were slight, while Strömbergsson (2016) stated that PRAAT outperformed ESPS and YIN. The comparative evaluation of f0 tracking for singing voices by Babacan *et al.* (2013) revealed that the efficacy of different algorithms depended on the error metric and recording condition. Specifically, PRAAT and RAPT excelled in determining voicing boundaries, RAPT reached the lowest incidence of gross pitch errors, YIN achieved the best accuracy, and STRAIGHT was the most robust to reverberation. All these algorithms are sufficient for citation tone studies, among which PRAAT may be particularly favoured since it is the most widely used algorithm in citation tone research according to our present review.

a. Which algorithm/software to use? Given the considerable variation in the algorithms for f0 tracking, it is critical to report not only the specific algorithm used but also its associated parameters or configuration, in order to increase the replicability of the research. For example, in the most recent version of PRAAT (version 6.4.01) at the time of writing, multiple methods are available for f0 extraction. Since 2023, the preferred method in the PRAAT manual for measuring tone and intonation in PRAAT has changed from *Sound: To Pitch (raw ac)* to *Sound: To Pitch (filtered ac)*, which applies low-pass filtering to the signal prior to autocorrelation, with the benefits of outputting fewer unwanted octave drops and rises. The corresponding parameters for finding f0 candidates involve time step (in seconds), f0/pitch floor (in Hz), f0/pitch ceiling (in Hz), the maximum number of candidates, and window length, each of which can significantly impact the f0 output. Tailoring the f0/pitch floor and ceiling for male and female participants may greatly help reduce the f0 tracking artefacts, especially when using the classic autocorrelation *(raw ac)* method. Additionally, parameters for the post-processing algorithm that seeks the optimal path through the f0 candidates include silence threshold, voicing threshold, octave cost, octave-jump cost, and voiced/unvoiced cost.

b. Which part to measure? While conflicting views are expressed concerning the domain of tone, it is advisable to obtain f0 measurements over the entire voiced part of each syllable at the f0 extraction stage, instead of precluding the examination of some parts of an f0 contour. The phone-level time alignment allows further investigation into the effects of different segments on f0 contours.

The relationship between the abstract representation of citation tone and its realisation across various segmental contexts remains unclear. There has been a debate on whether the tone in Mandarin is carried by the whole voiced part of a syllable (Chao, 1968), solely by the vocalic part of a syllable excluding any initial voiced consonants and final nasal consonants (Kratochvíl, 1968), or by the syllabic vowel and any subsequent voiced segments (Howie, 1974). The practice of excluding certain voiced segments in citation tone analysis, as advocated in several textbooks [e.g., Zhu (2010) and Bei and Xiang (2016)], seems to be a hasty conclusion drawn from the observed pitch variations associated with syllable structure, especially considering different language varieties may present different patterns.

More empirical evidence from a wider range of language varieties is needed to reach a definite conclusion on this topic. However, we do recognise that researchers can decide on which part of the syllable to analyse after carefully considering the analytical framework, the relevant literature specific to the language variety in discussion and the tonal patterns in their data.

c. How many f0 points? There are two common approaches for sampling f0 and the choice between them depends on the method of f0 analysis and the specific research questions being addressed. One is to sample a tone contour in a fixed number of equidistant f0 estimates. The f0 sampling rate is recommended to be 11 or 21 per tone, so that f0 is sampled at every 10% or 5% of its time course, respectively, sufficient to capture the shape of a tone contour. This approach leads to time-normalised f0 data for each monosyllabic token and enables comparison of f0 across different utterances at various relative time points. The other is to sample a tone contour in a fixed time step, usually 10 ms. This approach preserves the original length of a tone contour in the f0 data and typically provides better time resolution, useful for identifying portions of problematic f0 in a tone contour and for curve-fitting.

3. F0 outlier treatment

a. What is an outlier? In citation tone studies, outliers may arise from f0 measurement errors or represent phenomena of lesser interest from the viewpoint of tone patterns, such as consonantal perturbations of f0. Rarely is the identification and treatment of f0 outliers documented in phonetic journal articles and most often it is reduced to a brief mention that f0 measurements were manually checked and corrected, without further details or insights. The core challenge, therefore, is to navigate the trade-off between the necessity and amount of manual correction and the detrimental impact outliers have on f0 analysis, especially on the shape of f0 contour. Sometimes manual correction may not be feasible in very large datasets. In such cases, other methods can mitigate the risk, such as data point interpolation and smoothing.

b. Issues with arbitrary trimming. Previous studies often trim an arbitrary portion of the f0 measurements at the edges of the vowel or syllable nuclei, serving as a quick automatic approach to reduce f0 perturbations from neighbouring consonants or in the initiation and termination of the speech. For example, in Keating and Kuo (2010), the f0 measurements for the first and last 2% of each target interval were discarded; in Tang *et al.* (2019), the initial and final 5% of the vowel were excluded; in Rose (1987), it was the first and last 10% of the duration of each tone track; and in Stanford (2008), as much as 25% of the beginning portion of a tone token was trimmed. This is also evident in the variable strategies employed for the exclusion of certain f0 measures by studies in our current database, as discussed in Sec. III C 2. The arbitrariness in deciding how many f0 values to omit does not ensure its effectiveness across tokens, and the variability across studies undermines the comparability of f0 contours.

c. How to identify outliers? It is crucial to first identify potential f0 outliers, so that bespoke methods can be used to address them. Any corrections made to f0 measures should be logged for version control, allowing researchers to track modifications and ensure replication. A two-stage screening of f0 measures is recommended as follows: first, summary statistics of f0 for each monosyllabic token, including mean, standard deviation, minimum, and maximum are calculated. Tokens with a very large standard deviation, or an unusually high maximum or low minimum can be flagged for further check of their waveform and spectrogram. However, we do not recommend trimming the f0 values solely based on these summary statistics. Second, f0 tracks of each speaker, grouped by tone categories, are plotted in an overlaying manner, preferably in semi-transparent colours, for visual inspection.

The visualisation is demonstrated via scatter plots in R in Fig. 14, where the entire f0 track of two tokens: one [u] in Tone 2 and one [ma] in Tone 3 deviate from the typical f0 tracks in their respective tone categories. Their f0 estimates are approximately half those of other repetitions of the same syllables, suggesting possible pitch-halving errors or diplophonia. These two tokens can then be further examined and corrected. Additionally, other potential f0 outliers such as sudden jumps over 300 Hz or drops below 150 Hz are markedly shown in Fig. 14.

In particular, the interim visual inspection of the raw f0 data should focus on the following aspects:

(1) **Presence of octave shifts**. Octave shifts can manifest as pitch-halving or doubling errors, or indicate a diplophonic voice characterised by two concurrent periodicities. Common f0 tracking issues such as the octave shifts are readily identifiable through visual inspection, as shown in Fig. 14. The shifted f0 outliers may be manually corrected, for instance, in PRAAT, by selecting alternative f0 candidates in the *Pitch Object*, or by adjusting the parameters of the post-processing algorithm.

https://doi.org/10.1121/10.0032356

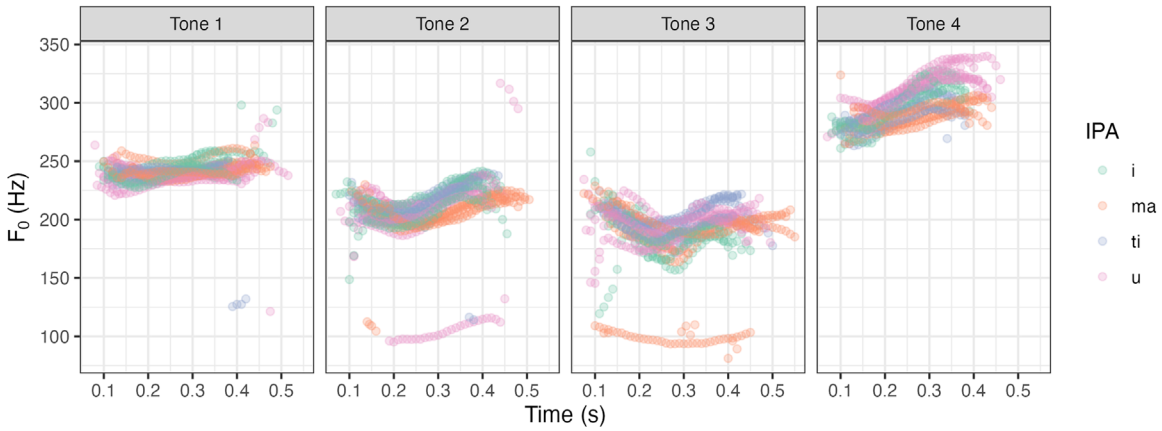

FIG. 14. (Color online) Visual inspection of the f0 tracks of monosyllables /i/, /ma/, /ti/, and /u/ produced by a speaker of Plastic Mandarin.

(2) **Acoustic validation of tone elicitation**. The plot also allows us to check whether speakers of the same variety indeed produce utterances that exhibit similar tone patterns and to identify any speakers who demonstrate noticeable deviations from the majority. This serves as a validation of the successful elicitation of the tones in the target language variety. For speakers with deviating pitch patterns, further auditory and perceptual validation of these utterances is needed to determine whether they are equivalent pronunciations, individual idiosyncrasies, style shift, or accent variation. Utterances with style shift or accent variation should be excluded from the analysis.

(3) **Consistency of intraspeaker tone patterns**. It is important to assess the uniformity of intraspeaker tone patterns within an assumed tone category, particularly in non-standard varieties. A ubiquitous phenomenon among Chinese varieties, especially in non-Mandarin regions, is that a single sinogram or morpheme may have differentiated literary and vernacular readings (Ho, 2015). The doublets of readings result from language contact and reflect lexical stratification under the influence of the standard language of education and literary instruction at various historical periods (Norman, 1979). If two distinct tone patterns are elicited, it will be necessary to represent them separately in the analysis.

Non-modal voice, particularly creaky voice, tends to be the leading cause of problematic or unreliable f0 measurement. Often characterised by its lower frequencies, creaky voice may not be accurately estimated when the f0 values are below the predefined pitch floor setting. In such cases, pitch trackers might mistakenly rely on minor individual peaks within the heavily damped long pulses typical of creaky voice, yielding substantially high f0 estimates. This is why sometimes there is an upsurge in f0 at the right edge of a token, reflecting creaky phonation related to the termination of an utterance. The first-pass f0 measures of extremely low-pitched tokens can be corrected by adjusting the pitch floor setting.

d. How to deal with laryngealised tokens? Creakiness or laryngealisation is also typically accompanied by relatively high jitter and shimmer. Sometimes encountered are tokens containing portions with irregular pulses, where f0 is likely to be not available or unreliable. Figure 15 demonstrates three naturally occurring instances of [i:] "ant" with a low dipping tone by a Standard Mandarin speaker, two of which feature partial non-modal phonation: the middle token shows modal-creaky-modal pattern, with the creaky portion characterised by a low rate of vocal fold vibration and irregular f0; the bottom token includes a full closure in the middle of the utterance. These variable acoustic realisations, while potentially having an equivalent percept in cuing the same tone category, present challenges in f0 measurement and analysis.

One challenge lies in accurately computing f0 estimates when sampling the f0 time course at a fixed number of equidistant data points. The f0 estimate at a particular time is usually computed by interpolating between the surrounding f0 measures of the pitch tracking frames. The interpolation function, hence, affects the f0 estimation. Rose (2019), for example, proposed a dual-model approach for f0 extraction in laryngealised tokens, exemplified by the middle token in Fig. 15. Separate models were used for the central creaky portion and the peripheral modal portions: a *lowess* model for the former and a low order polynomial model for the latter. The f0 was then sampled from these smoothed curves. This method resulted in a close fit to the f0 measures.

Another challenge relates to the fact that the precise f0 of the laryngealised portion of a token may not correspond to our perceived pitch in the same way as it does for the f0 of modal voice. Kuang and Liberman (2015) suggests that the perceived pitch is much higher than the actual f0 in cases of creaky phonation. Figure 16 illustrates the f0 data points extracted every 10 ms for the three variants of [i:] in Fig. 15. For each token, a *loess* fit, a smooth curve fitted to all f0 points using local regression, is depicted with a dotted line, and a quadratic polynomial, fitted specifically to the modal voiced f0 points in the dipping portion of the curve, is drawn with a solid line. Although with only peripheral modal portions in the two laryngealised tokens, the shape of their fitted

quadratic polynomials closely mirrors that of the fully modal token. In contrast, the *loess* fit, sensitive to the central irregular f0 measures, results in a noticeably different shape in one of the laryngealised tokens. The close fit to the creaky f0 measures, in turn, greatly influences the claims we made about the pitch pattern of a tone. Considering that the strikingly large dip in the f0 curve modelled by the *loess* fit in Fig. 16 appears not to be perceptually significant, does the quadratic polynomial model based on modal portions provide a more perceptually valid capture of the tone pattern?

Tone is not only realised through changes in pitch but also frequently by changes in voice quality. This additional phonetic realisation has been documented across a variety of languages such as Vietnamese, Hmong (Garellek *et al.*, 2013), Yucatec Maya (Frazier, 2013), and Valley Zapotec (Esposito, 2010). Yet the traditional approach to lexical tone analysis tends to exclude non-modal tokens due to the absence of reliable automatic f0 data. While f0 data of tokens with partial non-modal voice portions may be obtained via modelling the f0 curve and interpolating as previously described, it is crucial to report the f0 sampling approach and the relevant smoothing methods. In addition, the variability in acoustic realisation demonstrated in Fig. 15 reflects the importance of auditory analysis, despite the decline of impressionistic auditory transcription in the era of automatic f0 tracking. The auditory analysis provides a sensible heuristic to the perceived tone pattern. It is recommended to document the perceived pitch pattern as well as any non-modal voice quality observed in citation tones.

4. F0 Normalisation.

In speech, the f0 conveys not only the linguistically relevant information, but also the physiological, emotional, and sociolinguistic characteristics of the speaker. In citation tone studies where the interest is to map out the general tone patterns of all language users in a speech community, individual anatomical differences and idiosyncrasies are treated as noise (Genette *et al.*, 2023). The same contour tone produced by a man and a woman, for instance, is highly likely to have different pitch spans when plotted on a Hz scale. Normalisation is a mathematical tool to eliminate or reduce such noise and to allow between-speaker comparison.

There are a variety of normalisation methods for f0 from published studies, as shown in our database (Sec. III C). Zhang (2018) and Genette *et al.* (2023) curated comprehensive lists of specific methods. These methods can be categorised into four broad strategies:

(1) **Perceptually motivated rescaling**. This category involves transforming an f0 measured in Hz to a perceptually motivated non-linear psychoacoustic scale such as semitone [e.g. Chiang (1967)], mel, Bark, and ERB [e.g., Nolan (2003)]. The semitone conversion (log to base 2) often employs a fixed reference f0, such as 1, 50, or 100 Hz.

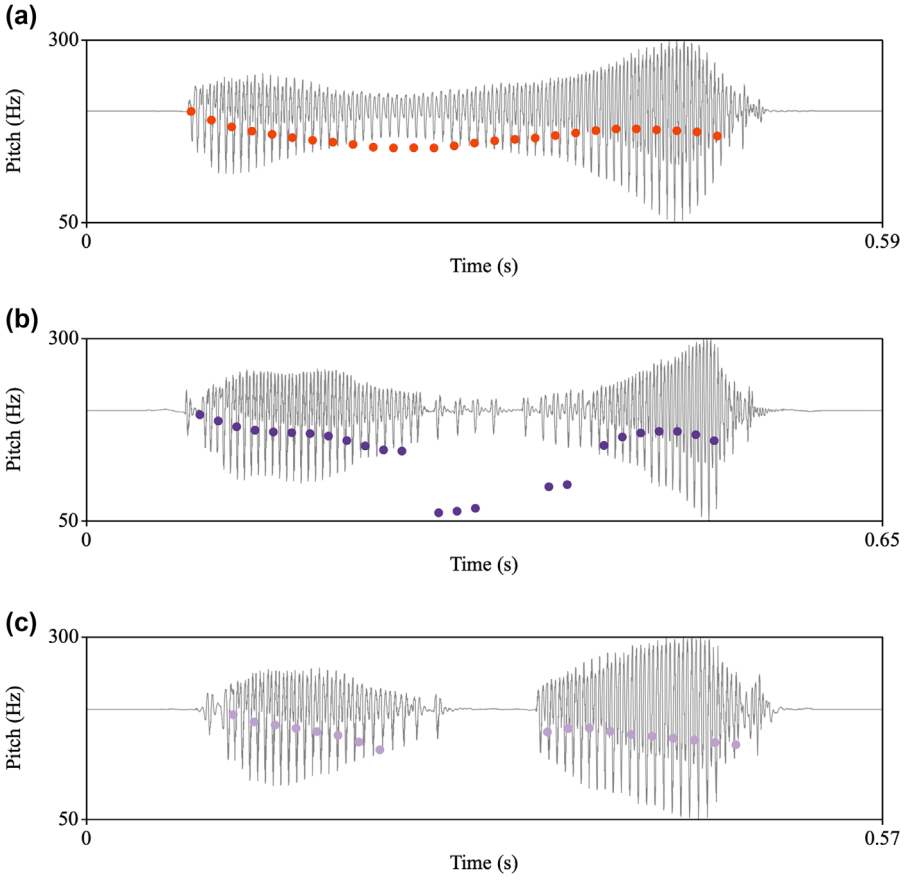

FIG. 15. (Color online) The waveform of three tokens of [i:] ant in Standard Mandarin dipping tone by the same speaker, with f0 tracked by Praat superimposed. (a) modal voice with regular pulses throughout the syllable; (b) creaky voice in the middle of the syllable; (c) a short "pause" (glottalisation) in the middle of the utterance.

https://doi.org/10.1121/10.0032356

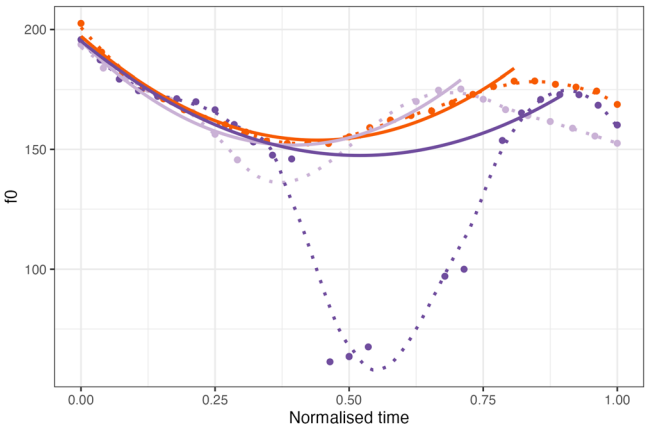

FIG. 16. (Color online) The fitted quadratic polynomial of the dipping portion based on the modal f0 points (solid line, the five lowest purple dots were excluded) and the fitted local polynomial regression based on all f0 points (dotted line, smoothing = 0.45). The f0 measurements of the three tokens are shown as a reference. The tokens match those in Fig. 15 by colour.

(2) **Range normalisation**. This category includes Fraction of Range transforms [e.g. Ladd *et al.* (1985)], representing an f0 as a fraction of the difference between the maximum and minimum f0 values, and other linear scaling methods such as rescaling an f0 between 0 and 1 [e.g., Hoole and Honda (2011)].
(3) **Centroid normalisation**. This category normalises f0 relative to a central tendency of the f0 distribution such as medians [e.g., Chen *et al.* (2021)] and means. The most common method is Z-scores [e.g., Menn and Boyce (1982)], transforming an f0 as a multiple of a measure of dispersion sway from a mean f0 (Rose, 1987).
(4) **Log-based combined normalisation**. This category combines logarithmic transformation with other centroid and/or range normalisation methods. For instance, Andruski and Costello (2004) converted f0 in Hz to semitones relative to each speaker's average f0 (log-mean normalisation). Other examples include the T score and LZ score introduced in Sec. III C.

Having mentioned the reasons for normalisation, we would also like to flag the risks of doing so. The *a posteriori* rescaling of the original frequency measurements in Hz can impact the variability in the data. Consequently, normalisation potentially affects hypothesis testing and the power of statistical tests (Genette *et al.*, 2023). There are different opinions regarding the choice of normalisation methods. Rose (1987) preferred Z-score normalisation to Fraction of Range normalisation because Z-scores more effectively reduced the between-speaker variance of f0 samples. Similarly Zhu (1999) considered logarithmic Z-score the best among six methods since it resulted in the most tightly clustered normalised f0 contours. Upon evaluating the effectiveness of 16 tone normalisation methods from three perspectives including their ability to preserve phonemic variation, minimise anatomical variation, and maintain sociolinguistic variation, Zhang (2018) concluded that the semitone transformation relative to each speaker's average pitch (log-mean) is the best method. Genette *et al.* (2023) suggested that f0 measures normalised by the log-mean, centroid, and range methods with per-subject reference levels achieve higher statistical power with smaller sample sizes compared to unnormalised and perceptually rescaled data. In particular, the log-mean method enhances the statistical power the most when the sample size is small.

a. How to calculate the speaker mean f0? It is important at this point to discuss a caveat associated with the calculation of the speaker mean f0 in a tone language, which is utilised in many f0 normalisation methods presented here. When the dataset is unbalanced, that is, the number of syllables per tone category is not the same, the composition of different tones in the speech materials influences the grand mean (averaging f0 across all syllables). For instance, if the speech materials contained a disproportionately high number of high tones, the estimated speaker mean based on the grand mean would also be artificially elevated. One way of defining the speaker mean f0 was averaging the f0 mean of each tone in a language variety using the monosyllabic dataset [Eqs. (4) and (5)],

$$\bar{f}_T = \frac{1}{n}\sum_{j=1}^{n} f_{0j}, \tag{4}$$

$$\bar{f}_s = \frac{1}{N}\sum^{N} \bar{f}_T. \tag{5}$$

Here, $\bar{f}_T$ is the average of all n number of f_0 values of Tone T syllables in a given language variety by a given speaker (T denotes an arbitrary tone category; j is the index of each f_0 measurement of a tone T syllable in a given variety by a given speaker). In this way, the speaker mean, denoted as $\bar{f}_s$, represents an equally weighted mean across N number of tone categories and can be interpreted as the centre of the tonal space, encompassing every type of citation tone.

Built on the discussion of f0 normalisation, selecting an appropriate f0 normalisation method should take into account the research objectives, the size of a dataset (statistical power), and the intent to compare findings with other relevant tone studies.

b. How to normalise time? When examining f0 contours, the time dimension cannot be neglected. With regard to the normalisation of time in f0 contours, the following should be considered: (1) whether the duration of citation tones is part of the research objective, (2) whether duration significantly and systematically varies among the citation tones, and (3) the analysis and visualisation method of f0 contours.

Often, we are interested in the shape of a tone contour, especially in Chinese varieties, regardless of the duration variation. In most cases where f0 values are equidistantly sampled with a fixed number (e.g., 11), the lengths of the f0

contours of the different tones are equalised. The time indices of f0 points can then represent a linear progression in duration expressed as percentages (e.g., 0%, 10%, , 100%, when there are 11 sample points).

Sometimes a particular analysis method necessitates data to be preprocessed in a particular format. For instance, the use of Lengendre polynomials in modelling f0 contours requires shifting and scaling the time axis to range between −1 and 1 because Legendre polynomials were defined as an orthogonal system over the interval $[-1, 1]$ (Grabe *et al.*, 2007). There are nevertheless cases where duration is a salient feature of citation tone patterns. Duration can be measured and integrated into f0 contour analysis, such as with generalised additive mixed models (GAMM).

5. Statistical analysis and visualisation

The f0 data of each monosyllabic token are dynamic trajectories over time, and the overall dataset is hierarchical in nature: while individual measurements at various time points were arranged into f0 contours, the contours can be grouped according to syllables, tones, and speakers, resulting in complex dependencies among individual measurements. The main goal of f0 analysis and visualisation is to establish the prototypical pitch pattern of citation tones, taking into careful consideration the hierarchical data structure.

Functional data analysis has emerged as a valuable and effective collection of tools across various fields, including modelling f0 data registered on a continuum of time. Hadjipantelis *et al.* (2012), for instance, treated f0 contours of Mandarin syllables as bounded continuous curves. These smooth curves can then be modelled as a sum of weighted basis functions, including quadratic spline basis functions (Hirst and Espesser, 1993), cosine functions (Watson and Harrington, 1999), or functional principal components (Gubian *et al.*, 2015). Grabe *et al.* (2007) and Xu (2022), for instance, employ parametric orthogonal Legendre polynomials as basis functions. This approach projects f0 curves in the lower dimensional space where the Legendre polynomials served as the axis system. The f0 contour of a single syllable, which is not expected to be very wiggly, can be easily represented by a small number of interpretable and meaningful coefficients: the first three coefficients correspond to the average f0, the best-fitting slope, and central curvature, respectively. It is desirable to keep the degree of polynomials low so that all coefficients are straightforward to interpret and understand, and potentially filter out unstructured information. The coefficients, serving as proxy data for the f0 curves, can be used as the dependent variables in a series of linear mixed effects models, thereby effectively modelling the shape features of various tone prototypes.

Alternative approaches such as growth curve analysis (GCA) (Mirman, 2014) and GAMM (Wood, 2017) integrate mixed models with the modelling of nonlinear patterns. While the former is polynomial-based, the latter is non-parametric, inherently more flexible in modelling nonlinearity. GAMMs can also accommodate the autocorrelation observed among consecutive time points in time series data, though GAMMs do not have directly interpretable coefficients, in contrast to GCA (Winter and Wieling, 2016). For accessible introductions to GCA, see Mirman (2014); GAMMs, see Sóskuthy (2017, 2021) and Chuang *et al.* (2021). For more f0 contours modelling techniques, see review in Zhang and Chen (2024).

a. How to report findings of citation tones?

Visualising the prototypical f0 contour of each tone is a straightforward and effective way to communicate the citation tones of a language variety. Typically, the average normalised f0 contour of each tone is plotted against normalised time. When mixed models such as GAMMs are used in the analysis, model predictions of normalised f0 over normalised time for each tone can be plotted instead. Visual methods are in fact key to interpreting GAMMs output (Sóskuthy, 2021). We recommend incorporating confidence intervals in the f0 contour plot to capture variability across tokens and speakers. Alongside the visualisation, we also encourage a brief textual summary (e.g., high rising) of the prototypical citation tones. If tone patterns are to be summarised in Chao's tone letter or numeral system, it is important to explain the f0 rescaling scheme and the boundaries used to define the tone categories.

D. Open and reproducible research

We encourage researchers to adopt open and reproducible practices such as sharing study materials, data, and analysis scripts, which enable replication and evaluation, as well as greatly increase the comparability of citation tones cross-linguistically or longitudinally.

V. CONCLUSION AND FUTURE ENDEAVOURS

This paper consists of two main components: (1) a systematic review in Sec. III that identifies methodological issues in citation tone production studies and (2) a thorough collection of recommendations in Sec. IV to guide future research.

Our review reveals that many studies in our database lack sufficient methodological details. Methods on determining and distilling prototypical citation tones also differ significantly across studies, hindering reliable comparisons. In every stage of the research pipeline—data collection, processing, analysis, and interpretation—various analytical decisions researchers made can affect the shape of f0 contours and, in turn, the conclusions regarding the citation tones of the language variety.

Our recommendations identify the potential pitfalls of certain practices and highlights the advantages of a variety of approaches, instead of aiming to offer a one-size-fits-all solution. The discussions on f0 analysis, such as f0 outlier treatment and normalisation in Sec. IV, are also beneficial to the broader field of linguistic research, where f0 plays a critical role. We hope that the recommendations will encourage

https://doi.org/10.1121/10.0032356

more careful experiment designs and more thorough reports of the findings in future studies. Adhering to these recommendations will likely enhance the reliability of results within a language variety and facilitate typological comparison and discussion. We have also prepared a checklist in the Appendix to guide researchers throughout the research pipeline.

There are, nevertheless, still unresolved issues in citation tone research that call for the attention of phoneticians, phonologists, and typologists in future endeavours.

One issue concerns the effectiveness of the notation system of citation tones, particularly the widely used tone numerals. The comparability of citation tones across languages is greatly compromised by how tone numerals were derived in different varieties: different studies used different formula and cut-off points for obtaining the final tone numbers. Apart from this caveat, we do recognise that the tone numeral system has been working well in capturing the contrastive lexical tones within a variety and is easy to conceptualise. However, despite its effectiveness and simplicity, the 5-level numerical system has its weaknesses in capturing surface variation: the distinction of one degree (e.g., 44 and 55) is sometimes not significant and the choice of five levels is not motivated by any phonological principles, which makes it lie in an ambiguous status between a phonetic system or a phonemic one [see Duanmu (2007) for discussion]. An alternative approach is to present the lexical tones in various language varieties through a visual database of the prototypical citation tones. This may facilitate cross-linguistic or diachronic comparison. Each prototypical tone can be summarised as a vector of (at least) three numeric values, its central curvature, best-fitting slope, and mean height, using orthogonal polynomial modelling, thereby each tone is projected as a dot in the three-dimensional space where a similarity index can be created. An example can be found in Xu (2022).

Another issue lies in the interpretation of the f0 contour of a whole syllable and the controversial domain of a tone (also discussed in Sec. IV C 2). In the database, a segment of varying size may have been excluded from an f0 contour. How do we prove that these excluded f0 segments present in the acoustic signal of the citation tone production are not part of a tone? The (surface) f0 contour could manifest an interaction among tone, intrinsic segmental effects, and citation intonation? While Tone 3 in Standard Mandarin is known for its shape of a falling-rise contour, all tokens illustrated in Fig. 16 also have an extra falling tail (a falling-rising-falling shape). To what extent do we hear such a tail in a citation tone? More perception studies are needed to see what role these controversial f0 segments play in our perception of a tone.

ACKNOWLEDGMENTS

This research was supported by the Leverhulme Trust Early Career Fellowship.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The catalogue of articles reviewed in this paper and the annotations is publicly available at OSF repository https://osf.io/7h3ar/.

APPENDIX: CHECKLIST FOR CITATION TONE RESEARCH

The following checklist is designed to guide researchers through the critical stages of conducting rigorous and reproducible research on the production of lexical tones. It will serve as a valuable roadmap for your production study, from the initial planning phase to the analysis and interpretation of results.

(1) Language Variety

Which language variety are you studying and why?

☐ Introduce the variety you are researching and its language family.

☐ Introduce the geographical location of the variety and the speech community.

☐ Report any previous findings in the literature about the variety.

☐ Report a summary of the tonal system of the variety in conclusion.

(2) Research Ethics

Does the research comply with ethical guidelines?

☐ Apply for ethical approval from your research institute before data collection.

☐ Assess the risks involved in field trips and data collection.

☐ Consider how to recruit and compensate participants.

☐ Prepare recruitment materials, consent forms, and other supporting documents.

☐ Check out relevant laws and regulations concerning data protection and privacy.

☐ Store all data in secure environments, using encryption and controlled access.

(3) Speaker Selection

Are the sample population representative of the target speech community?

☐ Consider the number of speakers.

☐ Consider the gender balance of the speakers.

☐ Consider the age range of the speakers.

☐ Administer a sociolinguistic questionnaire to learn about speakers' language use and language backgrounds.

(4) Speech Materials Design

Are the speech materials designed in a well-controlled and motivated way?

☐ Consider the number of monosyllabic words in the speech materials.

☐ Consider the number of repetitions of each word.
☐ Consider the syllable structure of the selected monosyllabic words.
☐ Consider the segmental composition of the selected monosyllabic words.
☐ Consider the usage frequency of the selected monosyllabic words.
☐ Consider the presence of tonotactic gaps in the chosen materials and discuss the strategies for addressing them.

(5) Experiment Setup
Have you considered how to present your speech materials?
☐ Decide on the presentation format of the speech materials.
☐ Consider the presentation sequence of the monosyllabic words.
☐ Add pauses between utterances.
☐ Use the same language variety as the medium of instructions.
☐ Standardise the instructions to ensure that each participant receives identical information.

(6) Recording
Where do you conduct the recording sessions and what equipment do you use?
☐ Conduct the speech recordings in a phonetics lab or a quiet small room with soft furnishings.
☐ Use a head-mounted unidirectional microphone.
☐ Locate the microphone head to the side of or below the mouth.
☐ Consider the choice of the recorder and recording software.
☐ Take a picture of the recording setup.
☐ Configure the recording setting including the sampling rate, bit depth, and channel.
☐ Save the recordings in WAV format.

(7) Data Management
Have you considered how to organise your corpus of monosyllables?
☐ Backup all data securely, preferably in multiple locations.
☐ Assign a unique ID to each participant to anonymise the data.
☐ Log the relevant metadata including the original filename of the recordings, recording date and time, and participant ID, after each recording session in a spreadsheet.
☐ Devise a file-naming system for the recordings and the relevant processing files such as the TextGrids.
☐ Keep track of the processing steps and do not overwrite the original data.

(8) Acoustic Measurement
How to do you extract f0 contours?
☐ Select an acoustic analysis interface and algorithm for f0 measurement.
☐ Check if any pre-processing is needed.
☐ Set appropriate parameters for the f0 estimation algorithm.
☐ Inspect the first-pass f0 measurements and identify potential outliers.
☐ Examine the f0 data points flagged as potential outliers and address them accordingly.
☐ Consider normalisation strategies for f0 and time.

(9) Pitch Contour Analysis
How do you analyse f0 contours and present findings of citation tones?
☐ Validate the number of lexical tones (distinct contour shapes).
☐ Describe the tone contours qualitatively.
☐ Provide some descriptive statistics of f0 contours and/or their duration by tone, such as the mean, standard deviation, maximum, minimum, and f0 range.
☐ Document any voice quality variations.
☐ Select an f0 contour analysis method that handles hierarchical data structure.
☐ Visualise the prototypical f0 contour of each tone based on f0 data from all speakers.
☐ Summarise the prototypical f0 contour of each tone in textual description.
☐ Report the specific rescaling strategy and decision boundaries if summarising the f0 contours in tone numerals.

(10) Open and Reproducible Research
How to adopt open and reproducible research practices?
☐ Share the study materials, f0 data, preprocessing logs, and analysis scripts via open platforms such as Open Science Framework (OSF) and Github.
☐ Where appropriate and consented to, openly share the monosyllabic corpus including audio recordings and anonymised metadata.

[1]To achieve a reliable, informative and sustainable database, cited authors are encouraged to submit corrections if we have inaccurately interpreted or represented any aspects of the methods and/or results presented in their published articles. Furthermore, we invite scholars who have published work on citation tones not logged in the present database to share their results with us for inclusion in the database.
[2]See an example of a brief questionnaire on speakers' linguistic backgrounds in the Appendix A.5 in Xu (2022).
[3]A sample recording protocol for prosody research, including equipment setup, can be found in the Supplementary Material 1 in Zhang *et al.* (2024).

Andruski, J. E., and Costello, J. (**2004**). "Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong," J. Int. Phon. Assoc. **34**(2), 125–140.
Babacan, O., Drugman, T., d'Alessandro, N., Henrich Bernardoni, N., and Dutoit, T. (**2013**). "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *ICASSP 2013—38th IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, pp. 1–5.
Bei, X., and Xiang, N. (**2016**). *Shiyan Yuyinxue de Jiben Yuanli yu Praat Ruanjian Caozuo* (*Introduction to Experimental Phonetics and Praat*) (Hu nan shi fan da xue chu ban she, Changsha).
Boersma, P. (**1993**). "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, p. 14.
Bowern, C. (**2015**). *Linguistic Fieldwork: A Practical Guide* (Palgrave Macmillan, Basingstoke, UK).

Bradley, C. B. (**1916**). "On plotting the inflections of the voice," Am. Archaeol. Ethnol. **12**(5), 195–218.
Butcher, A. (**2013**). "Research methods in phonetic fieldwork," in *The Bloomsbury Companion to Phonetics*, edited by M. J. Jones and R.-A. Knight (Bloomsbury Academic, London), pp. 57–78.
Chao, Y. R. (**1922**). "Zhongguo yanyu zidiaodi shiyan yanjiufa" ("Experimental approach to citation tones of Chinese languages"), Ke xue **7**(9), 27–36.
Chao, Y. R. (**1930**). "A system of tone letters," Maitre Phonet. **30**, 24–27.
Chao, Y. R. (**1968**). *A Grammar of Spoken Chinese* (University of California Press, Berkeley, CA).
Chen, W.-R., Whalen, D. H., and Tiede, M. K. (**2021**). "A dual mechanism for intrinsic f0," J. Phon. **87**, 101063.
Chen, Y. (**2015**). "Neutral tone," in *Encyclopedia of Chinese Language and Linguistics*, edited by R. Sybesma, W. Behr, Y. Gu, Z. Handel, C.-T. J. Huang, and J. Myers (Brill, Leiden, Netherlands).
Chen, Y., and Guo, L. (**2022**). "Zhushan Mandarin," J. Int. Phon. Assoc. **52**(2), 309–327.
Chiang, H. T. (**1967**). "Amoy-Chinese Tones," Phonetica **17**(2), 100–115.
Chuang, Y.-Y., Fon, J., Papakyritsis, I., and Baayen, H. (**2021**). "Analyzing phonetic data with generalized additive mixed models," in *Manual of Clinical Phonetics* (Routledge, Milton Park, UK).
Clark, H. H. (**1973**). "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research," J. Verbal Learn. Verbal Behav. **12**(4), 335–359.
Collins Online Dictionary (**2024**). "Citation form," https://www.collinsdictionary.com/dictionary/english/citation-form (Last viewed July 18, 2024).
de Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**(4), 1917–1930.
Duanmu, S. (**2007**). "Phonology of the world's languages," in *The Phonology of Standard Chinese*, 2nd ed. (Oxford University Press, Oxford, UK).
Entropic Research Laboratory (**2006**). "Entropic Signal Processing System (ESPS)."
Esposito, C. M. (**2010**). "Variation in contrastive phonation in Santa Ana Del Valle Zapotec," J. Int. Phon. Assoc. **40**(2), 181–198.
Frazier, M. (**2013**). "The phonetics of Yucatec Maya and the typology of laryngeal complexity," STUF–Lang. Typol. Universals **66**(1), 7–21.
Fung, R. S. Y., and Lee, C. K. C. (**2019**). "Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception," J. Acoust. Soc. Am. **146**(5), EL424–EL430.
Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (**2013**). "Voice quality and tone identification in White Hmong," J. Acoust. Soc. Am. **133**(2), 1078–1089.
Gedney, W. J. (**1972**). "A checklist for determining yones in tai dialects," in *Studies in Linguistics in Honor of George L. Trager*, edited by M. E. Smith (Mouton, The Hague, Netherlands), pp. 423–37.
Genette, J., Verhoeven, J., and Gillis, S. (**2023**). "Fundamental frequency normalization and statistical power: An assessment of 15 normalizing techniques," in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, pp. 644–648.
Grabe, E., Kochanski, G., and Coleman, J. (**2007**). "Connecting intonation labels to mathematical descriptions of fundamental frequency," Lang. Speech **50**(3), 281–310.
Gu, W., Hirose, K., and Fujisaki, H. (**2007**). "Analysis of tones in Cantonese speech based on the Command-Response Model," Phonetica **64**(1), 29–62.
Gubian, M., Torreira, F., and Boves, L. (**2015**). "Using functional data analysis for investigating multidimensional dynamic phonetic contrasts," J. Phon. **49**, 16–40.
Hadjipantelis, P. Z., Aston, J. A. D., and Evans, J. P. (**2012**). "Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models," J. Acoust. Soc. Am. **131**(6), 4651–4664.
Hirst, D., and Espesser, R. (**1993**). "Automatic modelling of fundamental frequency using a quadratic spline function," Trav. Inst. Phon. Aix **15**, 75–85.
Ho, D-a. (**2015**). "Chinese dialects," in *The Oxford Handbook of Chinese Linguistics*, Vol. 1 of Oxford Handbooks, edited by W. S. Y. Wang and C. Sun (Oxford University Press, Oxford).
Hoole, P., and Honda, K. (**2011**). "Automaticity vs feature–enhancement in the control of segmental F0," in *Where Do Phonological Features Come From?: Cognitive, Physical and Developmental Bases of Distinctive Speech Categories*, edited by G. N. Clements and R. Ridouane (John Benjamins Publishing Company, Amsterdam).
Hou, X., Zhao, L., and Chodroff, E. (**2023**). "Intermingling tone systems: The relationship of Nanning Mandarin to Nanning Cantonese and Standard Mandarin," in *The Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, pp. 1935–1939.
House, A. S., and Fairbanks, G. (**1953**). "The influence of consonant environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. **25**(1), 105–113.
Howie, J. M. (**1974**). "On the domain of tone in Mandarin," Phonetica **30**(3), 129–148.
Howie, J. M. (**1976**). *Acoustical Studies of Mandarin Vowels and Tones* (Cambridge University Press, Cambridge, UK).
Huang, S., Liu, J., Wu, X., Wu, L., Yan, Y., and Qin, Z. (**1998**). "1997 Mandarin broadcast news speech (HUB4-NE)," available at https://catalog.ldc.upenn.edu/LDC98S73 (Last viewed July 18, 2024).
Hyman, L. M., and Leben, W. R. (**2020**). "Tone systems," in *The Oxford Handbook of Language Prosody*, edited by C. Gussenhhoven and A. Chen (Oxford University Press, Oxford, UK).
Jones, D. (**1909**). *Intonation Curves: A Collection of phonetic texts, in Which Intonation is Marked throughout by Means of Curved Lines on a Musical Stave* (B.G. Teubner, Leipzig, Germany).
Jones, M. J., and Knight, R.-A. (**2013**). "Bloomsbury companions," in *The Bloomsbury Companion to Phonetics* (Bloomsbury Academic, London).
Kawahara, H., Estill, J., and Fujimura, O. (**2001**). "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *The Proceedings of the 2nd International Workshop of Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, Florence, Italy.
Keating, P., and Kuo, G. (**2010**). "Comparison of speaking fundamental frequency in English and Mandarin," in *UCLA Working Papers in Phonetics*, Vol. 108, pp. 164–187.
Kirby, A., and Sonderegger, M. (**2018**). "Model selection and phonological argumentation," in *Shaping Phonology*, edited by D. Brentari and J. L. Lee (University of Chicago Press, Chicago).
Kratochvíl, P. (**1968**). *The Chinese Language Today: Features of an Emerging Standard* (Hutchinson, London).
Kuang, J., and Liberman, M. (**2015**). "Influence of spectral cues on the perception of pitch height," in *Proceedings of ICPhS*, Glasgow, UK, pp. 0435.1–0435.5.
Ladd, D. R., Silverman, K. E. A., Tolkmitt, F., Bergmann, G., and Scherer, K. R. (**1985**). "Evidence for the independent function of intonation contour type, voice quality, and *F*0 range in signaling speaker affect," J. Acoust. Soc. Am. **78**(2), 435–444.
Ladefoged, P. (**1997**). "Instrumental techniques for linguistic phonetic fieldwork," in *The Handbook of Phonetic Sciences*, number 5 in Blackwell Handbooks in Linguistics, edited by W. J. Hardcastle and J. Laver (Basil Blackwell, Oxford), pp. 137–166.
Lakens, D. (**2022**). "Sample size justification," Collabra: Psychol. **8**(1), 33267.
Lee, W.-S., and Zee, E. (**2009**). "Hakka Chinese," J. Int. Phon. Assoc. **39**(1), 107–111.
Li, Q., Chen, Y., and Xiong, Z. (**2019**). "Tianjin Mandarin," J. Int. Phonetic Assoc. **49**(1), 109–128.
Li, X. (**2019**). "Gannan Lintanxian (Xincheng zhen) danzidiao shiyan yanjiu" ("An experimental study on citation tones in Gannan Lintanxian Xincheng town"), Zhaozhuang xueyuan xuebao **36**(6), 32–37.
Liang, X., and Tang, Q. (**2017**). "Guiping Madonghua danzidiao shiyan yanjiu" ("An experimental study on Guiping Madong dialect"), Ningxia daxue xuebao (renwen shehui kexue ban) **39**(3), 8–17.
Liberman, M., and Pierrehumbert, J. (**1984**). "Intonational invariance under changes in pitch range and length," in *Language Sound Structure*, edited by M. Aronoff and R. T. Oehrle (MIT Press, Cambridge, MA), pp. 157–233.
Maddieson, I. (**1990**). "The transcription of tone in the IPA," J. Int. Phon. Assoc. **20**(2), 28–32.
Maddieson, I. (**2001**). "Phonetic fieldwork," in *Linguistic Fieldwork*, edited by P. Newman and M. Ratliff, 1st ed. (Cambridge University Press), pp. 211–229.
Maddieson, I. (**2023**). "Tone is not predominant: Tone is not primordial," in *Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, pp. 1901–1905.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017**). "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Interspeech 2017*, pp. 498–502.
Menn, L., and Boyce, S. (**1982**). "Fundamental frequency and discourse structure," Lang. Speech **25**(4), 341–383.
Mirman, D. (**2014**). "Chapman & Hall/CRC the R series," in *Growth Curve Analysis and Visualization Using R* (CRC Press, Boca Raton, FL).
Nolan, F. (**2003**). "Intonational equivalence: An experimental evaluation of pitch scales," in *Proceedings of the 15th International Congress of Phonetic Sciences*., Barcelona, Spain, Vol. 771.
Norman, J. (**1979**). "Chronological strata in the Min dialects," Fang yan **4**, 268–274.
Oh, K., and Un, C. (**1984**). "A performance comparison of pitch extraction algorithms for noisy speech," in *ICASSP '84*, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 9, pp. 85–88.
Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (**2019**). "PsychoPy2: Experiments in behavior made easy," Behav. Res. **51**(1), 195–203.
Pike, K. L. (**1948**). *Tone Languages: A Technique for Determining the Number and Type of Pitch Contrasts in a Language, with Studies in Tonemic Substitution and Fusion* (University of Michigan Press, Ann Arbor, MI).
Qu, Z., and Yang, C. (**2019**). "Acoustic analysis of tone in Benna Hani: Tone sandhi and neutralisation in an atypical Tibeto-Burman language," in *The Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, pp. 1967–1971.
Ratliff, M. (**2015**). "Tonoexodus, tonogenesis, and tone change," in *The Oxford Handbook of Historical Phonology* (Oxford Academic, Oxford, UK).
Reichardt, B., Zapata, O. J. B., Saurty, K., and Fließwasser, E. (**2016**). "Comparison of different algorithms for Pitch Tracking," https://nats-www.informatik.uni-hamburg.de/pub/SLP16/WebHome/POSTER-pitch-tracking-poster.pdf (Last viewed October 8, 2024).
Roettger, T. B., and Gordon, M. (**2017**). "Methodological issues in the study of word stress correlates," Linguist. Vanguard: Multimodal Online J. **3**(1), 20170006.
Rose, P. (**1987**). "Considerations in the normalisation of the fundamental frequency of linguistic tone," Speech Commun. **6**(4), 343–352.
Rose, P. (**1993**). "A linguistic–phonetic acoustic analysis of Shanghai tones," Aust. J. Ling. **13**(2), 185–220.
Rose, P. (**2019**). "Tonatory patterns in Taizhou Wu tones," in *The Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, pp. 2099–2103.
Sanker, C., Babinski, S., Burns, R., Evans, M., Johns, J., Kim, J., Smith, S., Weber, N., and Bowern, C. (**2021**). "(Don't) try this at home! The effects of recording devices and software on phonetic analysis: Supplementary material," Language **97**(4), e360–e382.
Selting, M. (**2007**). "Lists as embedded structures and the prosody of list construction as an interactional resource," J. Pragmatics **39**(3), 483–526.
Shen, X. S. (**1992**). "Mandarin neutral tone revisited," Acta Ling. Hafniensia **24**(1), 131–152.
Shi, B., and Zhang, J. (**1986**). "Vowel intrinsic pitch in Standard Chinese," Lund Work. Papers Linguistics **29**, 169–190.
Shi, F. (**1990**). *Yuyinxue Tanwei* (*An Investigation of Phonetics*) (Beijing University Press, Beijing).
Sóskuthy, M. (**2017**). "Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction," https://arxiv.org/abs/1703.05339 (Last viewed September 18, 2024).
Sóskuthy, M. (**2021**). "Evaluating generalised additive mixed modelling strategies for dynamic speech analysis," J. Phonetics **84**, 101017.
Stanford, J. N. (**2008**). "A sociotonetic analysis of Sui dialect contact," Lang. Var. Change **20**(3), 409–450.
Strömbergsson, S. (**2016**). "A comparison between three commonly used methods for pitch extraction in speech," https://cdn.dal.ca/content/dam/dalhousie/pdf/sites/icpla/postersjune17/61724.PDF.lt_9ddeaa1253afb7974ae5ffe9e8637664.res/61724.PDF (Last viewed October 8, 2024).
Talkin, D. (**1995**). "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, edited by W. B. Kleijn and K. K. Paliwal (Elsevier, Amsterdam), pp. 495–518.
Tang, P., Yuen, I., Xu Rattanasone, N., Gao, L., and Demuth, K. (**2019**). "Acquisition of weak syllables in tonal languages: Acoustic evidence from neutral tone in Mandarin Chinese," J. Child Lang. **46**(1), 24–50.
Tang, Z., Wang, D., Xu, Y., Sun, J., Lei, X., Zhao, S., Wen, C., Tan, X., Xie, C., Zhou, S., Yan, R., Lv, C., Han, Y., Zou, W., and Li, X. (**2021**). "KeSpeech: An open source speech dataset of Mandarin and its eight sub-dialects," in *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, p. 12.
Tian, Z., and Hong, X. (**2023**). "Jiyu shiyan yuyinxue de Xinfeng guanhua danzidiao yu shuangzidiao yanjiu" ("An experimental phonetic study on citation tone and disyllabic tone in Xinfeng Mandarin"), Gannan shifan daxue xuebao **44**(4), 80–85.
Vasishth, S., and Nicenboim, B. (**2016**). "Statistical methods for linguistic research: Foundational ideas–Part I," Lang. Ling. Compass **10**(8), 349–369.
Wan, M., and Wang, F. (**2023**). "A study on tones in Jiuhe Bai produced by Naxi speakers," in *The Proceedings of the 20th International Congress of Phonetic Sciences*, Prague, Czech Republic, pp. 3340–3344.
Wang, L., Gussenhoven, C., and Liang, J. (**2020**). "How to pronounce a low tone: A lesson from Kaifeng Mandarin," J. Int. Phon. Assoc. **50**(2), 199–219.
Watson, C. I., and Harrington, J. (**1999**). "Acoustic evidence for dynamic formant trajectories in Australian English vowels," J. Acoust. Soc. Am. **106**(1), 458–468.
Watt, D. (**2013**). "Research methods in speech acoustics," in *The Bloomsbury Companion to Phonetics*, edited by M. J. Jones and R.-A. Knight, Bloomsbury companions (Bloomsbury Academic, London), pp. 79–97.
Whalen, D. H., and Levitt, A. G. (**1995**). "The universality of intrinsic F_0 of vowels," J. Phon. **23**, 349–366.
Winter, B., and Wieling, M. (**2016**). "How to analyze linguistic change using mixed models, Growth curve analysis and generalized additive modeling," J. Lang. Evol. **1**(1), 7–18.
Wood, S. N. (**2017**). *Generalized Additive Models: An Introduction with R* (Chapman and Hall, London).
Xu, C. (**2022**). "Investigating the tonal system of plastic mandarin: A cross-varietal comparison," Ph.D. thesis, University of Oxford, Oxford, UK.
Xu, C. (**2023**). "ASR from Scratch II: Training models of Hong Kong Cantonese with MFA implementation" https://chenzixu.netlify.app/resources/3asr/sr4/ (Last viewed September 18, 2024).
Xu, C. (**2024**). "Cross-dialectal perspectives on Mandarin neutral tone," J. Phon. **106**, 101341.
Xu, C., and Zhang, C. (**2024b**). "A cross-linguistic review of citation tone production studies: Methodology and recommendations," https://osf.io/7h3ar/ (Last viewed September 18, 2024).
Xu, C., and Zhang, C. (**2024a**). "CitationTone," Zotero Group, available at https://www.zotero.org/groups/5660568/2178 (Last viewed September 18, 2024).
Xu, Y. (**1997**). "Contextual tonal variations in Mandarin," J. Phon. **25**, 61–83.
Yip, M. (**2002**). *Tone* (Cambridge University Press, Cambridge, UK).
Yu, K. M., and Lam, H. W. (**2014**). "The role of creaky voice in Cantonese tonal perception," J. Acoust. Soc. Am. **136**(3), 1320–1333.
Yuan, J., and Liberman, M. (**2008**). "Speaker identification on the SCOTUS corpus," J. Acoust. Soc. Am. **123**(5), 3878.
Zee, E. (**1991**). "Chinese (Hong Kong Cantonese)," J. Int. Phon. Assoc. **21**(1), 46–48.
Zhang, C., and Chen, Y. (**2024**). "The interface of intonation and lexical tone: Boundary phenomena in Mandarin varieties," in *Shaping Phonological and Morphological Representations: Diachrony, Acquisition, and Processing*, edited by S. Kotzor, P. Fikkert, and A. Wetterlin (Oxford University Press, in press).
Zhang, C., Jepson, K., and Chuang, Y.-Y. (**2024**). "Investigating differences in lab-quality and remote recording methods with dynamic acoustic measures," Lab. Phonol. **15**(1), 1–49.
Zhang, C., Jepson, K., Lohfink, G., and Arvaniti, A. (**2020**). "Speech data collection at a distance: Comparing the reliability of acoustic cues across homemade recordings," J. Acoust. Soc. Am. **148**(4), 2717.
Zhang, C., Jepson, K., Lohfink, G., and Arvaniti, A. (**2021**). "Comparing acoustic analyses of speech data collected remotely," J. Acoust. Soc. Am. **149**(6), 3910–3916.
Zhang, J. (**2018**). "A comparison of tone normalization methods for language variation research," in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Association for Computational Linguistics, Hong Kong.

Zhang, Q. (**2022**). "Yantaihua danzidiao de gejv yu bianyi" ("Yantai Mandarin citation tone system and variation"), Tongji daxue xuebao (shehui kexue ban) **33**(4), 111–119.
Zhao, Y., and Jurafsky, D. (**2009**). "The effect of lexical frequency and Lombard reflex on tone hyperarticulation," J. Phon. **37**(2), 231–247.
Zhou, Y. V., and Martin, B. A. (**2012**). "The role of amplitude envelope in Cantonese lexical tone perception: Implications for cochlear implants," in *Speech Prosody 2012*, pp. 629–632.
Zhu, J., Zhang, C., and Jurgens, D. (**2022a**). "ByT5 model for massively multilingual grapheme-to-phoneme conversion," in *Interspeech 2022*, pp. 446–450.
Zhu, J., Zhang, C., and Jurgens, D. (**2022b**). "Phone-to-audio alignment without text: A semi-supervised approach," in *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8167–8171.
Zhu, X. (**1999**). *Shanghai Tonetics* (Lincom Europa, Munich, Germany).
Zhu, X. (**2004**). "Jipinguiyihua—ruhe chuli shengdiao de suiji chayi?" ("f0 normalization: How to deal with between-speak tonal variations?"), yuyan kexue **3**(2), 3–19.
Zhu, X. (**2010**). *Yuyin Xue* (*Phonetics*), 1st ed. (The Commercial Press, Beijing).
Zotcard (**2023**). "zotcard," https://github.com/018/zotcard (Last viewed September 18, 2024).