IDEAH •

A Tone Perfect Story: How to Develop an Open Access Mandarin Chinese Audio Database as a Collaborative Digital Humanities Project

Catherine Ryu, Benjamin Furman, Devin Higgins

Published on: Aug 19, 2021

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Tone Perfect, a Mandarin Chinese (MC) audio database developed by an interdisciplinary team at Michigan State University (MSU), was launched in August 2017 as a digital collection and open access dataset to support teaching and research that pertain to MC. This article provides an overview of how the research team, the audio production team, and the digital repository team collaborated to build the database (tone.lib.msu.edu/), which includes an exhaustive catalogue of monosyllabic sounds in MC (410 in total) in four tones (410 x 4 = 1,640) pronounced by six native MC speakers. More specifically, this article illuminates the multifaceted process of implementing a collaborative digital humanities project, while highlighting the challenges that the collaborative team addressed to accommodate the unique nature of this database of 9,840 audio assets as a web-based open-source project. To that end, this article is comprised of four parts: (I) key challenges to building *Tone Perfect*; (II) designing the database of audio recordings; (III) transforming the database into a learning and research resource; and (IV) broader implications of the *Tone Perfect* project. We hope that this story of *Tone Perfect*, which was an entirely novel undertaking for all team members involved, will also serve as a resource for those who plan to develop similar digital humanities projects in the future.

Key Challenges to Building *Tone Perfect*

For native English speakers, MC—one of the most studied foreign languages in the U.S.—is particularly challenging to learn because it is a tonal language. The tones in MC differ intrinsically from the pitch, volume, and inflection used in English to add texture to semantic meaning. In MC, which has four main tones (1-4), the same sound “ma,” for instance, can mean “mother (1),” hemp (2), “horse (3),” or “to yell (4),” depending on the tone used as Figure 1 indicates.

Figure 1: The Sound “ma” in Four Tones with Different Meanings

Researchers and teachers have been developing and experimenting with various tools and approaches to aid learners’ Mandarin tone acquisition. In fact, the audio assets of *Tone Perfect* were originally produced for *Picky Birds*, a monosyllabic MC tone

perception app game, which was developed by another interdisciplinary team at MSU. This initial motivation behind the project dictated not only that the assets be of high quality for tonal perception training purposes, but also that they be of human voices instead of synthetic stimuli. It meant that throughout the entire process of developing *Tone Perfect*, we needed to capture and deliver the irregularities and rich textures of the six voice actors' tonal productions in a consistent manner. Another challenge we faced was the logistics of undertaking the *Tone Perfect* project itself; namely, how to establish (and to share in advance) communication norms, standards, and paths among all interdisciplinary team members engaged in research, audio production, and repository & web development. As will be discussed, these challenges and others impacted the design of all aspects of the project.

Design of the Database of Recordings

This section highlights key design aspects of *Tone Perfect* as an audio database, spanning across the audio pre-production, production, and post-production stages, as well as highlighting the metadata structure and the website interface layout.

Tone Distribution Chart

From the overall perspective of this collaborative project's design and implementation, the single most important document is a tone distribution chart, in this case an Excel sheet that lists all the tokens (discrete audio assets) to be recorded. As will be shown, this chart served not only as the backbone of the pre-recording, recording, and post-recording production but also as the connective tissue for all team members involved in developing what has ultimately become the audio database that is *Tone Perfect*. The distribution chart was developed in house by a research team comprised of three native Chinese researchers who were subject specialists in linguistics, second language teaching in Chinese, and sociolinguistics, and the project lead (Catherine Ryu), a digital humanities specialist. Designing the tone distribution chart itself entailed three steps: (1) establishing token selection criteria; (2) identifying resources for selecting monosyllabic sounds; and (3) designing the layout of the distribution chart itself.

As for the token selection criteria, the team decided to include both lexical tones (a sound-tone combination with a semantic meaning) and lexical gaps (a sound-tone combination without a semantic meaning). In light of studies that report that native speakers' tone production of lexical gaps differs from their production of lexical tones, we recognized that an investigation into such differences could lead to a fuller understanding of native speakers' tone perception and production, which could in turn shed light on how MC

learners acquire tone proficiency of unfamiliar sounds. Moreover, since *Picky Birds* was for tone perception training only, the distinction between lexical tones and lexical gaps would not be as important, and it would even be beneficial to expose users to *all* potential sound-tone combinations of monosyllabic sounds in MC.

For selecting monosyllabic sounds for the distribution chart, we utilized two open sources, the [Jun Da Corpus](#) and the [Lancaster Corpus of Mandarin Chinese](#). The former provides a sequence of monosyllabic words in the order of the frequency associated with a particular character. For example, ‘de’ (的) is no. 1, being the most frequently used character in Chinese, and ‘long 2’ (鸪), being the least frequently used, is no. 9933. We selected a near mid-point (no. 5000) as a threshold for the sounds to be included in the distribution chart (Figure 2).

Figure 2: Tonal Distribution Chart in the Making

By excluding any item ranked over 5000 and by further eliminating the chosen items not included in the Lancaster Corpus, we identified 410 monosyllabic sounds, which formed a set of 1,640 tokens (410 monosyllabic sounds in four tones). The size of this selection is comparable to that of monosyllabic words (416) included in *Xinhua Dictionary*, one of the most authoritative and popular MC dictionaries. Of these 410 sounds, we identified one quarter of them as lexical gaps (397).¹

The next step was to determine how to design the tonal distribution chart itself. Given that the audio assets were to be embedded in the *Picky Birds* app based on the game's level design, we organized the 410 sounds in accordance with the hierarchy of perceptual difficulty in acquiring non-native sounds. Determining the degrees of difficulty is, however, a complex task, and we designed the tonal distribution chart based on a hypothesis known as the Perceptual Assimilation Model (PAM) proposed by such scholars

as Catherine Best and Michael Tyler (Tyler 5-7). More specifically, we distributed 1,640 sounds over 11 levels, which were determined by phonetic and phonological similarities between English and Mandarin. For example, such sounds as “ai” (which sounds similar to the word “eye” in English) were included in level 1, whereas such sounds as “xiong” and “zhuan” were assigned to level 11.² Difficulty of sounds in the same level was not differentiated. Each level was further divided into sections (e.g., A, B, C, D) of ten sounds apiece. This unit of ten sounds in four tones was to serve later as a recording module, and 410 sounds were to be recorded ten at a time. To assist the voice actors’ tone production, the finalized distribution chart also included a Chinese character for each sound. These characters were later included in the *Tone Perfect* website with multiple characters for homonyms. On the finalized distribution chart shared with the digital repository team, each Chinese character was also notated with Pinyin (a transcription system of Chinese in Roman letters), accompanied by a tone number (1-4), and lexical gaps indicated as shown in Figure 3.

The image shows a highly detailed and dense table, which is the 'Finalized Tonal Distribution Chart'. It is organized into 11 main sections, one for each level. Each section contains a grid of characters, likely representing different sounds or tones, with corresponding Pinyin and tone numbers. The table is very complex, with many columns and rows, and it appears to be a reference tool for recording audio.

Figure 3: Finalized Tonal Distribution Chart (the Traditional Chinese Character Version).

Developing Recording Procedures

Audio recording is a multi-tiered process, involving several elements that must be in place prior to the recording session and the post-recording audio production stages. This section highlights key elements, including fielding a team of voice actors and training them, and forming the audio production team and training its members.

The surest way to produce audio assets of the highest quality is to get the best and most consistent quality in the original recording. The voice actors in this regard are of paramount importance. Given the wide range of spoken Chinese, we decided to produce audio assets by monolingual MC speakers only from the Beijing area. Since this project was carried out at MSU, we identified six speakers (three male and three female), through

a rigorous selection process among international Chinese students from Beijing.³ After a preliminary tone production test, the research team interviewed potential voice actors to confirm their eligibility. Since no native Chinese speakers speak the language one monosyllabic word at a time, the research team trained the six selected to be able to perform as competent voice actors who can produce each sound in four tones distinctly, deliberately, and consistently. These voice actors were homogenous in terms of age, education pursued (both in China and the U.S.), and the extent of tonal training received. As will be seen, their profiles are also included in *Tone Perfect* as an integral part of this database precisely because the history of their educational and linguistic backgrounds can be a factor when analyzing the outcomes of any empirical experiment. Yet this is the kind of data that is neither available nor widely shared in current practices of the field. That is why the research team intentionally designed the team of voice actors so as to provide their profiles as potential data points for researchers.⁴

Just as important as the voice actors for the success of the audio database project is the competence of the audio production team members. Benjamin Fuhrman headed this team comprised of three undergraduate students at MSU. Besides possessing all the credentials and experiences indispensable for a veteran sound engineer, Fuhrman is a coder, musician, and electroacoustic music composer. This MC audio production project presented to him an opportunity to work with a tonal language, a new challenge that could further expand his expertise. Moreover, he brought to the project his own recording equipment, which otherwise would have had to have been rented for the duration of the project over 20 months. The recording equipment included a MacBook Pro; a Focusrite Liquid Saffire 56 Audio Interface; a Gator portable rack; a factory-matched pair of Rode Nt5 microphones, which are much more consistent for recording in stereo than non-factory matched ones; two microphone stands; two XLR cables; and Beyerdynamic DT770 Pro headphones.⁵ Audio files were recorded in a digital audio workstation, Reaper, from the six voice actors using condenser microphones in an acoustically treated room. All were recorded in ‘CD quality’ (44.1kHz/16-bit) to ensure a standard level of audio quality, and to allow for segmentation across multiple machines with different hardware profiles. All the audio files were produced as stereo WAV files to maintain the highest level of file quality without the potential loss of higher frequencies associated with lossy or compressed formats. The team chose stereo recordings as these files are more versatile in terms of use, be it for Foley art or sound design, while easily being consolidated and re-rendered into mono files if needed.

Quality Control

Producing 9,840 audio assets in high quality required strict measures for quality control, especially since we held multiple recording sessions over several months. To begin with, the quality of a recording studio is the first and foremost factor for an audio database project. Each recording space possesses a character of its own, and in that sense, it should be treated as a veritable member of the audio production team and be chosen cautiously. Due to our limited budget, we could not rent a soundproofed professional studio, and the microphones used for the recording were sensitive enough to pick up the sounds of the room itself, despite our best efforts. We had several measures in place to ensure consistency in recording. A typical recording session required a set of preparations, including a special arrangement with the infrastructure planning and facility (IPF) unit to turn off all vents in the studio and its adjacent rooms. The condition of the small studio with no air conditioning impacted the flow of the recording session as more frequent breaks were needed than otherwise. As a rule, a full set was recorded in one session since the voice actor's vocal quality could vary from day to day. That meant, though, that each recording session would last for eight hours or longer. We also kept the same configuration of people in each recording session: the voice actor, a voice coach (one of the three native Chinese research team members), the sound engineer, and the project lead.

A typical day of recording started with the recording engineer setting up the equipment while the voice actor did vocal warming exercises. Following the sound engineer's signal, we checked the microphone level, and he took some recording samples to check the audio quality and volume. He recorded the room with Reaper to create a subtractive finite impulse response (FIR) filter to be used later for removing any ambient noise from the room without impacting the recorded voice actor's tone productions. Then, the recording would finally start, and the voice actor would read off ten syllables in each of the four tones. During each recording sprint, the project lead and the voice coach checked tonal accuracy, vocal quality, the correctness of pronunciation and pitch contour. On our individual distribution charts, we notated any irregularities or errors during the recording, while the sound engineer monitored the register, volume, and audio quality on his monitor, and inserted markers directly on the recording for further reviews. At the end of the tenth syllable, we went around the table identifying any sounds to be re-recorded. Even when only one tone was off, all four tones were re-recorded as a set. Then, we moved on to the next group of ten, until all 1,640 sounds were recorded. Since intensely focused attention was required all throughout the recording session, we usually recorded about 20 to 30 syllables at a time, which took about 30 minutes of recording. Even if the

voice actor exhibited no sign of vocal fatigue, we would break for ten to fifteen minutes as other members also needed to rest their ears. This routine was maintained for all recording sessions to ensure consistent conditions for all audio asset production.

Audio Production and Post-Production

The workflow and best practices for the audio asset production and post-production were established long before the actual recording took place. However, it was in the process of working through the first set of audio assets that we were able to put guidelines into practice. To ensure consistency in the audio production and post-production, all three undergraduate production assistants were trained by the sound engineer on how to segment audio tokens uniformly, while recognizing how the onset of a tone varied, depending on the sound in question, as well as which tone it was. Using iZotope RX5 Suite, the assistants followed the audio post-production workflow. They were trained on how to de-click (i.e., remove all sounds such as dental clicks and saliva bubbles popping, breath noises, etc., as well as any external noises, including doors closing, footsteps, etc.). This required that the assistants be able to differentiate these noises from the sounds that were an integral part of the vocal production, such as aspirated sounds, or a vocal fry in the third tone, a dipping tone. During the audio asset processing phase, the assistants were required to wear Sennheiser HD 280 headphones to ensure a base level of audio quality that could not be guaranteed with a laptop's integrated speakers.

The assistant training also included labeling the audio assets. Each file was notated with the syllable followed by the voice actor's ID, for example **ma1_FV1.wav** (Ma syllable, tone 1, Female Voice 1). This notation with an underscore was chosen because some software (especially on the internet) does not play well with hyphens, and this also allowed the sound engineer to cluster each individual tone for analysis across multiple speakers more efficiently than other naming schemes. The production assistants would tag each audio clip with a name in Reaper, while splicing them up into regions because Reaper, which includes FFT based noise reduction, as well as built-in batch processing/exporting, can batch export regions. These region names would become the file names when exported. There would not be any additional tags created in the process of generating the audio files, and the files were uploaded to a shared cloud depository.

To make certain that all production assistants were following the same procedure throughout the audio production and post-production process, each audio asset was segmented in the order it appeared on the original tone distribution chart. The token distribution chart was modified as a production log to report any problematic tokens. To that end, new columns were added to include token, voice ID, level, section, tone number

(1-4), issues, and feedback. In other words, the file naming convention enabled us to identify and communicate about any one specific item among 9,840 audio assets. As the assistants were new to this kind of work, the sound engineer meticulously checked their work logs with notes about any problematic tokens and explained why. If the assistants could not fix the issue themselves, the sound engineer resolved it and added his comments to their individual logs, as shown in Figure 4.

	A	B	C	D	E	F	G
1	Tone	Voice	Level	Section	Number	Issue	
2	fu	MVS	3		2	end aspiration (edited)	FINE
3	yang	MVS	3		2	quality check (edited); removed large click	FINE
4	li	MVS	3		3	The beginning has a click and I ignored it at first, but Haitian marked it and I'm having second doubts, if we could trim it...	FIXED AS MUCH AS POSSIBLE
5	ha	MVS	2		2	AS I can't remove click (not edited)	FIXED
6	lu	MVS	2		4	has lots of v. fry at the end (edited)	YUP
7	nai	MVS	2		2	quality check (edited)	FINE
8	wa	MVS	2		4	end aspiration (not edited)	YUP
9	hai	MVS	2		3	quality check (edited)?	MADE SOME CHANGES
10	lai	MVS	2		4	AS I think we should trim after this point (not edited)	THAT GETS TAKEN CARE OF IN MASTERING
11	nai	MVS	2	384		AS I (done 3) and (around the same end part) tone 4 are natural clicks I assume	FIX, NOT CLICK
12	lu	MVS	2		2	I trimmed the end to remove a pop because there seemed to be a	FIXED
13	lu	MVS	2		3	list of room	FIXED
14	lu	MVS	2		3	Some as I cut, but for a different "excess" noise	FIXED
15	lu	MVS	2		4	the front has something weird going on and it is not a click or anything, but it's distorted or something	FIXED
16	lu	MVS	2		2	251 natural "click"	YUP
17	lu	MVS	2		4	quality check, please (edited) removed a large section and the end buttons me	EDITED AND CHANGED
18	ling	MVS	3		4	360 onwards has some aspiration, I don't know if you would want to lessen it or not	FINE
19	miu	MVS	3		4	AS I was wondering (to cut out the aspiration) if cutting after this point would be ok (edited)	THAT'S NOT ASPIRATION, NO, IT WOULD DESTROY THE TONE.
20	ping	MVS	3		3	300 There's a click and then a weird aspiration/air noise is there (not edited)	FIXED
21	ping	MVS	3		3	There's a high pitched noise around 270 (edited) probably ok to ignore, but just in case...	FINE
22	ping	MVS	3		4	quality check (edited)?	FIXED
23	min	MVS	3		2	AS I onwards is has "static" noise in the background, should be able to trim, but I defer it to you (not edited)	NORMAL BG NOISE ON A CONSONANT
24	ning	MVS	3		3	There's a high pitched aspiration after ASO	NOT SEEMING/HARING THAT
25	lin	MVS	3		3	From 370 onwards there's a metallic noise (edited)	RECORD - UNFIXABLE
26	gen	MVS	3		3	345 natural click	FINE
27	lin	MVS	3		3	375 onwards has a metallic noise and some aspiration (not edited)	RECORD - UNFIXABLE

Figure 4: A Sample of the Audio Production Log

Each time a whole set was prepared, the sound engineer and the project lead checked it in its entirety to identify any remaining problematic tokens. After our multiple reviews, the native Chinese research team members carried out, each independently, their auditory review of the accuracy and consistency of the set and compared their findings, using the original tone distribution chart as their review sheet. They were not provided with the production log in order to ensure that they had no advance awareness of problem tokens. Audio assets that did not pass the reviews at any stage were flagged for re-recording. The same process was repeated for re-recorded assets. We held six main sessions (one per voice actor) and approximately two to three re-recording sessions per voice actor. Overall, about 10% of the full six sets were re-recorded.

Since the ultimate purpose of creating the audio assets was to embed them in *Picky Birds* (Figure 5), we used a subset (1,200 tokens) of the first set (pre-mastered) for a modified *Picky Birds*, which was used as an experiment instrument for MC Tone Perception and Production (M-ToPP) led by Ryu.

Figure 5: Screen Shots of Picky Birds (a non-experiment version)

Prior to the experiment, we held a tone listening session in which all audio production members sat around the table and listened to each token embedded in the game, paying close attention to their uniformity in quality and accuracy. This process yielded a few more problematic assets and we re-recorded them.

For the rest of the five sets, the team became more efficient and the workflow more expedient. After all six sets were completed in this manner, the sound engineer mastered them together in terms of peak normalization. Otherwise, the audio assets were minimally treated (mainly removing noises) to retain the vocal quality and texture of natural human voices. There was no set time limit for the audio assets due to a wide range of variables determining the length of an audio asset, for example, a particular sound-tone combination and a voice actor's unique way of tone production. They ranged from a few hundred milliseconds to about 2 seconds for some of the third tones. While there was no conscious decision to make them a set length, silence from either end of an audio asset was stripped as part of the mastering process in order to ignore false positives when using the tokens for analysis or tracking of a non-native speakers' pitch contour (i.e., for research or experiments.) This resulted in some interesting mid-tone file separations in some third tone files, where the speaker would seemingly pause between the descending and ascending parts of the tone. While these were few and far between, it nevertheless required the sound engineer to do additional quality control checks on all third tone files during the mastering process. The mastered sets were also reviewed for consistency, quality, and accuracy. Once the reviews were completed, all the audio assets, logs, and review sheets were shared and archived in the project's cloud depository.

Overall, the audio production team did not experience any unique challenges due to the acoustic property of Chinese as a tonal language. A potential challenge due to the audio production team lead's and one production assistant's lack of familiarity with this language was proactively addressed by the inclusion of Pinyin in the tone distribution

chart, which facilitated the recording process and the audio file production. Following best practices, the sound engineer controlled the amount of bleed when recording the voice actors' tone productions. When mastering all the reviewed and approved audio assets, various steps were taken to ensure their consistency, including evenly spacing out each token; checking any phase cancellation between the left and right channels in all tokens; and normalizing each token's amplitude, just to mention a few. As the sound engineer commented, "Mandarin is surprisingly regular in the pronunciation of almost all tones (to the point that I can now read the tone number by looking at the waveform)" (Fuhrman 4).⁶ The most crucial aspect of managing audio productions of this nature and scope, however, is to automate and expedite the production process as much as possible by using scriptable tools such as Reaper, which allows an identical setting for amplitude control and facilitates the processes for tagging the tokens.

Metadata and Manner of Storing Obtained Information

The MSU Libraries' digital repository team, then headed by Devin Higgins, joined the project upon the audio production team's completion of the first two sets of audio assets (pre-mastered). With these sets, the repository team started to design a prototype metadata structure for a mini-database. The original idea had been to build a relational database that would house information about the audio files' content: the voice actor ID, the sound, and tone. The search would be bidirectional from filename to information or information to filename. It would include the voice actor's profile (hometown, age, gender, education, and so on). From the onset, we envisioned the structure of this database to be flexible so that more data could be added in the future, including an initial vowel, a vowel cluster, an initial consonant, and a consonant cluster; segment and tone information of the sounds (i.e., initial consonants, rimes, and tones); and the boundary labeling for each sound (i.e., the boundary of the beginning/end of the consonant/vowel), which would enable researchers to extract acoustic information such as pitch values, duration, intensity, etc.) for their own purposes. This section focuses on key technical aspects of the current metadata structure of *Tone Perfect*, which provides users with multiple channels of interaction and query.

Embarking on a project like *Tone Perfect* presented some challenges for the digital repository team. A collection of monosyllabic spoken word recordings of only a few seconds' duration differs in significant ways from other collections of audio materials held in the digital repository, which were typically longer speeches, oral histories, and interviews more suited to a sustained engagement with each piece. The *Tone Perfect* collection, by contrast, is better served by an interface that allows users to play

multiple files quickly in succession and make comparisons between related audio files across a range of different factors. Comparing the same syllable as pronounced by different speakers, or the same syllable sound across all four tones, were crucial functions for a database like *Tone Perfect*. The digital repository team works with a stack of software anchored by Fedora Commons, a widely used repository option, especially in larger academic libraries, that tends to enforce the notion that each digital object is a discrete entity, likely to be examined independently. To make *Tone Perfect* work for language learners, the team adapted the front-end interface to Fedora Commons, Islandora, to allow users to play audio files from any search result page, as shown in Figure 6. This way users can readily make connections between sounds without the need to load a new page.

Figure 6a: A Sample of the Sound “ai” Random Search Results

Figure 6b: A Sample of the Sound “ai” in Tone 1
Research Results Displayed in the Order of
Relevance

On a similar note, libraries are most familiar with metadata standards suitable for individual bibliographic materials, that is, books, or other media. These standards can be either generalist and designed to accommodate a wide range of materials, such as the commonly used Dublin Core, or more specifically designed for a particular purpose, such as to describe works of art (Visual Resources Association Core) or ecological data (Ecological Metadata Language) or any number of other specialist use cases. No existing metadata standard, however, was quite right for describing *Tone Perfect* audio recordings with the granularity and precision required to make it most useful. The repository team devised a simple XML structure to contain exactly the desired information and wrote a script in Python to convert the source spreadsheet of tonal information into metadata suitable for inclusion in the repository, as shown in Figure 7.

```

<custom>
1  <tone>l</tone>
2  <syllable>tingl</syllable>
3  <pinyin>tíng</pinyin>
4  <character_forms>听 聽</character_forms>
5  <character_forms>厅 廳</character_forms>
6  <character>
7    <simplified>听</simplified>
8    <traditional>聽</traditional>
9  </character>
10 <character>
11   <simplified>厅</simplified>
12   <traditional>廳</traditional>
13 </character>
14 <level>Unknown</level>
15 <frequent_character_count>4+</frequent_character_count>
16 <sound>ting</sound>
17 <title>ting by MV1</title>
18 <gender>male</gender>
19 <section>Unknown</section>
20 <initial>t</initial>
21 <lexical_gap>No</lexical_gap>
22 <speaker>MV1</speaker>
23 <identifier>tingl_MV1</identifier>
24 <final>ing</final>
25 </custom>

```

Figure 7: A Sample of the Custom Metadata Scheme Used by the Repository Team to Describe Each Audio Recording

The waveforms that serve as visualizations of each sound are produced using Wavesurfer.js, a JavaScript library for rendering and playing audio files within web browsers. The visual representation of each image is based on the mono signal that results from combining the two stereo channels of each audio file. The single waveform makes for a cleaner design element for each syllable but does lose some information that could otherwise be visible to users if both channels of the original audio recordings were rendered separately.

The *Tone Perfect* site plays back MP3 audio files, which were created using FFmpeg, utilizing its highest constant-bitrate compression setting, 320 Kbit/second. Given the relatively narrow dynamic range of voice recordings, as compared to most music, any loss in quality due to this compression is very unlikely to be appreciable to listeners. The compression produces a file about one quarter the size of the original .wav audio, which allows for faster page load. Given the brevity of *Tone Perfect* recordings, the savings there may be rather modest. But the smaller file sizes do render the dataset more portable than otherwise. The full downloadable dataset, available as a zipped package, contains the 9,840 MP3 files in roughly 300 MB. The zipped XML metadata is about 14 MB. Due to licensing restrictions on the original audio files, this data is available only for non-commercial projects, an intention that users must confirm prior to gaining access to the files. Upon approval of a user request, the files are delivered via FileDepot, an MSU data-

sharing service for large files. Figure 8 highlights the process and software used to ingest the audio files into the digital repository.

Figure 8a: The Flowchart of the *Tone Perfect* Metadata Structure (overview)

Figure 8b: The Flowchart of the *Tone Perfect* Metadata Structure (enlarged)

Figure 8c: The Flowchart of the *Tone Perfect* Metadata Structure (enlarged)

Transforming the Database into a Learning Resource

The impetus for the *Tone Perfect* development was to expand and share the materials and knowledge generated by the research team and the audio production team. When the project lead first approached the digital repository team, what she had in mind was a repository for 9,840 audio assets after their original use for *Picky Birds* was completed. Precisely because producing these assets was labor intensive and costly, she wanted other researchers, students, and game developers to have access to audio assets for their projects without having to duplicate work. The repository team shared this vision, and the outcome of our collaboration is an interactive web-based audio database, which is now

one of MSU Libraries’ online collections. This section illuminates three key aspects of further developing *Tone Perfect* specifically as a learning and research resource: the website design, accessibility, and possible uses.

Designing the Website

The multimodal interactive website of *Tone Perfect*, supported by the metadata structure previously discussed, was brought about by a sustained conversation among all those involved throughout its development. The project lead attended the digital repository team’s monthly meetings and reviewed what had been completed and their plans for the next sprint. The regularly held meetings enabled both teams to respond to questions, exchange information, and resolve any issues in a timely fashion, if not in real time. For example, in the early phase of the website design process, we discussed various ways of labeling the audio assets using such methods as Pinyin (Romanized Chinese pronunciation) or the International Phonetic Alphabet (IPA), as well as the merits and demerits of each. We ultimately chose Pinyin as that is how novice Chinese learners are introduced to Chinese pronunciation in the U.S., similar to how young students are taught in China. Throughout this agile approach, the research team members served as a user experience focus group, testing the navigation of the website and the browser functionality from their multiple perspectives as instructors and researchers who understand the challenges faced by students when acquiring tonal proficiency in Mandarin Chinese.

By the time of its initial launch in 2017, the *Tone Perfect* website was comprised of four pages as indicated by the four tabs in the navigation bar: “home”, “browse”, “related projects”, and “using and citing data” (Figure 9). The configuration of these pages remains the same, though expanded in terms of their contents.

Figure 9: Screen Shot of the Navigation Bar

The home page provides key information about Mandarin Chinese, its unique tonal structure, and the conventional notation of tonal information, either with numbers or tone contour marks (ma1 / mā). Moreover, colour-coded wave forms are included to represent the four tones visually as Figure 10 demonstrates.

Figure 10: The Colour-Coded Wave Forms of the Sound “ma 1” in Four Tones by Multiple Voices

The colour-tone association was based on the results of a large-scale empirical experiment, which was designed and implemented by ToPES (Tone Perception Efficacy Studies), another MSU research team led by the project lead. The main objective of the experiment was to assess the comparative efficacy of conventional tonal visualization methods and a novel method of using the colour-tone association in relation to the existing scholarship on synesthesia, a comingling of perceptions (Godfroid et al. 821-822, 825-826). Therefore, the colour-coded wave forms used in the database are not for decorative purposes but one of multimodal perception training input. By bringing together the Chinese characters (both traditional and simplified), Pinyin, and three ways of visualizing tonal information, *Tone Perfect* thus provides the user with multimodal representations and input of the tonal information, visually and aurally. The remaining three webpages of the *Tone Perfect* site—browse, related projects, and using and citing the data—will be discussed later in the section on possible uses.

Access and Accessibility

Building *Tone Perfect* as an open-source project required addressing limitations to the use of the data, both unique to this collection and more general in nature. As mentioned earlier, the audio assets in this database were originally produced for *Picky Birds*, which was to be a paid app game developed at MSU and managed by the licensing office of MSU known as MSUT (MSU Technologies). This necessitated MSUT’s approval before making the same audio assets available to the public, especially prior to the release of *Picky Birds*. Through a series of conversations, MSUT and the team agreed upon “Educational Use Permitted” (Rights Statements, “In Copyright - Educational Use Permitted”), with

an understanding that this might later be changed to a non-commercial use statement. This condition also influenced how the paths of interactions with potential users were designed. While users can download individual audio assets directly from the website, they need to contact the repository team to access the entire set. It is only provided to users who intend to use it for non-commercial or education projects.

Another important consideration was the accessibility of the database. After the initial launch of *Tone Perfect* in 2017, the website was reviewed by an accessibility consultant using NVDA screen reader (deque.com/blog/accessibility-testing-nvda-screenreader/). Based on the consultant's recommendations, the repository team enhanced the website for visually impaired users. Some of the issues addressed included the structural layout of each page with a clear hierarchy of information, modification of the clickable buttons throughout the website to make them visible to the screen reader, modification of language input (English and Chinese) in the code to make it visible to the screen reader, and making buttons interactivity visible to the screen reader to indicate when an audio asset stops playing in sync with its wave form. For example, the repository team added "aria-label" attributes to key interface elements, as shown in Figure 11.

```
1.<button aria-label="Play" class="btn btn-primary waveform-play"
  onclick="wavesurfer.playPause()"><i class="fa fa-play" style="font-
  size:16px color:white"></i></button>
2.<input aria-label="Search the Repository" type="text" id="edit-
  islandora-simple-search-query" name="islandora_simple_search_query"
  value="" size="15" maxlength="256" class="form-text">
```

Figure 11: Code Samples

This brought the website in line with the World Wide Web Consortiums WAI-ARIA (Web Accessibility Initiative –Accessible Rich Internet Applications) standards.

There still remains a known accessibility issue, that of color contrast. Since the colour yellow used for the wave representation of tone 1 does not have enough contrast against the white background, a gray background might be preferable. To date, we have not received any concerns from users regarding colour-related visibility, but this issue will be addressed in the future.

Possible Uses

Tone Perfect is thus designed as a teaching and learning resource with multiple channels of interaction and query. The browser tab leads to the landing page where users can search for specific tones through various search options: by speaker (speaker 1-6); gender of speaker (male or female); tone (four tones); sound (410 sounds); or lexical gap (yes or no). The user can, for example, choose a particular tone (ai 1) by all six speakers and

compare the shape of the wave. While these wave forms are not identical as previously shown in Figure 6-b, there is a discernable pattern among them. This demonstrates that even native speakers do not produce a tone in the same way and that there is an acceptable range of tonal intelligibility. Similar differences can be also observed between male and female speakers. The pedagogical benefits of seeing and hearing that range cannot be overemphasized.

The related projects tab lists all the projects that have utilized *Tone Perfect*'s audio assets. At the initial launch of the database, all the projects featured here were MSU's internal projects, including a tone-audio piece, *Lingua Incognita*, and a piece of experimental music, *Heart Doubt*. Since its launch, we have received a steady stream of requests for its audio assets. In response, we automated the system in July 2020, using a Qualtrics form. The requests come from researchers, students, and professionals from all over the world, including the U.S., China, Netherlands, and the Czech Republic. The list of completed projects featured on the website now includes *Kuai* (a pronunciation app), *Mandarin Sound Table*, *Mandarin Tone Machine Learning*, *Chinese Consonant Classifier Watch Your Tone*, and *Supervised Learning Models for Classifying Tones in Mandarin Chinese*. Several additional projects are currently in progress and will be featured on this page upon completion.

Finally, the using and citing data tab opens to a page where users can contact the *Tone Perfect* team to request a batch download by filling out a form available there (Figure 12).

MICHIGAN STATE UNIVERSITY

Welcome to Tone Perfect!

Thank you for visiting the Tone Perfect database.

If you are interested in using the entire set of Tone Perfect's audio assets for your own project, and **if you do NOT plan to commercialize the outcomes of your project**, please fill out this survey. After reviewing your answers, the Tone Perfect Team will contact you. The survey also includes a consent form that will allow us to feature the outcomes of your project on the Tone Perfect website for our community of users. If you have not already seen the projects shared, please take a look at them (<https://toneperfect.lib.msu.edu/related>) before filling out the survey.

If you have any questions or concerns about this survey, please contact Catherine Ryu (jryuc@msu.edu).

Thank you for your time and interest in Tone Perfect!
We look forward to hearing from you and learning about your project!

Tone Perfect Team
Michigan State University

Figure 12: Screen Shot of the Request Form

As mentioned before, those users who request a batch download are potential contributors and members of the growing *Tone Perfect* community built on our shared interests in MC. This page also included information about how to cite the data to encourage users to follow best practices when making use of open access data.

Broader Implications of Tone Perfect

Since the initial launch of *Tone Perfect* in August 2017, we have been continuously improving its performance and accessibility. The labor-intensive nature of producing nearly 10,000 audio assets of monosyllabic MC sounds in human voices aside, the greatest challenge to the project team was the novelty of undertaking the project itself. Without having the benefit of learning the steps from other projects, we had to approach each step and stage of the project with openness and flexibility, be it identifying monolingual MC speakers at MSU and training them to be voice actors; looking for a suitable recording space; carrying out an eight-hour recording session; maintaining the rigorous and reiterative review process; and securing funds to support the project, just to mention a few.

Another challenge of *Tone Perfect* specifically as a digital humanities project was getting various units on campus on board to support it. For example, to collaborate with the digital repository team, we first had to secure approval from the associate dean for Digital Information & Systems, MSU Libraries. Without this approval, the project could not be carried out as it would require their division's human resources and time. Fortunately, our database project aligned optimally with the new MSU libraries vision to house not only traditional scholarly materials but also raw data (i.e., a set of segmented human voices as data, which differs in nature from a collection of famous speeches), and our collaboration was approved. Our project was also positively received as a library-faculty collaboration, which was actively encouraged as a way of fulfilling MSU libraries' mission. Our advocate digital curation librarian soon started to organize a series of meetings with all units and individuals who could potentially be brought into the project—the Asian subject librarian, the linguistics subject librarian, and other members in the digital information & system division, as well as audio specialists in the Vincent Voice library at MSU. In this way, our collaborative audio database project came to be known to those who subsequently became directly or tangentially involved with the project. This kind of institutional support, including MSUT's approval as mentioned before and MSU's internal funding, was critical to successful completion and implementation of *Tone Perfect*.

We were therefore delighted and grateful when *Tone Perfect* was selected for the 2018 Esperanto “Access to Language Education” award (esperantic.org/en/education/awards-calico/) and the 2019 Open Scholarship award (etcl.uvic.ca/2019/01/16/announcement-open-scholarship-awards-2019/). We are even more pleased to see *Tone Perfect* in action. In fact, from the beginning of this database project, the interactions with users have been some of our most important considerations. The repository team wanted to place minimal barriers to accessing the audio assets. That is why we provided the option of downloading individual audio files directly from the website. Of late, this database appears to be favoured by users in the field of computer science (undergraduates, graduates, and researchers from a wide range of academic institutions, as well as independent researchers and developers) who are interested in building and testing machine learning models. Assessing tone perception and production through machine learning has not yet been established in second language studies and teaching. Digital humanities projects such as *Tone Perfect* therefore have the potential to pave a new path of inquiry and knowledge production.

Tone Perfect, still developing and transforming, is a concrete expression of our team’s shared experiences, accumulated knowledge, technical expertise, the network of enabling support (both financial and human resources), and the contributions of our users. Since its initial launch, we have received inquiries and requests to expand the scope of the database to include multisyllabic sounds in MC, various dialects of Chinese, and other tonal languages, as well as even a collection of MC learners’ own tone productions that are not *perfect*. Due to the intensive labor and the prohibitive cost required to build such an audio database, our team does not have a plan to expand *Tone Perfect* for the foreseeable future, but if researchers from various institutions could come together to form a grant-supported consortium and build a large-scale multilingual audio database that can function as a hub of standardized audio asset collections, it would indeed transform dramatically the existing linkages between the digital humanities and other disciplines such as second language studies and game studies, while building a powerful dataset to support research in machine learning and AI-enabled voice recognition. We now conclude the story of *Tone Perfect* with a hopeful look toward the arrival of just such an exciting future.

Works Cited

Fuhrman, Benjamin. “Mandarin Tone Perception and Production (M-TOPP) Analysis Tool Report” . *Collection of Catherine Ryu’s Cube2Cube Documents*, unpublished, 7 August

2016, East Lansing, benfuhrman.com/code/ToPP_Analysis_Tool.zip.

Godfroid, Aline, et al. “Hearing and Seeing Tone through Colour: An Efficacy Study of Web-based, Multimodal Chinese Tone Perception Training.” *Language Learning*, vol. 67, no. 4, 2017, pp. 819-857. *ProQuest*, doi.org/10.1111/lang.12246.

Gong, Donald Shuxiao. “Grammaticality and Lexical Statistics in Chinese Unnatural Phonotactics.” *University College London Working Papers in Linguistics*, vol. 29, 2017, pp. 1-23, ucl.ac.uk/pals/research/linguistics/linguistics-research/working-papers-linguistics/contents-uclwpl-29-2017.

Li, Xiaoshi, et al. “Empirical Studies on L2 Mandarin Chinese Production: What Can We Learn from Them?” *Researching and Teaching Chinese as a Foreign Language*, vol. 3, no. 1 2017, pp. 23-49, journal.equinoxpub.com/RTCFL/issue/view/1839/.

Rights Statements. “In Copyright - Educational Use Permitted.” rightsstatements.org/vocab/InC-EDU/1.0/.

Ryu, Catherine, et al. *Tone Perfect: Multimodal Database for Mandarin Chinese* at Michigan State University, 2017, tone.lib.msu.edu/.

Tyler, Michael D., et al. “Perceptual Assimilation and Discrimination of Non-native Vowel Contrasts.” *Phonetica*. vol. 71, no. 1, 2014, pp. 4-21, doi.org/10.1159/000356237.

Acknowledgments

The *Tone Perfect* project was supported by the Delia Koo Endowment; the Humanities and Arts Research Program Production Grant; and the Target Support Grant for Technology Development from Michigan State University, awarded to Catherine Ryu from 2015 to 2017.

The authors of the article thank the two anonymous reviewers for their thoughtful comments and suggestions. We would also like to recognize the invaluable contributions of the research team members (Xiaoshi Li, Qian Luo, Jie Liu); several repository team members and library supporters (Shawn Nicholson, Aaron Collie, Robin Dean, and Sruthin Gaddam); the audio production team members (Ashley Davis, Zijin Liu, and

Haitian Yan); and the six voice actors whose profiles are available on the *Tone Perfect* website, as well as the support staff in various units at MSU.

Footnotes

1. This relatively large proportion of lexical gaps needs to be understood in the context of Donald Shuxiao Gong' s 2017 study. Gong cites San Duanmu' s 2011 findings, which indicate that in Standard Chinese (SC), of 1,900 theoretically possible syllables, only 404 have been identified, which means 79% of these positions have not been (Gong 1). [↵](#)
2. The same research team conducted an empirical experiment in summer 2017 with a modified version of *Picky Birds* to assess the actual hierarchy of difficulty by analyzing the data generated from the experiment that utilized the theoretically constructed hierarchy of difficulty. The team is currently working on the final stage of data analysis and manuscript preparation. [↵](#)
3. Potential voice actors were initially asked to create audio files of the following Chinese characters to test the authenticity and the purity of their Beijing accent: 海, 该, 退, 微, 呆, 温, 湾, 日, 很, 乌, 堆, 爱, 黑, 翁, 凶, 穷 (Hǎi, gāi, tuì, wēi, dāi, wēn, wān, rì, hěn, wū, duī, ài, hēi, wēng, xiōng, qióng). [↵](#)
4. Li et al. (30-34) discuss several factors influencing Chinese learners' tone production performance such as stimuli, task type, and stimuli presentation, and assessment methods. The same factors also influence Chinese learners' tone perception performance. [↵](#)
5. The Nt5' s are small diaphragm condenser microphones, while voices are normally recorded on large diaphragm condensers. The choice was due to two factors: SDC' s have better transient responses (important when recording small, isolated samples), and LDC' s characteristically have both more self-generated noise and a tendency to “colour” the sound recorded through reflections in the mic grille creating a reflective surface around the capsule. This colouration is not necessarily an issue, but the sound engineer chose Nt5 microphones to keep the recording as sterile and clinical as possible. He also chose not to use a dynamic microphone largely because he could not anticipate how much frequency data would be required for any software created to track language learners. Dynamic microphones typically lose frequency response over 10 kHz, while condensers do not. That is why the condenser microphones were used. [↵](#)
6. While overseeing the audio production as the sound engineer, Fuhrman also worked on another related project, *A Tone Analysis Tool*, and submitted to the project lead (Ryu)

his unpublished report upon completion of the project in August 2016. [↗](#)