Phonetic Reconstruction of the Consonant System of Middle Chinese via Mixed Integer Optimization

Xiaoxi Luo
Yuanpei College,
Peking University, China
lxx_1900017744@pku.edu.cn

Weiwei Sun
Department of Computer
Science and Technology,
Cambridge University, UK
ws390@cam.ac.uk

Abstract

This paper is concerned with phonetic reconstruction of the consonant system of Middle Chinese. We propose to cast the problem as a Mixed Integer Programming problem, which is able to automatically explore homophonic information from ancient rhyme dictionaries and phonetic information from modern Chinese dialects, the descendants of Middle Chinese. Numerical evaluation on a wide range of synthetic and real data demonstrates the effectiveness and robustness of the new method. We apply the method to information from Guǎngyùn and 20 modern Chinese dialects to obtain a new phonetic reconstruction result. A linguistically motivated discussion of this result is also provided.[1]

1 Introduction

Phonological reconstruction is one main concern in historical linguistics. There are two fundamental goals: reconstructing phonological categories and reconstructing phonetic values of these categories or individual phonemes. The classic linguistic and philological approach applies a comparative strategy to solve both problems by connecting cognates in different languages. Previous research in computational linguistics demonstrates the possibility to automate the comparative approach to some extent. See, e.g., Bouchard-Côté et al. (2007a, 2009, 2013), List et al. (2022), and He et al. (2023), among others.

The comparative approach developed out of attempts to reconstruct Proto-Indo-European, the common ancestor of the Indo-European language family. However, the comparative method itself is not well equipped to handle the special challenges of reconstructing Chinese. On the one hand, a key step in the comparative method is to identify as many cognates as possible, which is relatively straightforward for Chinese but can be extremely challenging in other languages. On the other hand, documentary materials predominantly use Chinese characters to annotate other characters (such as Fǎnqiē 反切), a tradition that has continued for thousands of years. For example, rhyme dictionaries such as Qiēyùn 切韻 extensively use this unique annotation method and systematically represent the phonological system of Chinese during a specific period. Such precious materials are relatively rare in other languages. This information is invaluable for the reconstruction of proto-languages, but the comparative method itself cannot adequately handle it. In fact, throughout Chinese history, numerous works similar to the Qiēyùn have existed in different periods, each reflecting the phonological system of its time. The purpose of this paper is to address the question of how to systematically utilize these phonetic materials.

In the practice of phonological reconstruction for ancient Chinese, linguists have been overwhelmingly exploring alternative information to spelling and hence alternative methods to the comparative one (Handel, 2014, pp. 1–2). Their work heavily relies on philological documents, especially rhyme dictionaries, which have a unique way to record homophonic information, i.e., Fǎnqiē. A basic consensus on the phonological categories of Middle Chinese (MC)[2] has been reached—there were 35–38 initials in MC, with minor disagreement only on some categories' merging or splitting (Karlgren, 1926; Li, 1956; Wang, 1957). Ancient rhyme dictionaries, however, do not provide phonetic information, and phonetic reconstruction is still extremely challenging. The relevant research is limited and there

[1]Code and datasets are available at https://github.com/LuoXiaoxi-cxq/Reconstruction-of-Middle-Chinese-via-Mixed-Integer-Optimization.

[2]There are three basic periods: Old Chinese, Middle Chinese, and Old Mandarin (Wang, 1957).

Transactions of the Association for Computational Linguistics, vol. 13, pp. 424–441, 2025. https://doi.org/10.1162/tacl_a_00742
Action Editor: Micha Elsner. Submission batch: 10/2023; Revision batch: 10/2024; Published 4/2025.

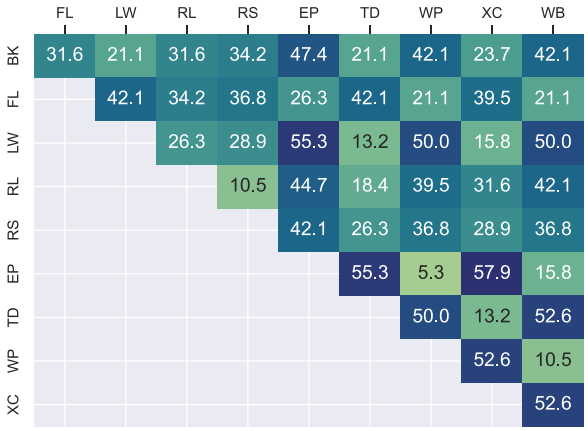


Figure 1: Disagreement among scholars on phonetic reconstruction. BK: Karlgren (1926), FL: Li (1971), LW: Wang (1957), RL: Li (1956), RS: Shao (1982), EP: Pulleyblank (1984), TD: Dong (2004), WP: Pan (2000), XC: Chen (2005), WB: Baxter (1992).[3] The number in each cell represents the percentage of initials on which the corresponding two scholars disagree.

is a lot of disagreement among scholars. Figure 1 shows the percentage of initials with which scholars disagree. There is significant inconsistency between any two scholars, l et alone a consensus among all of them.

This paper is concerned with developing a computational model for phonetic reconstruction of the consonant system of MC. We propose to cast phonetic reconstruction as a Mixed Integer Optimization problem (§4). A particular goal is to conveniently integrate heterogeneous information. We consider two major information sources: (1) philological documents and (2) modern Chinese dialects,[4] the descendants of MC. Following Generative Phonology (Chomsky and Halle, 1968), we introduce a novel compact set of Chinese-specific distinctive features to represent consonants (§3). Based on the feature-oriented precise phonetic representation, we formalize the optimization goal as minimizing the overall distance between possible homophonic characters and the overall distance between MC and modern dialects. Measuring the distance is a key element to the success of the new architecture. To this end, we design a new mathematically sound distance/metric function to suit our feature representation.

Evaluating the goodness of the reconstruction result is uniquely challenging because of the lack of ground-truth. Instead, we evaluate the reconstruction method. We consider two types of experiments: experiments on synthetic data (§5), where the ground-truth is known, and experiments with held-out data (§6), where partial information transformed from the ground-truth is known. To create representative synthetic data, we start from a pre-defined consonant system, derive homophonic information that matches Fǎnqiē, and derive varieties by introducing stochastic change as well as random noise. We consider three types of consonant systems: 1) purely artificial systems that randomly select elements, e.g., from an IPA chart, 2) natural systems of modern languages, including English, German, and Mandarin, and 3) the reconstructed system of Latin. Numerical evaluation demonstrates the effectiveness and robustness of the new method. It is able to successfully reconstruct most consonants when natural and reconstructed consonant systems are considered.

For the experiments with real data, we consider a wide range of representative Chinese characters with relevant information from Guǎngyùn and 20 modern dialects. Given the absence of ground-truth in phonetic reconstruction, to validate the effectiveness of the reconstruction method, we employ the strategy to hold out some Fǎnqiē information. In particular, we apply our method to 70% Fǎnqiē annotations and compare the automatically reconstructed result with the other 30%. The reconstructed phonemes predict around 68% Fǎnqiē. Considering that Fǎnqiē annotations themselves are not fully consistent, the result is quite promising and the method has a potential use to detect inconsistent Fǎnqiē annotations.

Based on the entire real data set, we provide a new phonetic reconstruction for Middle Chinese. We present both numerical and linguistic comparison to previous philologist work (§7). Our phonetic reconstruction aligns to the well-studied phonological category reconstruction to a great extent—it obtains an Adjusted Mutual Information (Vinh et al., 2010) score of over 0.8. A linguistic analysis of the reconstruction result suggests some future research venues.

[3]The original data comes from https://zh.wikipedia.org/wiki/%E4%B8%AD%E5%8F%A4%E9%9F%B3.

[4]The modern Chinese dialects are more like a family of languages (Handel, 2014), and many of them are not mutually intelligible. This paper uses the term 'dialect' instead of 'variety', because we focus on their common ancestor MC.

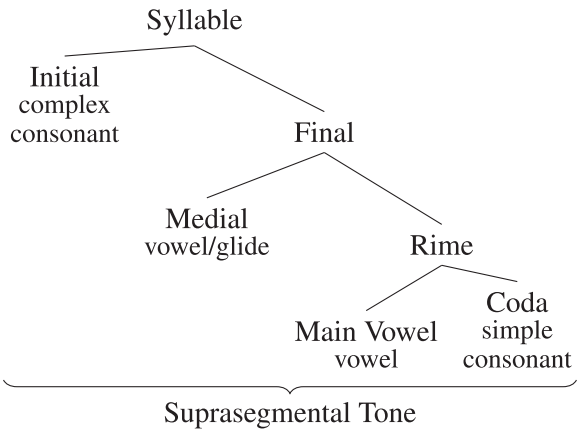


Figure 2: The syllabic structure of MC and Mandarin.

2 Linguistic and Philological Basis

2.1 Syllable Structure

Ancient documents overwhelmingly indicate that Chinese was, from the beginning of its recorded history, a monosyllabic language, in which morphemes are by and large represented by single syllables (Norman, 1988; Shen, 2020). Moreover, the sound pattern of the syllabic structure remained unchanged from Middle Chinese to modern Mandarin. The syllabic structure is composed of an initial segmental consonant (I), a medial (aka on-glide, denoted as M hereafter), a main vowel (V), a coda (or an off-glide), denoted by C hereafter, and a suprasegmental tone (T). The terms 'rime' and 'final' are also frequently used: Rime is the combination of the main vowel and the coda, while final is a combination of the medial, the main vowel and the coda. There are no consonant clusters, i.e., more than one consecutive consonant. See Figure 2 for the hierarchical organization of the above elements.

Below we list three examples:

- 巔 /[tian]: I = t, M = i, V = a, C = n, T = 55
- 眼 /[ian]: I = ∅, M = i, V = a, C = n, T = 214
- 暗 /[an]: I = ∅, M = ∅, V = a, C = n, T = 51

Consonants can only appear as I or C. Consonantal codas are rather simple and have been relatively clearly recorded in rhyme dictionaries. The reconstruction of the associated phonetic values is also clear: 6 categories in total, including nasals [m, n, ŋ] and stops [p, t, k]). This paper aims to complete the reconstruction of the entire consonant system by systematically studying initials.

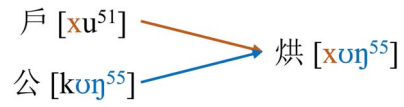


Figure 3: An example of the Fǎnqiē spelling.

2.2 Fǎnqiē Spelling

Fǎnqiē is a traditional method to indicate the pronunciation of a character in question. In the Fǎnqiē spelling, two characters are selected as two spellers to represent the pronunciation of the character (denoted as X) in question: The first character (X_u) is called the upper speller and shares the same I with X; the second character (X_l) is called the lower speller and shares the same M, V, E, and T with X. Take Figure 3 for example. To partially record the pronunciation of 烘, 戶 is employed as the upper speller, while 公 is used as the lower speller.

Zhíyīn 直音 is another method to partially annotate pronunciation. It uses a homophonic character to annotate the character in question. Both Fǎnqiē and Zhíyīn were frequently used.

A rhyme dictionary is a type of ancient Chinese dictionary that collates characters by tone and rime. In rhyme dictionaries, there are three types of important phonological information: rhyme categories, Fǎnqiē spellings, and Zhíyīn notations. The Qiēyùn is a renowned rhyme dictionary that encapsulates the phonology of MC. Chinese philologists have been working on it to derive phonological analysis for centuries.

2.3 Modern Chinese Dialects

Modern Chinese dialects are classified into seven groups in three geographic zones: Mandarin (northern zone); Wu, Min, Xiang (central zone); Gan, Hakka, and Yue (southern zone) (Norman, 1988; Li and Xiang, 2013). Most scholars believe that they are all descendants of MC, and therefore provide valuable information to reconstruct phonetic values.

3 Representing Phonemes

Representing phonemes in a formal way plays an essential role in computational reconstruction. Following Generative Phonology (Chomsky and Halle, 1968), we use distinctive features to represent phonemes. Hayes (2011) proposes a feature

set for all human languages. Any specific language only uses a subset of it to mark phonemic contrasts. To compactly represent Chinese and its modern varieties, we propose a Chinese-specific set. Reducing the total number of distinctive features can also boost the efficiency in solving the corresponding optimisation problem.

3.1 Distinctive Features

Distinctive features provide a systematic way to identify and represent phonemes. Each phoneme is represented as and collectively defined by a bundle of binary features (Hayes, 2011, p. 71). The negative (−) and the positive (+) annotations are used to indicate the absence or presence of a feature. Below is an example:

$$\text{Pom:}= \begin{bmatrix} - \text{ syllabic} \\ - \text{ sonorant} \\ + \text{ stop} \\ - \text{ nasal} \\ + \text{ labial} \\ - \text{ voice} \end{bmatrix} \begin{bmatrix} + \text{ syllabic} \\ + \text{ sonorant} \\ - \text{ stop} \\ - \text{ nasal} \\ + \text{ low} \\ + \text{ back} \\ - \text{ round} \end{bmatrix} \begin{bmatrix} - \text{ syllabic} \\ + \text{ sonorant} \\ + \text{ stop} \\ + \text{ nasal} \\ + \text{ labial} \\ + \text{ voice} \end{bmatrix}$$

It is straightforward to formalize the bundle of features as a vector, which can be used to measure the distance between phonemes.

3.2 Our Feature Set

We propose the following modification of Hayes (2011) to obtain a feature set for Chinese.

Remove Some Features Two types of features are not considered: (1) features that can be represented by other features, and (2) 'tap' and 'trill'.[5]

Merge Some Features The features in Generative Phonology are binary. By combining comparable and orderable features into multi-valued ones, we can reduce the number of features. For example, Hayes (2011) uses 4 features (syllabic, consonantal, approximant, sonorant) to describe the sonority hierarchy, while we combine them into one feature, 'sonority', with 5 graduable values.

Our feature set is summarized in Table 1. We have 14 features in total, reduced from 25 in Hayes (2011). Accordingly, we use a 14-dimensional vector to represent a phoneme for computation.

Independent vs Dependent Features In both the Hayes feature set and ours, some features are meaningful[6] only when some other features at higher levels take certain values. We refer to features that decide whether other features are meaningful as 'I-features' (independent feature), and those determined by I-features as 'D-features' (dependent feature).

Zero Value Hayes (2011) uses the digit 0 (zero feature) to represent meaningless D-features. Some syllables in Chinese lack initial consonants, and Chao (1968, pp. 18–23) suggests calling them 'zero initials'. Accordingly, we use digit 0 to represent zero initials. 0-valued I-features occur when and only when the corresponding character is initialless, while 0-valued D-variables occur when and only when they are meaningless. stop

4 The Optimization Model

Mixed Integer Programming (MIP) is an optimization problem in which some but not necessarily all variables are constrained to be integers. MIP has been widely applied in many NLP tasks, e.g., dependency parsing (Riedel and Clarke, 2006), semantic role labeling (Riedel and Clarke, 2005), coreference resolution (De Belder and Moens, 2012), as well as some more recent applications, e.g., exemplar selection for in-context learning (Tonglet et al., 2023)

We introduce our MIP model for phonetic reconstruction as follows. The objective function and constraints are detailed in §4.1 and §4.3 separately. An essential component of the objective function is measuring the distance between two phonetic feature vectors, for which we propose a mathematically sound distance function in

[5]Taps, flaps, and trills are uncommon in modern Chinese dialects (Zhu, 2008), and they do not appear in our dataset. No scholars have used taps, flaps, and trills to reconstruct MC. Existing research shows their close connection with the affix ér 儿 (Su, 2019), but there is still no consensus on the timing and process of their formation.

[6]The words 'meaningful' and 'meaningless' used here correspond to the description 'not to care' in Hayes (2011, p. 91): 'in most languages with plain /p/, the position of the tongue body during the production of this sound is simply whatever is most articulatorily convenient, given the neighboring sounds. ⋯ the tongue body does not adopt any particular position during the /p/; ⋯ In this sense, the /p/ could be said truly 'not to care' about values for dorsal features.'

	Name	Value
Manner Feature	sonority	5: vowel, 4: glide, 3: liquid, 2: nasal, 1: obstruent
	continuant	1: fricatives, liquids, glides, laterals -1: stops, affricates, nasals
	delayed release[1]	1: fricatives, affricates, -1: stops
Place Feature	labial	1: articulated with the lips
	labiodental[2]	1: articulated by touching the lower lip to the upper teeth
	coronal	1: articulated with the tongue blade and/or tip
	anterior[3]	1: (front) dental, alveolar, -1: (post) palato-alveolar, retroflex
	distributed[3]	1: (blade, laminal) dental, palato-alveolar -1: (tip, apical) alveolar, retroflex
	lateral	1: distinguishes [l] from other coronal liquids and [ɬ, lʒ] from other coronal fricatives.
	dorsal	1: articulated with the tongue body
	high[4]	3: velar, 2: uvular, 1: pharyngeal
	front[4]	3: fronted velar, 2: central velar, 1: back velar, uvular, pharyngeal
Laryngeal Feature	voice	1: voiced, -1: voiceless
	spread glottis	1: [h], breathy vowels, and aspirated consonants.

Table 1: Our feature set. In the 'Value' column, the number before the colon is the possible value of the feature, while the right of the colon is the condition for taking this value. If there is only value '1', it means that the feature is '-1' under all other circumstances. [1]Only meaningful for obstruents (i.e., when sonority is 1). [2]Only meaningful when [+labial]. [3]Only meaningful when [+coronal]. [4]Only meaningful when [+dorsal].

X	character of which the initial is to be reconstructed
X_u	upper speller of character X, with its initial to be reconstructed
S_{fq}	set of all character–speller pairs (X, X_u)
L/l	set of modern dialects/a modern dialect
F	14-dimensional phonetic feature vector
F^j	j-th dimension of phonetic feature vector F
$F_l(X)$	phonetic feature vector that encodes X's initial in dialect l (known)
$F_{\text{MC}}(X)$	phonetic feature vector that encodes X's initial in MC (to be solved)
S_I/S_D	set of independent/dependent features
τ	function that maps D-feature j to the corresponding I-feature $\tau(j)$
$d(F_1, F_2)$	distance between feature vectors F_1 and F_2
f	general distance function, e.g., p-norm
$g_{j,\tau(j)}(F_1, F_2)$	distance function between F_1 and F_2 according to D-feature j and I-feature $\tau(j)$

Table 2: A summary of mathematical notations used to illustrate our model.

§4.2. Mathematical notations used in §4.1–§4.3 are summarized in Table 2.

4.1 The Objective Function

To phonetically reconstruct MC, we consider two information sources: Fǎnqiē/Zhíyīn and varieties. Fǎnqiē/Zhíyīn reveals homophonic relationships between characters of MC, and each descendent dialect partially reflects MC's phonetic structure. Formally, assume we have a set of characters under consideration, denoted as S. The construction of S is discussed in §6.1. Each character $X \in S$ has at least one upper Fǎnqiē or Zhíyīn speller, denoted as $X_u \in S$. We collect all character–speller pairs and define the set $S_{\text{fq}} = \{(X, X_u) : X \in S\}$. Let L denote the set of modern dialects. The pronunciation of any character $X \in S$ in any dialect $l \in L$ is known. Accordingly, the phonetic

feature vector of X's initial in l, denoted as $F_l(X)$, is known. The goal is to infer its phonetic feature vector of MC, denoted as $F_{\text{MC}}(X)$, based on S_{fq} and all known $F_l(X)$ where $l \in L$.

We cast the goal as **minimizing** the overall *distance* between $F_{\text{MC}}(X)$ and $F_{\text{MC}}(X_u)$, and minimising the overall *distance* between $F_{\text{MC}}(X)$ and $F_l(X)$, for all $X \in S$. Assume d is a mathematically sound distance/metric function and $\lambda_{\text{fq}} \in (0, 1)$ is a coefficient then the objective is

$$\begin{aligned}\lambda_{\text{fq}} \sum_{(X,X_u)\in S_{\text{fq}}} & d(F_{\text{MC}}(X), F_{\text{MC}}(X_u)) \\ + (1-\lambda_{\text{fq}}) \sum_{l\in L, X\in S} & d(F_{\text{MC}}(X), F_l(X))\end{aligned} \tag{1}$$

Although the speller X_u is supposed to share the same initial with X in general, we should not model such homophonic relation with constraint $F_{\text{MC}}(X) = F_{\text{MC}}(X_u)$ due to the existence of a considerable number of counterexamples. Such inconsistency exists probably because the Fǎnqiē/Zhíyīn spellings were not devised by one individual but rather collected from various preexisting phonological works, and therefore encoded phonological information of a mixture of diachronically connected languages (Shen, 2020). Instead, we relax the identity restriction by employing a more general distance notion.

4.2 The Distance Function

For each $X \in S$, we set 14 **continuous** variables $F^j (0 \leqslant j \leqslant 13)$ to encode the phonetic value of its initial, each dimension corresponding to a feature. The range of F^j is $[\min\{0, l_j\}, u_j]$, where l_j and u_j are the upper and lower bounds of its corresponding feature in Table 1. Usually, we can use p-norm to measure the distance between two real vectors. However, in our problem, some features are not independent from each other—it is meaningless to discuss a D-feature if its corresponding I-feature does not take a particular value. To solve this problem, we design a new distance function. The mathematical proof of its soundness is provided in Appendix A.

In our solution, the distance w.r.t. I-features is characterized by a general distance function f, e.g., p-norm. We only consider the special case of D-features. We define τ as a function that maps each D-feature to its corresponding I-feature, e.g., maps 'labiodental' to 'labial'. Consider F_1 and F_2, two feature vectors to be compared. Assume $j \in S_D$ is a D-feature, and $\tau(j) \in S_I$ is the corresponding I-feature.

$$s_j \overset{\text{def}}{=\!=} \sup_{F_1, F_2 \in \Omega} f(F_1^j, F_2^j) \tag{2}$$

Denote the set of all valid feature vectors as Ω, which is a subset of $\mathbb{R}^{14}$. We define a function $g_{j,\tau(j)} : \Omega \mapsto \mathbb{R}$ as follows:

$$g_{j,\tau(j)}(F_1, F_2) = c \cdot s_j + (1-c) f(F_1^j, F_2^j) \tag{3}$$

where $c = \min\{f(F_1^{\tau(j)}, F_2^{\tau(j)}), 1\}$. The intuition of the design of $g_{j,\tau(j)}$ is as follows. It is reasonable to compare F_1^j and F_2^j with a normal distance f, when the corresponding I-features $F_1^{\tau(j)}$ and $F_2^{\tau(j)}$ are equal (or very near, since they are continuous). Otherwise, the distance between F_1^j and F_2^j should correspond to the maximum possible distance they can reach.

Now we are ready to define

$$d(F_1, F_2) = \sum_{k\in S_I} f(F_1^k, F_2^k) + \sum_{j\in S_D} g_{j,\tau(j)}(F_1, F_2) \tag{4}$$

4.3 The Restrictions

To obtain a proper phonetic feature vector, we need to ensure the values of its D-features to be consistent with its corresponding I-features. When a D-feature is meaningless w.r.t. its I-feature, we force the D-feature's value to be near 0 by some mathematical tricks. Three cases are consider separately.

Case I: Delayed Release Unless the corresponding I-feature *sonority* is around 1, the value of the *delayed release* feature is meaningless and thus should be around 0. Therefore, the following constraint is considered:

$$F^j \leqslant \max(0, \min(F^{\tau(j)}, 2 - F^{\tau(j)})) \tag{5}$$

Case II: High or Front Unless the corresponding I-feature *dorsal* is around 1, the value of a *high* or *front* feature should be around 0. Ideally, the following constraints are satisfied:

$$F^j \geqslant 1 \ (\text{if } F^{\tau(j)} > 0.5) \tag{6}$$

$$F^j = 0 \ (\text{if } F^{\tau(j)} \leqslant 0.5) \tag{7}$$

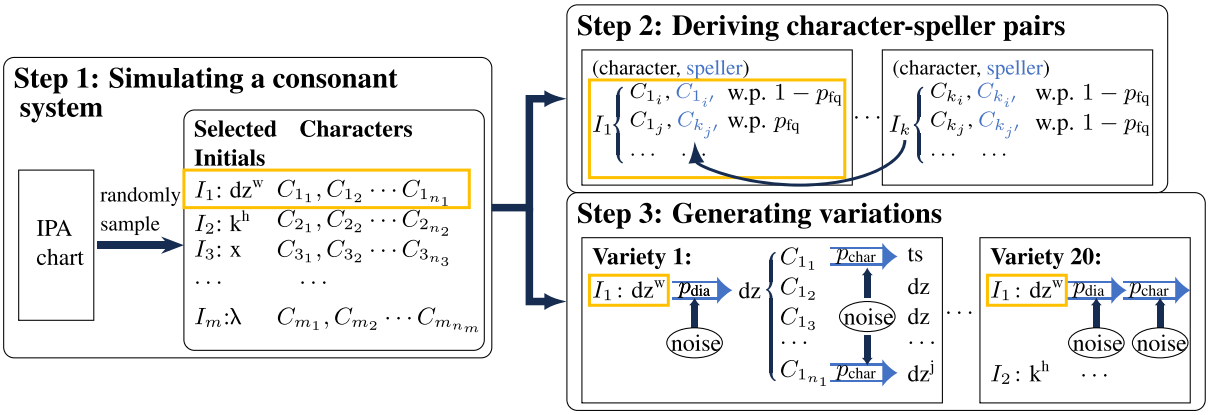


Figure 4: Overview of synthesis data.

To linearize, we define auxiliary variables b (binary, indicator of whether $F^{\tau(j)}$ is larger than 0.5), M (large enough), ϵ (small enough). We have:

$$F^{\tau(j)} \geqslant 0.5 + \epsilon - M \cdot (1 - b) \quad (8)$$

$$F^{\tau(j)} \leqslant 0.5 + M \cdot b \quad (9)$$

$$\max(0, 1 - F^{j}) = 1 - b \quad (10)$$

Case III: Other D-features When the value of the corresponding I-feature is around 1, the absolute value of a D-feature F^j should be around 1. Otherwise the absolute value should be close to 0. We apply the same linearizing trick, with only (10) changed into:

$$|F^{j}| = b \quad (11)$$

To sum up, our model is to minimize Eq. (1) subject to constraints characterized by Eq. (5)–Eq. (11).

5 Validation Experiments on Synthetic Data

Since language reconstruction lacks a definitive ground-truth, it is challenging to discuss the 'correctness' of any reconstruction result. We validate our method on a wide range of synthetic datasets, which hopefully mirror diachronic phonetic change. Starting from a predefined consonant system, we create varieties of it by introducing systematic change and random noise. We then extract character–speller pairs to mimic the Fǎnqiē information. In order to evaluate the effectiveness in reconstructing the predefined consonant system, we apply our model to the varieties as well as character–speller data. The idea of experimentation with synthetic data has been utilized to simulate lexical semantic change (Rosenfeld and Erk, 2018; Shoemark et al., 2019).

5.1 Generating Synthetic Data

Data synthesis is demonstrated in Figure 4. It consists of three steps, as explained as follows.

Step 1: Selecting/Simulating a Consonant System To generate the old stage initial system, we randomly sample an initial set $S_I = \{I_1, I_2, \ldots, I_m\}$ from an IPA chart of consonants,[7] namely, S_{IPA}, with $m \in [35, 40]$. For each initial $I_i (1 \leq i \leq m)$, we generate a set of characters $S_{C_i} = \{C_{i_1}, C_{i_2}, \ldots, C_{i_{n_i}}\}$, with the number of characters n_i falling within $[20, 80]$.

In addition to purely artificial consonant system, we also utilize modern English, German, Mandarin, and reconstructed Latin.[8]

Step 2: Deriving Character–Speller Pairs Following the model of rhyme dictionaries, we assign an artificial Fǎnqiē spelling to each character. Given that not all characters and their spellers share the same initial (§4.1), we introduce variability by randomly assigning a portion p_{fq} of characters to have their upper spellers randomly selected from S_I.

Step 3: Generating variations We generate 20 varieties based on S_I and S_{C_i}s to simulate sound

[7] The IPA chart is based on Hayes's feature spreadsheet (https://brucehayes.org/120a/index.htm#features). Diacritics [$^{\text{h}}$ $^{\text{w}}$ $^{\text{j}}$] are additionally considered.

[8] See https://en.wikipedia.org/wiki/English_phonology, https://en.wikipedia.org/wiki/Standard_German_phonology, https://en.wikipedia.org/wiki/Standard_Chinese_phonology, and https://en.wikipedia.org/wiki/Latin_phonology_and_orthography, respectively.

Example	I-feature $\tau(j)$	D-feature j	Valid Combinations	Shortest Distance
(1.048, 0.905)	sonority	delayed release	**(1, 1)**, (1, −1), (2/3/4/5, 0), (0, 0)	0.143
(0.946, −0.919)	labial	labiodental	(1, 1), **(1, −1)**, (−1, 0), (0, 0)	0.135
(0.499, 0.988)	coronal	anterior	**(1, 1)**, (1, −1), (−1, 0), (0, 0)	0.503
(0.499, 0.499)	coronal	distributed	(1, 1), (1, −1), (−1, 0), **(0, 0)**	0.998
(0.992, 1.952)	dorsal	high	(−1, 0), (**1**, 1/**2**/3), (0, 0)	0.056
(0.992, 2.889)	dorsal	front	(−1, 0), (**1**, 1/2/**3**), (0, 0)	0.119
			Total Distance:	**1.954**

Table 3: Demonstration of how to calculate the 'total distance' from a reconstructed vector. The 'Example' column contains all the $(\tau(j), j)$ value pairs in a reconstructed vector. For each pair, we highlight the shortest L_1 distance between it and all valid combinations with bold font. 'Total Distance' is the sum of all shortest distances over j.

change, denoted as $S_I^v = \{I_1^v, I_2^v, \ldots, I_m^v\}$ and $S_{C_i}^v = \{C_{i_1}^v, C_{i_2}^v, \ldots, C_{i_{n_i}}^v\} (1 \leq i \leq m, 1 \leq v \leq 20)$. We assume that most sound changes are regular, where the phonetic value of an initial influences all characters with that initial in a given variety. To simulate regular sound change, initial I_i can change to any $I_i^v \in S_{\text{IPA}}$ in variety v with probability p_{dia}.

Exceptions to regular change can occur due to various causes, e.g., Wexical borrowing and grammatical analogy, and we model such irregular change by allowing the initial of character $C_{i_j}^v$ to change from I_i^v to any consonant in S_{IPA} with probability p_{char}.

Denote the L_1 distance between I_a and I_b (both in S_{IPA}) as d_{I_a, I_b}.

5.2 Experimental Setup

We use the Gurobi[9] MIP solver for our empirical investigation. We set `MIPGap` to 1e-4, and the `TimeLimit` to 8 hours as the maximal time for calculation. Notably, the obtained solutions usually are not optimal. Nevertheless, they are of relatively good quality to verify the reliability of our model.

We consider the following three metrics to evaluate the goodness of a reconstruction result, when ground-truth is available. The ground-truth consonant system of the experiments in this section is either stochastically sampled from IPA, the reconstructed Latin, or modern English, German, and Mandarin.

Average L_1 Since we represent phonemes by vectors, a straightforward way to evaluate the goodness of reconstruction is to calculate the overall distance between reconstructed vectors and their corresponding ground-truth vectors. To this end, we report the average L_1 distance.

Equal rate A more strict evaluation metric is to reward only when the reconstructed vector is extremely close to its ground-truth. Here, we consider a phoneme as successfully reconstructed only when the L_1 distance between the reconstructed vector and its predefined value is smaller than 10^{-4}, Accordingly, we report the proportion of successfully constructed initials as **equal rate**.

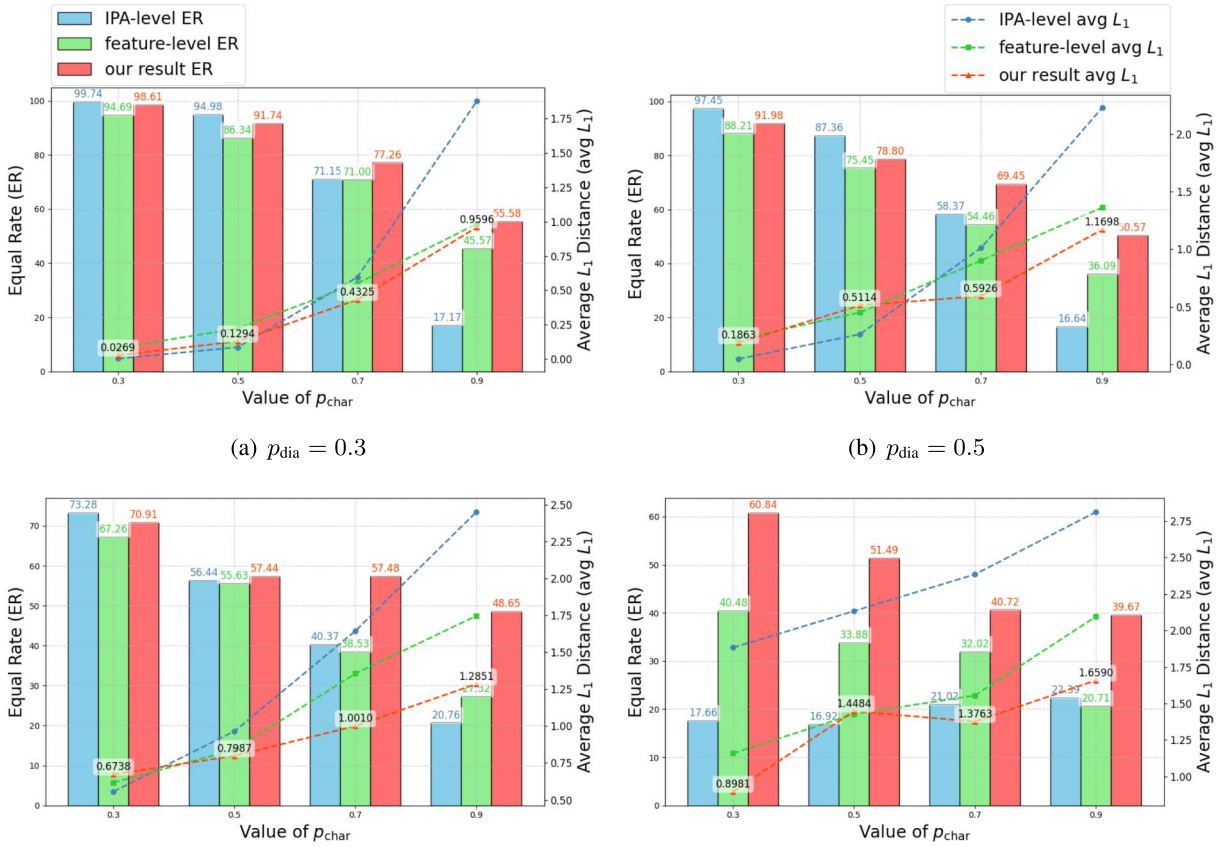
Soundness of Phonetic Feature Vector The reconstructed results should be valid phonemes that satisfy the constraints on D-features listed in §4.3. Our features are continuous, and we thus propose to measure their deviation from the constraints rather than classifying them as strictly 'valid' or 'invalid'. For each D-feature j and its corresponding I-feature $\tau(j)$, we consider the shortest L_1 distance between our result and all the valid values of $(j, \tau(j))$. Table 3 serves as an example, listing all $(j, \tau(j))$ pairs and their valid combinations. We report **sound rate**—the proportion of characters with a total distance less than 10^{-4}.

5.3 Results and Analysis

Our main results are shown in Figure 5, where we compare our results with two baselines. Since the phenomenon of characters having different initials from their upper spellers is not common in real data, we set $p_{\text{fq}} = 0.1$.

We report two versions of majority vote results as baseline: IPA-level and feature-level. Considering the randomness in generating consonant systems and their variations, we conduct the experiment three times and report the average for each setting. In the IPA-level majority vote, for

[9] `https://www.gurobi.com`.



(a) $p_{\text{dia}} = 0.3$ (b) $p_{\text{dia}} = 0.5$

(c) $p_{\text{dia}} = 0.7$ (d) $p_{\text{dia}} = 0.9$

Figure 5: Comparison between our results and baselines with synthetic data that starts from Latin consonant system, with respect to equal rate (ER) and average L_1 distance (avg L_1).

each character, we select the most frequent IPA phoneme from all 20 dialects to reconstruct its initial. For feature-level voting, we choose the most frequent value for each feature of each character.

The IPA-level majority vote achieves the highest equal rate when p_{fq} and p_{char} are small, but its equal rate declines rapidly as randomness increases. In contrast, the feature-level majority vote performs better under high p_{fq} and p_{char} settings. Compared with the baselines, our model significantly outperforms both in terms of equal rate and average L_1 distance across most settings, particularly when p_{fq} and p_{char} are large, which highlights the robustness of our model.

To estimate a proper value of the change rate p_{dia} is challenging. We use the following geometric method to estimate a lower bound of p_{dia} from the ancestral form (MC) to modern dialects based on the differences among all modern dialects. First, we measure how different any two dialects are by calculating the percentage of characters with different initial pronunciation. For example, this proportion between Beijing and Guangzhou is 65.70%. Intuitively, halving such a difference degree gives a lower bound though the estimation based on any single pair of dialects should be far from being tight. We then leverage the concept of high-dimensional sphere to integrate all possible pairs of dialects. The key idea is as follows. Each dialect is viewed as a point in a high-dimensional space and the percentage of characters with different pronunciation of initials is viewed as the distance between the corresponding pair of dialects. It is easy to see that such a distance measurement satisfies the triangle inequality. The radius of the minimal high-dimensional sphere that covers all dialects serves as a (loose) lower bound. Based on the data from Zihui, we first convert the distance matrix into coordinates using the algorithm proposed by Crippen (1978), then apply the algorithm proposed by Fischer et al. (2003) to determine the radius of the minimal sphere, obtaining an empirical value of 0.4180. The maximum of such lower bound is 0.8844, when any

Setting	SR_1	n_1	SR_2	n_2	stat z
(.3, .3)	1.0000	1078	0.9733	3138	5.4191
(.3, .5)	0.9847	1110	0.9515	3290	4.8737
(.3, .7)	0.9991	1107	0.9338	3436	8.6460
(.3, .9)	0.9950	1001	0.9586	3160	5.6477
(.5, .3)	0.9872	1173	0.9579	3228	4.7228
(.5, .5)	0.9943	1047	0.9558	3109	5.9036
(.5, .7)	0.9895	954	0.9422	3216	6.0635
(.5, .9)	0.9826	1037	0.9731	3158	1.7152
(.7, .3)	0.9829	994	0.9671	3108	2.5809
(.7, .5)	0.9804	1172	0.9245	3174	6.8637
(.7, .7)	0.9942	1030	0.9389	3165	7.2458
(.7, .9)	0.9889	1169	0.9247	2985	8.0104
(.9, .3)	0.9865	1040	0.9678	3190	3.1966
(.9, .5)	0.9952	1048	0.9424	3179	7.1880
(.9, .7)	0.9904	1038	0.9315	3272	7.2954
(.9, .9)	0.9873	1004	0.9141	3339	8.0252

Table 4: Two-proportion z-test between our result and feature-level majority vote with respect to sound rate (SR). SR_1 and SR_2 represent the sound rates of our model and the baseline, respectively, while n_1 and n_2 indicate the sample sizes (number of characters) of our model and the baseline.

two dialects have totally different pronunciation (in other words, the distance is 1). The empirical result suggests to utilize a high ratio of sound change.

Since the vectors derived from feature-level voting do not necessarily correspond to valid phonemes, we compare its sound rate with that of our model. Although our model maintains a high sound rate across all settings, we perform a two-proportion z-test to determine whether the difference between our model and the feature-level majority vote in terms of SR is statistically significant. The null hypothesis is that the SR of our method is equal to that of the baseline. The results are reported in Table 4. If we test this hypothesis at a significance level of 95%, for all settings except (0.5, 0.9), the statistic z exceeds $z_{0.975} = 1.96$, and we can reject the null hypothesis. For the (0.5, 0.9) setting, $z = 1.7152$ is still larger than $z_{0.95} = 1.65$. Therefore, we conclude that our model performs better than the baselines in terms of sound rate.

Though remarkable, we should exercise caution when interpreting the results. Naturally occuring sound changes display greater regularity in some cases but are much less regular in others.

Base Distance Function f The general distance function f (defined in §4.1) is also a hyperparameter. In Figure 5, it is set as $f(x_1, x_2) = |x_1 - x_2|$. We did a number of auxiliary experiments with different p-norm distance functions. Even the quadratic function significantly increases the difficulty to the corresponding optimisation problems, without substantial improvement in performance. It seems that the only practical option for f is L_1. All following experiments are based on this choice.

Weight of Fǎnqiē (λ_{fq}) We adjust the weight of terms related to Fǎnqiē in the objective function, i.e., λ_{fq} in Eq. (1). By default, λ_{fq} is set to 0.5, and we explore its effect with respect to equal rate in Table 5. Setting λ_{fq} to 0 (i.e., not using Fǎnqiē information) results in a decrease in the equal rate, while increasing it to 0.75 improves the equal rate. However, further increasing it to 0.95 leads to a decline. These experiments suggest that the choice of λ_{fq} is empirical, and we will adjust it accordingly when working with real data (§7.1).

Results with English, German, and Mandarin To evaluate the robustness, we experiment with synthetic data that starts from a consonant system in natural phonology. We choose modern standard English, German, and Mandarin as representatives. We also present results from a random consonant system for comparison. The results are in Table 6, under the setting of (0.1, 0.5, 0.3), with λ_{fq} set to 0.5. The remarkable results reaffirm the reliability of our model.

It is worth noting that reconstructing natural consonant systems is much easier than artificial ones.

5.4 Modeling the Influence of Context

The phonetic value of phonemes can be influenced by its context. For initials in Chinese syllables, the major influential context is the medial that follows. To integrate medial information in MC[10] into our model, we adjust the weight of terms related to Fǎnqiē in the objective function. For each character and speller pair, i.e., $(X, X_u) \in S_{\text{fq}}$, we assign a weight of k to $d(F_{\text{MC}}(X), F_{\text{MC}}(X_u))$ if they share the same medial, otherwise 1. In the basic setting, $k = 1$, and increasing it aims to improve the likelihood of pairs with matching medials sharing the same initial.

Setting k to 3, we present representative results in Table 5. Changing k from 1 to 3 has

[10]The data is provided by Peking University.

	$k=1$				$k=3$
Setting	$\lambda_{\text{fq}}=0$	$\lambda_{\text{fq}}=0.5$	$\lambda_{\text{fq}}=0.75$	$\lambda_{\text{fq}}=0.95$	$\lambda_{\text{fq}}=0.5$
(0.1, 0.3, 0.3)	96.85%	**98.61**%	98.89%	90.35%	98.70%
(0.1, 0.5, 0.5)	75.45%	**78.80**%	79.94%	68.77%	79.27%
(0.1, 0.7, 0.7)	54.37%	**57.48**%	59.13%	53.69%	57.67%

Table 5: Results on Latin consonant system with parameters adjusted. The numbers in the 'Setting' column correspond to $(p_{\text{fq}}, p_{\text{dia}}, p_{\text{char}})$ respectively. k represents the weight assigned to character and speller pairs with matching medials, as defined in §5.4.

	Ger	**Man**	**Eng**	**RND**
ER(%)	96.10	94.48	93.02	84.73
Avg. $L1$	.1894	.0884	.1519	.2549

Table 6: Results with synthetic data that starts from German, Mandarin, English, and the random system.

little influence on equal rate, indicating that medial information has already been well captured. Notably, our model is inherently conditional, as we always consider an initial in a particular character, where the medial, main vowel, and coda are all fixed. When human scholars encounter overlapping heterogeneous information sources, decision-making becomes challenging, while our model provides a possible technique for such challenging issues.

The results demonstrate the flexibility of our model—heterogeneous information can be seamlessly integrated as constraints or terms in the objective function, and be integrated into our model conveniently. Furthermore, the flexibility to adjust weights of different terms allows fine-tuning according to specific requirements.

6 Validation Experiments on Real Data

6.1 Collecting Real Data

We examine our model with the spelling information in Qiēyùn and the phonetic information of 20 dialects in Zihui (1989). Polyphonic characters (characters with multiple pronunciations) are common in both MC and dialects. We treat different pronunciations of the same character as different entries and correlate different vectors to them. The final dataset consists of 1960 different characters and 2661 entries in total.[11]

[11]We will release the dataset for research.

The Qiēyùn Information Only fragments of the original Qiēyùn survived, and most commonly used documents are its revisions. The most accurate revision is Guǎngyùn 廣韻. Though published in the Song Dynasty, Guǎngyùn is commonly believed to record the Qiēyùn system and reflect the status of MC. Guǎngyùn was heavily used in traditional philological research, including Karlgren (1926) and Wang (1957). We collect and integrate information from two electronic versions of Guǎngyùn, separately provided by Peking University and Beijing Normal University.

The Dialect Information Zihui (1989) is a workbook for fieldwork on Chinese dialects. There is information for 20 modern Chinese dialects: Beijing, Jinan, Xi'an, Taiyuan, Wuhan, Chengdu, Hefei, Yangzhou (Mandarin), Suzhou, Wenzhou (Wu), Changsha, Shuangfeng (Xiang), Nanchang (Gan), Meixian (Hakka), Guangzhou, Yangjiang (Yue), Xiamen, Chaozhou, Fuzhou, and Jianou (Min). For each of these dialects, it documents both the phonological system and the phonetic values of representative characters.

Selecting Representative Characters The original Qiēyùn dataset has 25333 entries, but a large proportion of them are rarely used. In contrast, Zihui (1989) contains less than 3000 frequently used characters. We denote the characters included in Guǎngyùn and Zihui (1989) as S_{gy} and S_{zh}, respectively, and denote all characters used as Fǎnqiē spellers of characters in Guǎngyùn as S_{fq} ($|S_{\text{fq}}|$ =1462). Instead of using all available characters, we aim to select a set of representative characters that comprehensively reflect the entire phonological systems. Our selection process involves the following steps:

1. Subtract a smaller set S_* from S_{gy} for subsequent selection. Since Fǎnqiē spellings

λ_{fq}	0.5	0.75	0.95
Matching Rate	66.38%	65.33%	67.96%
Avg. L_2	1.1062	1.1046	1.2432

Table 7: Evaluation with held-out Fǎnqiē.

connect different characters and encapsulate valuable relationships between them, they are essential for deriving phonological categories. Therefore, We define $S_\cap$ to include common characters as well as Fǎnqiē spellers. Specifically, $S_\cap = S_{\text{gy}} \cap (S_{\text{fq}} \cup S_{\text{zh}})$ with 3990 entries.

2. For each character in $S_\cap$, if both its upper and lower spellers are also in $S_\cap$, this character is considered of particular interest. This set is denoted as S^*, which contains 2461 entries.
3. Among S^*, if several characters share the same Fǎnqiē, indicating that they are homophones, we select the first character only, which is often the most frequent character. We denote this set as S_1^*.
4. Finally, we include Fǎnqiē spellers themselves into the selected set to link different entries. Our final representative character set is $S_1^* \cup (S_{\text{fq}} \cap S_\cap)$, with 2661 entries.

6.2 Results and Analysis

Matching to Held-out Fǎnqiē Data Ideally, each character should share the same initial with its Fǎnqiē/Zhíyīn speller. We randomly take 70% of Fǎnqiē/Zhíyīn material for MC reconstruction, and use the remaining 30% for evaluation. We consider a character–speller pair as having matching initials if the L_2 distance between a character's reconstructed initial vector and that of its upper speller's is smaller than 10^{-4}. We report the average L_2 distance between the reconstructed initials in character–speller pairs and the rate of pairs with matching initials as the **matching rate**.

The results are shown in Table 7. A large portion of held-out character–speller pairs have matching reconstruction, affirming the self-consistency of our results.

7 Reconstruction Results and Discussion

We obtain our final reconstruction result by applying the method to all available data introduced in §6.1. Our reconstruction is based on individual characters, while existing results are based on phonological categories. A straightforward way to obtain category-centric result is averaging phonetic feature vectors of all characters belonging to the same phonological category. The nearest IPA phonemes to the averaged vectors can be directly used for comparison to previous manual results by philologists.[12]

7.1 Numerical Evaluation

The phonetic vectors resulting from our model should form clusters that align with phonological categories in Guǎngyùn to some extent. Based on this assumption, we develop a clustering-based method to evaluate the overall quality of a reconstruction result. We cluster the phonetic vectors with KMeans with a predefined number of clusters equal to 37.[13] We then report the *adjusted mutual information* (AMI; Vinh et al., 2010), an information-theoretic measure, between the automatic clustering and predefined phonological categories. Given two clusterings U and V,

$$\text{AMI}(U, V) = \frac{\text{MI}(U, V) - \mathbb{E}[\text{MI}(U, V)]}{\text{avg}(\text{H}(U), \text{H}(V)) - \mathbb{E}[\text{MI}(U, V)]}$$

where H is the Shannon entropy, MI is the mutual information, and $\mathbb{E}$ is expectation. Perfectly matched clusters yield an AMI of 1, while random cluster assignment yields 0. The numbers of samples and clusters are not necessarily the same.

Figure 6 shows our results, along with two baselines (majority vote and single best dialect). AMI is largely influenced by the value of λ_{fq}, indicating the effectiveness of Fǎnqiē in deriving categories.

Since we are dealing with 20 dialects but only a single set of Fǎnqiē spellings, setting λ_{fq} to 0.95 is a natural choice. However, since the information obtained from the dialects and Fǎnqiē lacks a common scale, it is difficult to make direct comparisons. As a result, we cannot theoretically determine the optimal weighting for each information source, and the choice of λ is therefore largely empirical.

[12] A comprehensive summary of the result can be found at `https://github.com/LuoXiaoxi-cxq/Reconstruction-of-Middle-Chinese-via-Mixed-Integer-Optimization`.

[13] There are 38 categories in the manual categorial reconstruction of Guǎngyùn. Our dataset excludes characters with the 俟 category.

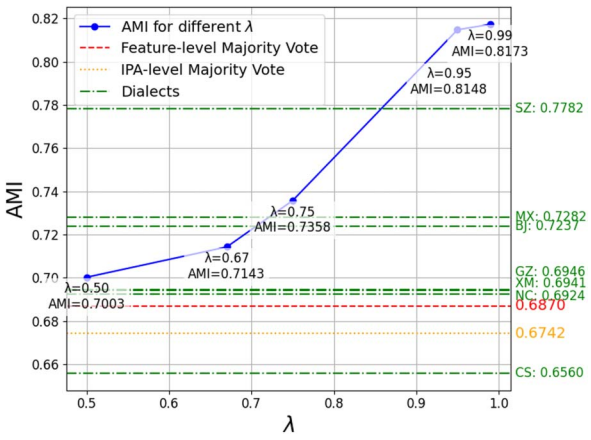


Figure 6: AMI values with different λ compared to baseline results. AMI values for seven individual dialects (BJ: Beijing, SZ: Suzhou, CS: Changsha, NC: Nanchang, MX: Meixian, GZ: Guangzhou, and XM: Xiamen) are presented, each representing one dialect group (see §6.1). The single best dialect is Suzhou, with the highest AMI (0.7782). In feature-level majority vote, features are aggregated and voted individually. The KMeans algorithm is then applied to the voted feature vector, yielding an AMI of 0.6870.

When λ_{fq} is set to 0.95, our model achieves an AMI of 0.8148 and outperforms the baselines, indicating a high degree of similarity between the phonetic reconstruction by us and the manual phonological reconstruction by philologists. When λ_{fq} is set to 0.99, the AMI increases to 0.8173, although the change is minimal.

In contrast to Table 5, where setting λ_{fq} to 0.95 results in a decrease in equal rate, increasing λ_{fq} actually improves AMI when applied to real data. This phenomenon reflects the difference in synthetic and real data, emphasizing the importance of adjusting λ_{fq} based on specific situations.

It is worth noting that when λ_{fq} is set to 0 (i.e., without using Fǎnqiē), AMI drops to 0.5892. This further reflects the critical role of Fǎnqiē information when dealing with real data.

In some auxiliary experiments that are not reported in this paper, we also used hierarchical clustering to further study the impact of the number of clusters. Results show that the difference in AMI between hierarchical clustering and KMeans is within 0.05, regardless of the specific setting.

7.2 Comparison to Existing Results

Our model successfully reconstructs most categories with consensus among philologists, such as bāng 幫 pāng 滂 míng 明 duān 端 tòu 透 ní 泥.[14] Similar to the computational operationalization of the historical comparative approach to Indo-European languages (List et al., 2022), our study confirms the usefulness of computation in linguistic inquiry. The differences between different reconstruction results may provide new evidence for philologists and linguists to consider and therefore are useful too. Such differences are mainly attributed to two factors.

Different dialects changed in different directions. For example, it is generally believed that Wu dialects retained all the voiced stops, while most other dialects became devoiced (Handel, 2014, pp. 224–225).[15] In phonology, devoicing refers to a sound change where a voiced consonant becomes voiceless due to the influence of its phonological environment. This process is common across many languages and is a part of Grimm's law. Our current model, however, cannot differentiate in what aspects a dialect changed most and in what aspects it stayed constantly. It treats different dialects with equal weight on different phenomena. Consequently, a character's reconstructed initial tends to be closer with the phonetic value that is more commonly observed across various dialect pronunciations. For example, our model fails to reconstruct the 'voiced' feature for categories that are assumed to be voiced by philologists, e.g., bìng 並 dìng 定 cóng 從 xié 邪. Our model also has difficulty distinguishing the zhī 知 zhuāng 莊 zhāng 章 groups, which have similar pronunciations in most modern Chinese dialects.

Our current model only contains the most basic information—Fǎnqiē spellings and modern Chinese varieties. Other types of information, including rhyme tables, e.g., Yùnjìng 韻鏡, and sino-xenic[16] pronunciations are not integrated into our model at present. Rhyme tables provide additional information about the voiced/voiceless

[14] All the Chinese characters used in §7.2 are categorical labels representing initial categories in traditional Chinese phonology. For example, characters fāng 方, fǔ 府, bó 博, bǐ 彼, and many other characters are assumed to have the same initial in MC, and philologists use bāng 幫 to represent their common initials.

[15] For example, the character tóng 同 is believed to had a voiced initial dìng 定 in MC. In Wu dialect, its initial is [d], while in most other dialects, it is [t^h].

[16] Sino-xenic vocabularies are large-scale and systematic borrowings of the Chinese lexicon into the Japanese, Korean, and Vietnamese languages. See https://en.wikipedia.org/wiki/Sino-Xenic_vocabularies for details.

feature of initials, which is crucial for philologists' manual reconstruction. It is another reason why our model fails to reconstruct voiced initials.

Because of the limitation in information sources, our model cannot provide definitive answers to some debatable problems, such as whether categories ní 泥 and níang 孃 are the same initial. It is generally believed that there is no distinction between the two categories in most modern Chinese varieties (Tseng, 2019, p. 228, Li, 1956, p. 126). Though Li (1956, pp. 125–126) and Shao (1982, pp. 98–101) have opposite opinions about this problem, they both used Sanskrit-Chinese pronunciations as the main evidence. However, in our model, with materials restricted to dialects, the reconstruction of ní 泥 and níang 孃; appears similar.

7.3 Extension

In principle, our method can be generalized to other languages. However, in practice, our model requires phoneme-level alignment between each protoform's reflexes. For Chinese, this alignment occurs naturally, as each Chinese character typically corresponds to a morpheme, and morphemes are largely represented by single syllables that follow specific patterns, as described in §2.1.

Sound change is a central focus in linguistic research, and our model can engage with it in two ways. First, incorporating common patterns of sound change as constraints into our model is a possible future direction. Second, by analyzing $F_l(X) - F_{\text{MC}}(X)$ in Eq. (1), we may identify potential sound changes in terms of distinctive features, such as devoicing.

8 Related Work

8.1 Computational Reconstruction

Bouchard-Côté et al. (2007a; 2007b; 2009; 2013) offered a series of influential work about unsupervised proto-word reconstruction, which requires an existing phylogenetic tree to infer the ancient word forms based on probability estimates for all the possible phoneme-level edits on each branch of the tree. The edit model parameters and unknown ancestral forms are jointly learned with an EM algorithm.

Following this series of work, He et al. (2023) also used Monte-Carlo EM algorithm but neural networks to parameterize the edit models, in order to express more complex phonological and the nonadjacent changes, achieving a notable reduction in edit distance from the target word forms. However, his highly parameterized edit models were designed for large cognate datasets with few languages, and may not be possible to train them on datasets with more languages but fewer datapoints per language.

In supervised protolanguage reconstruction, the models are easier to evaluate. Meloni et al. (2021) trained a GRU-based encoder-decoder architecture on cognates from five Romance languages to predict their Latin ancestors, and achieved low error from the ground-truth. Kim et al. (2023) updated Meloni et al.'s model with the Transformer and achieved better performance.

List et al. (2022) proposed a new framework for supervised reconstruction that combines automated sequence comparison with phonetic alignment analysis, which deals with the losing reflexes problem, and sound correspondence pattern detection, which models phonetic environments of sound change.

Lu et al. (2024) proposed a multi-model reconstruction system that improves its reconstructions via predicting the reflexes given a protoform. Their system consists of a beam search-enabled sequence-to-sequence reconstruction model and a sequence-to-sequence reflex prediction model that serves as a reranker, surpassing state-of-the-art protoform reconstruction methods on three of four Chinese and Romance datasets.

8.2 Middle Chinese Phonology

Phonetic reconstruction of phonological categories was pioneered by Karlgren (1926). Following the methodology of Karlgren (1926), subsequent scholars, including Li (1971), Wang (1957), Pulleyblank (1984), and Baxter (1992), made modifications to the methodology and proposed their reconstructions of MC.

In recent decades, some scholars have questioned the assumptions, methodology, and conclusions of Karlgren's approach. A critical view is exemplified by Norman and Coblin (1995). Norman advocated a data-centered approach to Chinese historical phonology, predicated on the collection, analysis, and comparison of spoken-language data. His controversial reconstruction of Proto-Min (Norman, 1973, 1974) is an example.

9 Conclusion

We propose a novel, MIP-based method for phonetic reconstruction for Middle Chinese, and validate its effectiveness on a wide range of synthesis and real data. Similar to the automation of the historical comparative approach to Indo-European languages, our study confirms the usefulness of computation in linguistic inquiry. The optimization-based architecture is flexible—different information can be integrated as either an element in the objective function, constraints, or both. It is also applicable to the reconstruction problem of other languages. We leave both for future work.

Acknowledgments

We would like to express our sincere gratitude to the reviewers for their valuable comments, which greatly broadened our perspective and significantly improved the quality of our work. We would also like to thank Kechun Li for her suggestions.

References

William H. Baxter. 1992. *A Handbook of Old Chinese Phonology*. De Gruyter Mouton, Berlin, New York. https://doi.org/10.1515/9783110857085

Alexandre Bouchard-Côté, David Hall, Thomas Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229. https://doi.org/10.1073/pnas.1204678110, PubMed: 23401532

Alexandre Bouchard-côté, Percy S. Liang, Dan Klein, and Thomas Griffiths. 2007b. A probabilistic approach to language change. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Alexandre Bouchard-Côté, Percy Liang, Thomas Griffiths, and Dan Klein. 2007a. A probabilistic approach to diachronic phonology. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 887–896, Prague, Czech Republic. Association for Computational Linguistics.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved reconstruction of protolanguage word forms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 65–73, Boulder, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/1620754.1620764

Yuen ren Chao. 1968. *A Grammar of Spoken Chinese*. University of California Press.

Xinxiong Chen. 2005. *Shengyunxue [Chinese phonology]* 聲韻學. Taipei: Wenshizhe Publishing House.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Studies in English. Harper & Row.

G. M. Crippen. 1978. Note rapid calculation of coordinates from distance matrices. *Journal of Computational Physics*, 26(3):449–452. https://doi.org/10.1016/0021-9991(78)90081-5

Jan De Belder and Marie-Francine Moens. 2012. Coreference clustering using column generation. Martin Kay and Christian Boitet, editors, In *Proceedings of COLING 2012: Posters*, pages 245–254, Mumbai, India. The COLING 2012 Organizing Committee.

Tonghe Dong. 2004. *Hanyu Yinyunxue [Chinese Phonology]* 漢語音韻學. Beijing: Zhonghua Shuju.

Kaspar Fischer, Bernd Gärtner, and Martin Kutz. 2003. *Fast smallest-enclosing-ball computation in high dimensions*. pages 630–641. https://doi.org/10.1007/978-3-540-39658-1_57

Zev Handel. 2014. *Historical Phonology of Chinese*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118584552.ch22

Bruce Hayes. 2011. *Introductory Phonology*. Blackwell Textbooks in Linguistics. Wiley.

Andre He, Nicholas Tomlin, and Dan Klein. 2023. Neural unsupervised reconstruction of protolanguage word forms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 1636–1649, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.91

Bernhard Karlgren. 1926. *Zhongguo Yinyunxue Yanjiu [Study on Chinese phonology]* 中國音韻學研究. Shangwu Yinshuguan.

Young Min Kim, Kalvin Chang, Chenxuan Cui, and David R. Mortensen. 2023. Transformed protoform reconstruction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–38, Toronto, Canada. Association for Computational Linguistics.

Fang-Kuei Li. 1971. Shangguyin yanjiu [studies on old Chinese pronunciation] 上古音研究. *The Tsing Hua Journal of Chinese Studies*, pages 1–61.

Rong Li. 1956. *Qieyun Yinxi [Phonological System of Qieyun]* 切韻音系. Beijing: Kexue Chubanshe.

Xiaofan Li and Mengbing Xiang. 2013. *Hanyu Fangyanxue Jichu Jiaocheng [Fundamentals of Chinese Dialect Studies]* 漢語方言學基礎教程. Beijing: Peking University Press.

Johann-Mattis List, Robert Forkel, and Nathan Hill. 2022. A new framework for fast automated phonological reconstruction using trimmed alignments and sound correspondence patterns. In Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, editors, *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 89–96, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.lchange-1.9

Liang Lu, Jingzhi Wang, and David R. Mortensen. 2024. Improved neural protoform reconstruction via reflex prediction. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8683–8707, Torino, Italia. ELRA and ICCL.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.353

Jerry Norman. 1973. Tonal development in Min. *Journal of Chinese Linguistics*, 1(2):222–238.

Jerry Norman. 1974. The initials of proto-Min. *Journal of Chinese Linguistics*, 2(1):27–36.

Jerry Norman. 1988. *Chinese*. Cambridge Language Surveys. Cambridge University Press.

Jerry Norman and South Coblin. 1995. A new approach to Chinese historical linguistics. *Journal of the American Oriental Society*, 115(4):576–584. https://doi.org/10.2307/604728

Wuyun Pan. 2000. *Hanyu LiShi Yinyunxue [Chinese Historical Phonology]* 漢語歷史音韻學. Shanghai: Shanghai Educational Publishing House.

Edwin G. Pulleyblank. 1984. *Middle Chinese: A Study in Historical Phonology*. Vancouver: University of British Columbia Press. https://doi.org/10.59962/9780774854580

Sebastian Riedel and James Clarke. 2005. Integer linear programming inference for conditional random fields. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 736–743, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1102351.1102444

Sebastian Riedel and James Clarke. 2006. Incremental integer linear programming for non-projective dependency parsing. In Dan Jurafsky and Eric Gaussier, editors, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 129–137, Sydney, Australia. Association for Computational Linguistics. https://doi.org/10.3115/1610075.1610095

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In Marilyn

Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1044

Rongfen Shao. 1982. *Qieyun Yanjiu [Studies on Qieyun]* 切韻研究, Beijing: Chinese Academy of Social Sciences.

Zhongwei Shen. 2020. *A Phonological History of Chinese*. Cambridge University Press. https://doi.org/10.1017/9781316476925

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1007

Junbo Su, Central China Normal University. 2019. Revisiting the trill in Hubei dialects. *Dialect*, (2):228–232.

Jonathan Tonglet, Manon Reusens, Philipp Borchert, and Bart Baesens. 2023. SEER: A knapsack approach to exemplar selection for in-context HybridQA. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13569–13583, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.837

Wen-Hsing Tseng. 2019. Research on the Evolution of [n-], [ɳ-]and [ɳʐ-] in Ming-Qing Rhyme Books and Rhyme Charts and Chinese Dialects 泥娘日三母於明清韻書韻圖及漢語方言的演變研究. Master's thesis, National Chengchi University. https://doi.org/10.6814/THE.NCCU.CHI.001.2019.A08

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.

Li Wang. 1957. *Hanyu Shigao [A Draft History of the Chinese Language]* 漢語史稿. Beijing: Zhonghua Shuju.

Xiaonong Zhu. 2008. On liquids. *Studies in Language and Linguistics*, 28(4):1–13.

Zihui. 1989. *Hanyu Fangyin Zihui [A Pronouncing Dictionary of Han Dialects]* 漢語方音字彙. 2 edition. Beijing: Wenzi Gaige Chubanshe.

Appendix

Here, we prove that the distance function (3) defined in §4.2 is mathematically sound.

It is easy to see:

1. $g_{j,k}(F_1, F_2) \geqslant 0$.
2. $g_{j,k}(F_1, F_2) = g_{j,k}(F_2, F_1)$.

Now we consider the triangle inequality.

Proposition 1. $\forall F_1, F_2, F_3 \in \Omega$, $\forall$ *indices of paired D-feature and I-feature* j *and* k*,*

$$g_{j,k}(F_1, F_2) + g_{j,k}(F_1, F_3) \geqslant g_{j,k}(F_2, F_3). \quad (12)$$

Proof. Let $c_{st} = \min\{f(X_s^k, X_t^k), 1\}$. We have

$$\begin{aligned}
\Delta =& g_{j,k}(F_1, F_2) + g_{j,k}(F_1, F_3) - g_{j,k}(F_2, F_3) \\
=& (c_{12} + c_{13} - c_{23}) \cdot s_j + (1 - c_{12}) f(F_1^j, F_2^j) \\
&+ (1 - c_{13}) f(F_1^j, F_3^j) - (1 - c_{23}) f(F_1^j, F_2^j) \\
\geqslant& (c_{12} + c_{13} - c_{23}) \cdot s_j + (1 - c_{12}) f(F_1^j, F_2^j) \\
&+ (1 - c_{13}) f(F_1^j, F_3^j) \\
&- (1 - c_{23})[f(F_1^j, F_2^j) + f(F_1^j, F_3^j)] \\
=& (c_{12} + c_{13} - c_{23}) \cdot s_j - (c_{12} - c_{23}) f(F_1^j, F_2^j) \\
&- (c_{13} - c_{23}) f(F_1^j, F_3^j)
\end{aligned} \quad (13)$$

There are three cases: If $c_{23} < c_{12}$ and $c_{23} < c_{13}$, then

$$\begin{aligned}
\Delta \geqslant& (c_{12} - c_{23})[s_j - f(F_1^j, F_2^j)] \\
&+ (c_{13} - c_{23})[s_j - f(F_1^j, F_3^j)] \geqslant 0.
\end{aligned}$$

If $c_{23} > c_{12}$ and $c_{23} > c_{13}$, then

$$\Delta \geqslant (c_{12} + c_{13} - c_{23}) \cdot s_j \geqslant 0.$$

feature	[z]	[f]	dist.	Note
continuant	-1	1	**2**	
delayed release	0	1	*2*$^{\clubsuit}$	$^{\clubsuit}$ j = delayed release, $\tau(j)$ = sonority. $c = \min\{1, f(F_1^{\tau(j)}, F_2^{\tau(j)})\} = 1$,
sonority	2	1	**1**	$s_j = 2,\ g_{j,\tau j}(F_1, F_2) = c \cdot 2 + (1-c) f(F_1^j, F_2^j) = 2$
voice	1	-1	**2**	
spread glottis	-1	-1	**0**	
labial	1	1	**0**	$^{\diamondsuit}j$ = labiodental, $\tau(j)$ = labial. $c = \min\{1, f(F_1^{\tau(j)}, F_2^{\tau(j)})\} = 0$,
labiodental	-1	1	*2*$^{\diamondsuit}$	$s_j = 2,\ g_{j,\tau j}(F_1, F_2) = c \cdot 2 + (1-c) f(F_1^j, F_2^j) = f(F_1^j, F_2^j) = 2$
coronal	-1	-1	**0**	
anterior	0	0	*0*	
distributed	0	0	*0*	
lateral	-1	-1	**0**	
dorsal	-1	-1	**0**	
high	0	0	*0*	
front	0	0	*0*	
			total distance: 9	

Table 8: An example of calculating the distance between [z] and [f] with our distance function. The distance between I-features (in bold) is calculated using general distance function $f(x_1, x_2)$, while the distance between D-features (in italic blue) is calculated using $g_{j,k}(F_1, F_2)$. The 'total distance' is the sum of the distances across all dimensions. Details of the calculation are provided in the 'Note' column.

If c_{23} lies between c_{12} and c_{13}, without loss of generality, assume $c_{12} \leqslant c_{23} \leqslant c_{13}$. Then,

$$\begin{aligned}\Delta &\geqslant (c_{13} - c_{23}) \cdot s_j - (c_{13} - c_{23}) f(F_1^j, F_3^j) \\ &\geqslant (c_{13} - c_{23})[s_j - f(F_1^j, F_3^j)] \geqslant 0.\end{aligned}$$

□

Table 8 provides an example of calculating the distance between [z] and [f] using our distance function $d(F_1, F_2)$ defined in Eq. (4). The phonetic feature vectors of [z] and [f] are denoted F_1 and F_2, respectively. The general distance f is set as $f(x_1, x_2) = |x_1 - x_2|$.