Automated Tone Transcription and Clustering with Tone2Vec

Anonymous ACL submission

Abstract

Lexical tones play a crucial role in Sino-Tibetan languages. However, current phonetic fieldwork relies on manual effort, resulting in substantial time and financial costs. This is especially challenging for the numerous endangered languages that are rapidly disappearing, often exacerbated by limited funding. In this paper, we introduce pitch-based similarity representations for tone transcription, named `Tone2Vec`. Experiments on dialect clustering and variance show that `Tone2Vec` effectively captures fine-grained tone variation. Utilizing `Tone2Vec`, we develop the first automatic approach for tone transcription and clustering by presenting a novel representation transformation for transcriptions. Additionally, these algorithms are systematically integrated into an open-sourced and easy-to-use package, `ToneLab`, which facilitates automated fieldwork and cross-regional, cross-lexical analysis for tonal languages. Extensive experiments were conducted to demonstrate the effectiveness of our methods. Experiment implementations are available at https://anonymous.4open.science/r/Tone2Vec-E5D4 [1].

1 Introduction

As the second-largest language family in the world, the Sino-Tibetan languages comprise over 400 languages, nurturing the cultural and communicative bonds of 1.4 billion speakers (Wikipedia). Given the prevalence of lexical tones in most Sino-Tibetan languages (Thurgood and LaPolla, 2003), phonetic fieldwork typically involves conducting tone transcription for each word in the survey lexicon across unexplored regions, followed by categorizing these transcriptions into the respective tone categories of the region. Exploring lexical tones enriches both linguistic and historical research, including migration patterns (LaPolla, 2013), contact between languages (LaPolla, 2010), and their evolution over time (LaPolla FAHA, 2001; LaPolla, 2006; Jacques and Michaud, 2011).

However, existing methodologies face two primary obstacles that hinder further investigation, research, and documentation of Sino-Tibetan languages.

1. **Obstacles in Documenting.** In practice, tone transcription relies on manual effort, and the recorders involved must undergo extensive and prolonged training, which typically lasts several months. Subsequently, the tone categories of a region are discerned based on these transcriptions. The absence of an automatic tone transcription and clustering system leads to substantial time and financial costs, especially for the vast number of endangered languages that are rapidly disappearing (Hale, 1992), often with limited funding.

2. **Obstacles in Analysis.** Although tones can be transcribed using a five-scale system, analyzing tones across different regions is challenging due to the varying lengths (2 or 3 units) of these transcriptions and the differing number of tones in each area. Moreover, extensive fieldwork, represented by the Chinese Language Resources Protection Project, has gathered abundant tone transcription data—exceeding one million records—from thousands of dialect regions within the Sino-Tibetan language family. This has created an urgent need to develop comparable features for different tone transcriptions and to use computational methods to analyze variations across these dialect regions.

In this paper, we systematically addressed the above problems from three angles: feature construction, algorithm design, and the development

[1] This IPYNB file contains all the experimental details presented in this paper. The official package will be released upon acceptance.

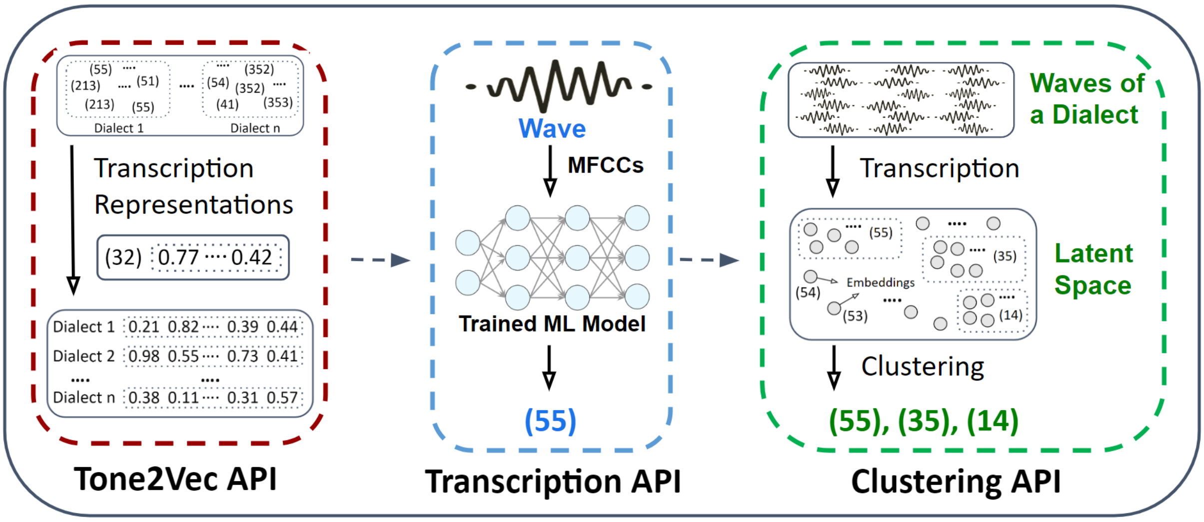


Figure 1: Overview of our proposed methods. From left to right: `Tone2Vec` API for feature construction, Transcription API for automated tone transcription, and Clustering API for clustering tonal data.

of an easy-to-use tool. As illustrated in Figure 1, our contributions can be summarized as follows:

- *Our first contribution* is the proposal of `Tone2Vec`, which maps diverse tone transcriptions to a comparable feature space. `Tone2Vec` constructs pitch-based similarity representations by mapping each transcription to a simulated smooth pitch variation curve. We also propose methods to construct tonal representations for dialect regions. By analyzing these representations across different dialect areas, we show that `Tone2Vec` captures tonal variations and clusters dialects more accurately than methods that treat each tone as an isolated category.

- *As our second contribution*, we developed the first automated algorithms for tone transcription and clustering. These algorithms are especially beneficial for endangered tonal languages. Experiments demonstrate that our models perform well in cross-regional tone transcription with less than 1,500 samples. Notably, our algorithms can accurately cluster tones using fewer than 60 speech samples for a given dialect.

- *As our third contribution*, all these algorithms are systematically integrated into `ToneLab`, a user-friendly platform designed for both lightweight fieldwork and subsequent analysis in Sino-Tibetan Tonal Languages. Users can choose to use pretrained models or train new models with their own data for different scenarios. Researchers can also leverage `ToneLab` to propose new computational methods and conduct evaluations.

2 Related Work

2.1 Representation

The learning process can be viewed as a means of compressing original information to extract effective representations, similar to converting tone signals into concise transcription sequences. The success of machine learning relies on distilling complex entities like words, graphs, and speeches into computable, comparable representations, typically in the form of multi-dimensional vectors, exemplified by notable works like word2vec (Mikolov et al., 2013), graph2vec (Narayanan et al., 2017), and speech2vec (Mikolov et al., 2013). Represented by the GPT series (Radford et al., 2018, 2019; Brown et al., 2020), large language models automatically extract the complex structures and semantic representations of language from vast text corpora. In contrast to treating different tones as atomic units, `Tone2Vec` offers fine-grained tonal representations for tone transcriptions and tone analysis.

2.2 Automated Tone Classification

In recent years, automated tone classification methods (Ryant et al., 2014; Chen et al., 2016; Yuan et al., 2021; Baevski et al., 2020; Yuan et al., 2023) have achieved accuracy rates surpassing those of human listeners, nearing 100% in Standard Mandarin. One approach involves preprocessing the raw signals into features using mel frequency cepstral coefficients (MFCCs), followed by classification prediction using models such as SVM (Ryant et al., 2014), MLP (Ryant et al., 2014), and Convolutional Neural Networks (CNNs) (Chen et al., 2016). Another strategy (Yuan et al., 2021, 2023) leverages more powerful pre-trained models like Wav2Vec 2.0 (Baevski et al., 2020) for fine-tuning. However, tone classification, primarily used in Standard Mandarin, predicts only the categorical information of tones rather than their transcription, making it inapplicable for representing cross-dialect tones.

3 Preliminary

3.1 Lexical Tones

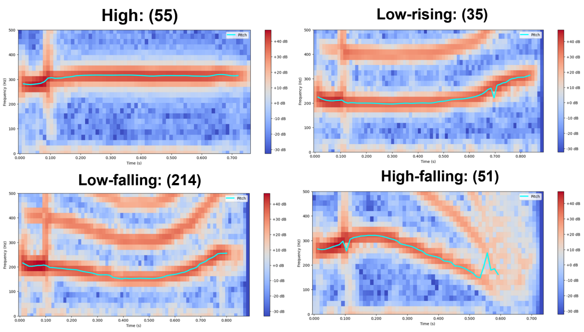


Figure 2: Fundamental frequency (F0, represented with solid lines) and transcription (e.g., (55) indicating a High tone) for the four basic Mandarin tones.

In tonal languages such as Standard Mandarin, lexical meanings are differentiated by pitch variations. These lexical tones are annotated using a scale from 1 (lowest) to 5 (highest), in accordance with Chao's Tone Letter system (Chao, 1930). The four basic lexical tones and pitch variations are visually expounded in Figure 2 by the fundamental frequency, F0.

3.2 Five-scale Marking System

The Five-scale Marking System, developed by Yuen-Ren Chao (Chao, 1930), is the most widely used method for transcribing tones in the Sino-Tibetan language family. In this system, the pitch of a person's speech is divided into five relative levels: (1), (2), (3), (4), and (5), where (1) indicates the lowest pitch and (5) the highest. Tones are then transcribed using sequences of two or three numbers to represent the pitch contour over time. For example, a tone that starts at the mid-level pitch and rises to the high level might be transcribed as (35). The relative changes between these numbers indicate the pitch movement. For example, the tones (53) and (42) both represent a falling pitch, but the first starts at the highest level (5) and ends at a mid-level (3), while the second starts one level lower, beginning at (4) and ending at (2). It is worth noting that transcription represents relative pitch, not absolute pitch. Different speakers may produce the same relative pitch at different absolute levels; for example, one person's lowest pitch might not be the same as another's, but listeners can still identify it as the lowest pitch in their speech (Honorof and Whalen, 2005).

3.3 Tone Classification, Transcription and Clustering Tasks

Let $S(t)$ be a speech signal and $T = < n_1, n_2, \ldots, n_k >$ as the corresponding transcription, where t represents time. We denote a set of speech signals from a dialectal region as $\mathcal{S} = [S_1(t), S_2(t), \ldots, S_m(t)]$, where each $S_i(t)$ represents a speech signal,

Tone Classification Task: Given a dialect area with a certain number of tone categories, for instance, there are M categories, the tone classification task l can be defined as shown in Equation 1.

$$\begin{aligned} l : \mathcal{S} &= [S_1(t), S_2(t), \ldots, S_m(t)] \\ \rightarrow \mathcal{T} &= [t_1, t_2, \ldots, t_m], \\ t_i &= \{1, 2, \ldots, m\} \end{aligned} \tag{1}$$

Tone Transcription Task: Unlike tone classification, the tone Transcription task f takes speech from any dialect as input and outputs a five-scale transcription rather than categories. This process can be defined as shown in Equation 2.

$$\begin{aligned} f : S(t) \rightarrow T &= < n_1, n_2, \ldots, n_k >, \\ n_i &\in \{1, 2, 3, 4, 5\}, \\ k &\in \{2, 3\} \end{aligned} \tag{2}$$

Note that, without any prior knowledge (e.g., speaker's highest/lowest pitch, all tone categories), it is hard to distinguish between a level tone (55) and a level tone (44), or a (41) and a (51) from a single speech signal. However, tones like (523) and (51) can be distinguished due to their different variations. In our subsequent tone evaluation, we will also take this into account, using only the relative pitch as the criterion for assessment.

Tone Clustering Task: The objective of the tone clustering task g is to group these signals into N distinct tonal categories $\mathcal{T} = [T_1, T_2, \ldots, T_N]$, defined as Equation 3, where N is not known and needs model automatic judgment.

$$
\begin{aligned}
&g : \mathcal{S} = [S_1(t), S_2(t), \dots, S_m(t)] \\
&\rightarrow \mathcal{T} = [T_1, T_2, \dots, T_N], \\
&\quad T_i = < n_{i,1}, n_{i,2}, \dots, n_{i,k(i)} >, \\
&n_{i,j} \in \{1, 2, 3, 4, 5\}, \\
&k(i) \in \{2, 3\}
\end{aligned} \tag{3}
$$

4 Data

The majority of publicly available speech data labeled for tones are limited to the four tone categories (T1-T4) in standard Mandarin (Ryu et al., Accessed 1 January 2022; Bu et al., 2017). There is a lack of comprehensive, cross-regional speech data transcribed using the five-scale marking system. To address these limitations, we managed to collect a speech dataset to develop models for automatic tone transcription and clustering, and a second, transcription-only dataset to demonstrate the application of the `ToneLab` tone analysis tool.

Both datasets are in Jianghuai Mandarin. which boasts approximately 70 million speakers and has been extensively studied (Tang, 2023; Zeng, 2018). Jianghuai Mandarin contains many dialect regions that differ from each other in their tonal systems (Chen, 1991; Wang and Sun, 2015; Ho, 2003). With its rich tonal resources, Jianghuai Mandarin serves as a valuable testbed for training and evaluating tone transcription and clustering systems, especially at an early stage where open-source speech with five-scale tone transcription labels is scarce.

Below, we provide a detailed introduction and preprocessing steps for the two datasets.

2238 Recordings from 11 Jianghuai Mandarin Dialects (`Dataset1`): We managed to compile a carefully curated dataset from a previous study(Tang, 2023), which includes 2238 speech recordings across 11 Jianghuai Mandarin dialects. Each speech sample was transcribed by experienced Sino-Tibetan linguists using the five-scale marking system. The dataset categorizes speakers into four groups for each dialect: young males(`YM`), young females(`YF`), older males(`OM`), and older females(`OF`). Tone clusterings are meticulously defined for each group in every region. Each Jianghuai Mandarin dialect is accompanied by detailed descriptions of geographical locations, tone classifications, and dialect regions, all detailed in Appendix A. In subsequent experiments, we randomly selected data from 7 regions for training, 2 regions for validation, and 2 regions for testing, out of a total of 11 regions. The best-performing parameters on the validation set were then used for the final test set evaluation.

Transcriptions with Dialect Cluster Labels (`Dataset2`): In the study of Chinese tones, `Hongchao` and `Huangxiao` clusters of dialect regions in Jianghuai Mandarin are often used to investigate tone evolution, such as the lengthening of entering tones (Tang, 2023), tone sandhi (Wang and Sun, 2015; Coblin, 2005), and tonal inventories (Wang and Sun, 2015). We obtained transcription data from 19 dialect areas in the `Hongchao` cluster and 12 dialect areas in the `Huangxiao` cluster from the Chinese Language Resources Protection Project, which is the largest language resource database in the world. Each dialect area includes 1000 tone transcriptions from the same survey word list, totaling 31,000 transcriptions. Detailed information is provided in Appendix A.

5 Tone2Vec: From Tones to Vectors

In this section, we propose pitch-based similarity representations by quantifying the differences in pitch variations inherent in tones, which we call `Tone2Vec`. `Tone2Vec` is an easy-to-use, simple, and effective method for measuring similarity distance. `Tone2Vec` not only enables the comparison of tonal variations across dialects but also provides a straightforward loss function for training automatic tone transcription and clustering models.

5.1 From Categories to Pitch-based Similarity Representations

In `Tone2Vec`, we map each transcription l, such as `(55)`, to a simulated smooth pitch variation curve $p_l(x)$. As shown in Figure 3, for transcriptions with two units, a linear curve is employed to represent pitch variations, while for those of three units, such as `(312)`, we employ a quadratic curve to smoothly interpolate the points $(1, 3)$, $(2, 1)$, and $(3, 2)$. The divergence between any pair of tone transcriptions, l_1 and l_2, is quantitatively assessed by calculating the area between their pitch variation curves, expressed as $D(l_1, l_2) = \int_{[1,3]} |f_{l_1}(x) - f_{l_2}(x)| dx$. This measure quantifies the differences in pitch variations. Given n transcription sequences $l_1, ..., l_n$, we can construct a $n \times n$ distance matrix $\mathcal{C} = (D(l_i, l_j))_{i,j} \in \mathbb{R}^{n \times n}$, where each row represents the features of a transcription, capturing the subtle pitch variation dif-

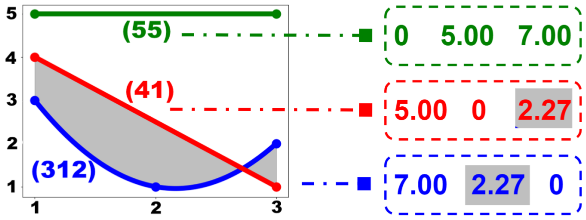


Figure 3: **Left**: Visual simulations using transcription sequences l_1 = (55) (green linear curve), l_2 = (41) (red linear curve), and l_3 = (312) (blue quadratic curve). Grey shading denotes the area between (41) and (312). **Right**: The number 2.27 with grey shading represents the calculated distance between (41) and (312).

ferences among them.

5.2 Case Study: Dialect Clustering and Variance

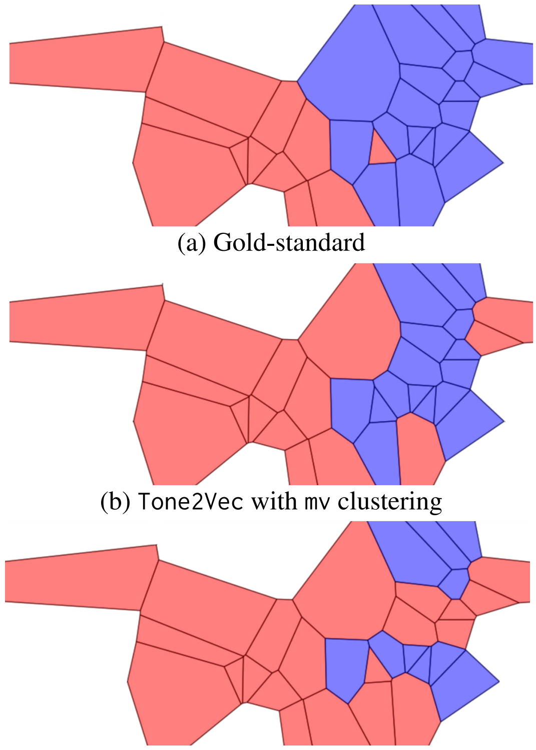
To better introduce and prove the effectiveness of our methods, we conducted experiments on Dialect Group Clustering and Variance using Dataset2. The Dialect Clustering task involves classifying 31 dialect regions, each with 1,000 transcription entries, into two clusters, and the metric accuracy is reported. The task of dialect variance aims to quantify the differences between dialect regions. A good representation should hierarchically reflect dialect variance. We compared Tone2Vec with the baseline model, Baseline. For Baseline, the difference between two transcriptions is 0 if they are identical, and 1 otherwise.

For the dialect clustering task, we calculated the average transcription difference for each pair of dialect areas to derive their tonal features, then performed clustering and evaluated the accuracy of the predicted labels against the true labels. To account for the influence of clustering techniques, we employed seven different methods following the study (Bartelds and Wieling, 2022): single link (sl), complete link (cl), group average (ga), weighted average (wa), unweighted centroid (uc), weighted centroid (wc), and minimum variance (mv) clustering (Heeringa et al., 2012; Prokić and Nerbonne, 2008). The best results are reported in Table 1 and the results of all seven methods are available in Appendix B.

For the dialect variance task, we use multidimensional scaling (MDS) (Torgerson, 1952; Bartelds and Wieling, 2022) to reduce the dimensionality of the dialect representations to 1. The value differences between regions intuitively reflect the variance across different areas and are depicted with varying color intensities in Figure 5.

Method	Accuracy (%)	Clustering
Baseline	70.97	wa
Tone2Vec	<u>83.87</u>	mv

Table 1: Accuracy of Tone2Vec and Baseline method in Dialect Group Clustering with the best clustering method. The underlined value represents the higher accuracy.

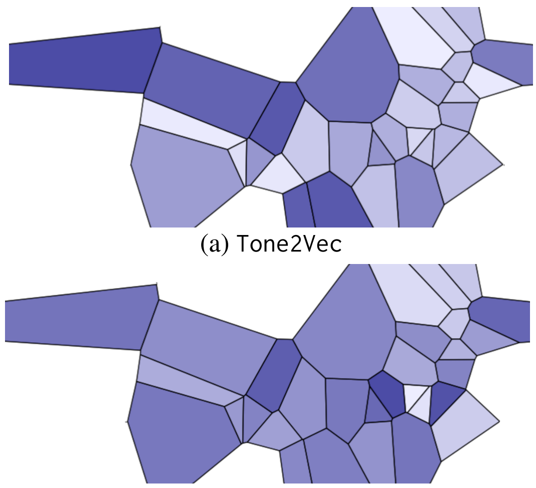


(a) Gold-standard

(b) Tone2Vec with mv clustering

(c) Category with wa clustering

Figure 4: Cluster maps visualizing the Huangxiao and Hongchao dialect clusters. Red represents Huangxiao and blue represents Hongchao.



(a) Tone2Vec

(b) Category

Figure 5: MDS maps visualizing pronunciation differences across dialects. Similar colors indicate similar pronunciations.

Discussion The accuracy results in Table 1 show that Tone2Vec outperforms the Baseline by 12.90%. Additionally, the visualization in Figure 4 indicates that clustering constructed by Tone2Vec is more balanced, whereas the Baseline method tends to classify most dialect areas into a single cluster. Figure 5 demonstrates that Tone2Vec better captures dialect variation, while the baseline method is more influenced by outliers, resulting in most areas having colors within a smaller range.

6 Automatic Tone Transcription

6.1 Pitch-based Loss Function

In contrast to CTC's explicit handling of transcriptions with variable lengths (Graves et al., 2006), our model $\mathcal{M}$ implicitly discerns the length of the transcription sequence during the inference stage. We first fix the model $\mathcal{M}$'s output to consistently produce three float points. For each training instance x_j, the model yields an output $z_j = (z_{j,1}, z_{j,2}, z_{j,3})$, where every z_i falls within the pitch range [1,5]. When viewed through the lens of pitch variations, a sequence of length two—whether a level tone like (55), an ascending tone like (35), or a descending tone like (53)—exhibits a linear relationship among the three predicted components $\mathcal{M}(x_0) = z_0 = (z_{0,1}, z_{0,2}, z_{0,3})$. Sequences of length three, characteristic of contour tones such as (352) or (334), lack this linearity. By establishing a threshold β, we can determine the linearity of a sequence. For speech data x_0, the inferred transcription $\hat{y_0}$ can be formulated as shown in Equation 4:

$$\hat{y_0} = \begin{cases} (\lfloor z_{0,1} \rceil, \lfloor z_{0,3} \rceil) & \text{if} \\ & |z_{0,1} + z_{0,3} - 2 \times z_{0,2}| < \beta \\ (\lfloor z_{0,1} \rceil, \lfloor z_{0,2} \rceil, \lfloor z_{0,3} \rceil) & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\lfloor \rceil$ denotes the operation of rounding to the nearest whole number. In ToneLab, the default value for β is 0.5.

Building on Tone2Vec, we propose a pitch-based loss function, designated $\mathcal{L}_{pitch}$, to automate the transcription of tones and represent signals as tonal representations. By recognizing that each numeral in a transcription sequence, ranging from 1 to 5, symbolizes a different pitch level, and the metric $D(l_1, l_2)$ mirrors the discrepancy between sequences, the metric itself can be directly employed as the loss function for training. For simplicity, we use the MAE loss $\hat{D}(\mathcal{M}(x_j), y_j)$, which approximates $D(\mathcal{M}(x_j), y_j)$ in Equation 5. The relevant properties and motivations of the loss function and evaluations are discussed in Appendix C carefully, demonstrating that the mean absolute error (MAE) loss $\hat{D}(\mathcal{M}(x_j), y_j)$ is essentially based on piecewise linear fitting of pitch variance.

$$\mathcal{L}_{pitch}(\mathcal{X}, \mathcal{Y}) = -\sum_{j=1}^{N} \hat{D}(\mathcal{M}(x_j), y_j) \tag{5}$$

To introduce this concept more intuitively, We denote $\mathcal{M}(x_j)$ as $(z_{j,1}, z_{j,2}, z_{j,3})$. If y_j is a sequence of length three, i.e., $(y_{j,1}, y_{j,2}, y_{j,3})$, then the distance $\hat{D}(\mathcal{M}(x_j), y_j)$ is defined as:

$$\begin{aligned} \hat{D}(\mathcal{M}(x_j), y_j) =& |z_{j,1} - y_{j,1}| + |z_{j,2} - y_{j,2}| \\ &+ |z_{j,3} - y_{j,3}| \end{aligned} \tag{6}$$

If y_j is a sequence of length three, i.e., $(y_{j,1}, y_{j,2}, y_{j,3})$, then the distance $\hat{D}(\mathcal{M}(x_j), y_j)$ is defined as:

$$\begin{aligned} \hat{D}(\mathcal{M}(x_j), y_j) =& |z_{j,1} - y_{j,1}| + |z_{j,3} - y_{j,2}| \\ &+ |z_{j,2} - \frac{1}{2}(y_{j,1} + y_{j,2})| \end{aligned} \tag{7}$$

6.2 Experiments

The experiments were conducted using Dataset1. In the absence of a baseline, we noted that linguists could record tone transcriptions by observing the fundamental frequency (F0) curves (Figure 2), as indicated by (Chen et al., 2016). We use quadratic fitting to regress twenty evenly sampled points from the F0 curve, using the values regressed from the second, middle, and second-to-last points as the predicted tone sequence. We first normalize these values and then use Equation 4 to infer the transcription. Although this method is not a standard automatic tone transcription system (since none currently exists), using F0 curves is a common practice in tone research.

Beyond metric accuracy, we propose a new metric, Variance, to describe the average discrepancy between model predictions and labeled transcriptions by calculating normalized pitch variation. Lower variance indicates better model performance. For a more intuitive presentation, Table 2 shows the Variance values for the transcription (445) compared to six other transcriptions.

Seq.	Variance	Seq	Variance	Seq	Variance
(445)	0.0000	(45)	0.1225	(245)	0.1608
(255)	0.2311	(154)	0.2829	(251)	0.5243

Table 2: `Variance` values for transcription `(445)` compared to `(45)`, `(245)`, `(255)`, `(154)` and `(251)`.

We tested our method on three models: ResNet (He et al., 2015), VGG (Simonyan and Zisserman, 2015), and DenseNet (Huang et al., 2017). Hyperparameters, such as the learning rate, were selected through grid search. Signals were preprocessed using Mel Frequency Cepstral Coefficients (MFCCs) before training the models. Each result is based on three separate experiments, and the averages are reported.

Model	Method	Accuracy (%)	Variance
	F0	10.07	0.2165
ResNet	Tone2Vec	55.99	0.1222
VGG	Tone2Vec	<u>56.08</u>	**0.1052**
DenseNet	Tone2Vec	**61.01**	<u>0.1083</u>

Table 3: Accuracy and `variance` of tone transcription using F0 extraction and `Tone2Vec` on ResNet, VGG, and DenseNet models. Higher accuracy or lower variance indicates better model performance. The bold value represents the best result, and the underlined value represents the second-best result.

Discussion As illustrated in Table 3, our automatic tone transcription method significantly outperforms the F0 extraction-based approach in both Accuracy and `Variance` metrics. Combined with the examples in Table 2, our model maintains consistently high performance across three models, with DenseNet showing the best in Accuracy and the VGG model excelling in `Variance`. These findings collectively indicate that using `Tone2Vec` to train models for automatic tone transcription effectively captures pitch variations.

7 Automatic Tone Clustering

7.1 Clustering on Transcription Features

Many studies (Yuan et al., 2023; Pepino et al., 2021; Zerveas et al., 2021) have shown that well-trained machine learning models not only perform well on targeted tasks but also provide hierarchical embeddings. Therefore, by extracting intermediate layer features, the automatic tone transcription model $\mathcal{M}$, has already assigned tonal representations for each speech instance. Hence, the task of Tone Clustering can be regarded as a clustering task on transcription features. We then employ the clustering algorithm DBSCAN (Ester et al., 1996) on these representations to determine the number of tone categories automatically, selecting the most probable predicted label in each cluster as a tone category.

7.2 Experiments

The experiments were conducted using `Dataset1`. We still use the 7:2:2 data split strategy for training and model selection, following the transcription experiments in Subsection 6.2. Each region has at most four clusterings from four speakers: young males (`YM`), young females (`YF`), older males (`OM`), and older females (`OF`). Each speaker's speech, consisting of fewer than 60 samples per dialect, is manually labeled for tone categories. We select the best-performing model, `DenseNet`, for tone transcription tasks. Tonal embeddings are visualized using UMap (McInnes et al., 2020), with DBSCAN parameters `eps` set to 0.6 and `min_samples` set to 4.

SPK	Type	Tone 1	Tone 2	Tone 3	Tone 4
OF	Lab.	(213)	(24)	(41)	(53)
	Pred.	(313)	(45)	(51)	(42)
YF	Lab.	(212)	(24)	(51)	(55)
	Pred.	(213)	(34)	(52)	(44)
OM	Lab.	(213)	(24)	(41)	(51)
	Pred.	(212)	(34)	(31)	(32)

Table 4: Comparison of manually labelled (Lab.) and automatically predicted (Pred.) tone categories for young females (`YF`), older males (`OM`), and older females (`OF`) in the `Wuhu` dialect area. Pred values indicate the transcriptions, with each non-dash value representing a predicted category.

SPK	Type	Tone 1	Tone 2	Tone 3	Tone 4
YM	Lab.	(41)	(24)	(31)	(55)
	Pred.	-	(24)	(32)	(44)
OF	(Lab.)	(41)	(24)	(31)	(55)
	Pred.	(41)	-	(212)	(45)
YF	Lab.	(51)	(24)	(32)	(55)
	Pred.	(51)	(24)	(43)	(33)
OM	Lab.	(41)	(24)	(32)	(55)
	Pred.	(52)	(23)	(31)	(44)

Table 5: Comparison of manually labelled (Lab.) and automatically predicted (Pred.) tone categories for young males (`YM`), young females (`YF`), older males (`OM`), and older females (`OF`) in the `Yangzhou` dialect area. Pred values indicate the transcriptions, with each non-dash value representing a predicted category.

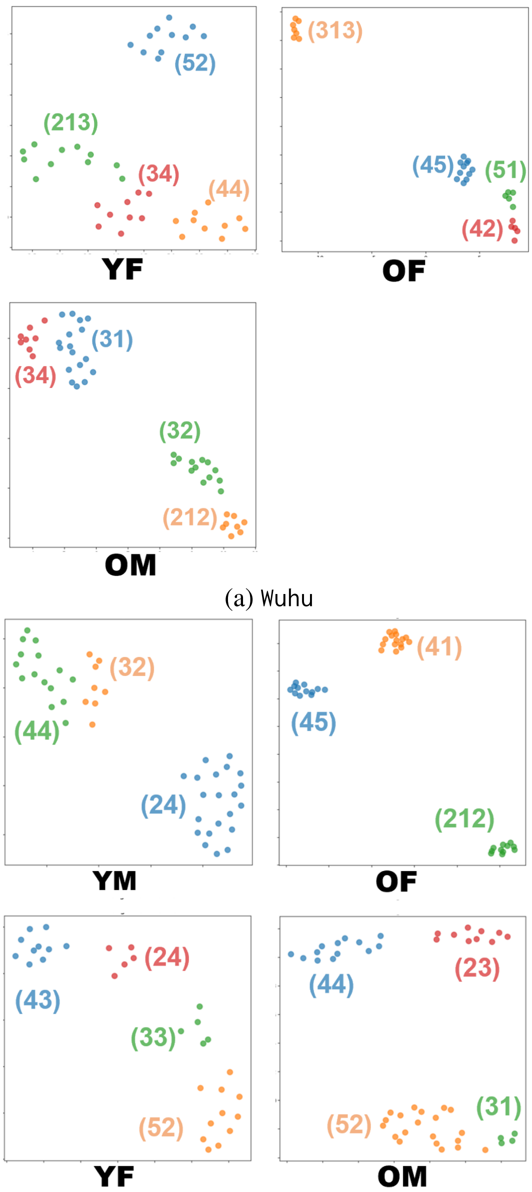


Figure 6: Visualization of automatic clustering for young females (`YF`), older males (`OM`), and older females (`OF`) in the `wuhu` dialect areas and young females (`YF`), older males (`OM`), and older females (`OF`) in the `wuhu` dialect using UMAP for dimensionality reduction and DBSCAN for clustering.

Discussion As illustrated in Table 4 and Table 5, our model accurately determined the number of tone categories with 71% accuracy. Additionally, the model generally predicted rising tones as rising, falling tones as falling, and contour tones as contour. Differences between predictions and ground truth mainly stemmed from variations in pitch magnitude, such as predicting `(212)` as `(213)`. Overall, these differences are within an acceptable margin. Notably, tone categorization varies among different individuals. Simultaneously, as depicted in Figure 6, tonal features show clear clustering. The proximity of `(52)` to `(31)` rather than to `(23)` reflects inner similarities among different tones.

8 ToneLab: A User-friendly Platform for Tones

We have developed an easy-to-use package, `ToneLab`. We aim for `ToneLab` to be a user-friendly platform for documenting and studying tones. To sum up, two main modules are introduced.

8.1 `ToneLab.Document`: Automatic Tone Documentation Solutions

This module supports tone transcription, tone clustering, and lightweight tone classification for studying tonal languages. The lightweight tone classification function requires a predefined transcription list of all categories. During inference, we use transcription models to predict and find the closest category within the list, reducing the need for retraining a new classification model.

Input: MFCCs extracted from speech, either one (for transcription and lightweight classification) or multiple (for clustering).

Models: MLP and CNN models, including ResNet (He et al., 2015), VGG (Simonyan and Zisserman, 2015), and DenseNet (Huang et al., 2017). Users can use the provided models or train their own models with their own data.

8.2 `Tonelab.Analysis`: Large Scale and Cross Dialect Tone Analysis

In `ToneLab.Analysis`, representations can be easily queried from the pre-computed database for any tone transcriptions. `ToneLab.Analysis` supports inputting a set of transcriptions from a dialect region and returns the comparable tonal features of that region, which can be used to study dialect clustering and variance. Our package also supports investigating the influence of initials and finals on tones using methods such as the improved Levenshtein distance (Wieling et al., 2012).

9 Conclusion

In this paper, we proposed Automated Tone Transcription and Clustering with `Tone2Vec`. We hope our work could raise awareness about the importance and urgency of preserving and studying endangered Sino-Tibetan tonal languages, which have long been overlooked, and encourages more collaborative efforts in this crucial field.

10 Limitations

As a paper focused on computational social science, we discuss the limitations, potential improvements, and future directions from both social science and computational perspectives below.

10.1 Tone Transcription Systems

In Sino-Tibetan tonal languages, the Five-Scale Marking System provides a consistent way to transcribe tones by establishing five relative pitch levels, which is the most system. As a result, developing algorithms based on this system is both urgent and practical for broad use.

However, several limitations exist in this system . Firstly, the Five-Scale Marking System assumes human pitch can be divided into five relative levels, which isn't always accurate. For instance, the Analco Chinantec language has at least six pitch levels, while four levels suffice for Standard Mandarin. Secondly, the Five-Scale Marking System doesn't specify the proportion of each pitch contour's duration within a tone. For example, a tone transcribed as `(312)` indicates a pitch that falls and then rises, but the duration of the fall and rise can vary. Alternative systems like the Four-Domain Marking System, Nine-Scale Marking System, and Contour Tone Marking System have been proposed to address these issues. Lastly, special tones, such as checked tones, require additional markings. These considerations indicate that finding an optimal tone representation remains a significant challenge.

In this paper, we found that embeddings extracted from the intermediate layers of trained transcription models effectively reflect tonal representations in clustering experiments, suggesting a promising direction. However, these embeddings are typically high-dimensional, floating-point, and computationally based. How to establish a more detailed connection with existing phonological theories needs further consideration and acceptance.

10.2 Limited Open-sourced Data

For tasks more complex than tone classification, our models are currently built using only a few thousand labeled speech data points, whereas the tone classification dataset contains hundreds of thousands of labeled syllables. We hope that more open-sourced data will be made available in the future to facilitate the construction of higher performance and more user-friendly benchmarks. A feasible and cost-effective approach would be to release speech segments along with some corresponding transcriptions without requiring precise alignment. Many algorithms (Moritz et al., 2021; Wigington et al., 2019; Miao et al., 2015; Laptev et al., 2021; Cai et al., 2021; Pratap et al., 2022; Xiang and Ou, 2019; Huang et al., 2016) have been proposed to address the issue of misalignment.

10.3 From Single Syllables to Continuous Speech

The speech supported by `ToneLab` is currently based solely on single syllables, primarily because we only labeled single-syllable speech data. Additionally, tone transcription and clustering in practical applications are mostly based on single syllables. However, continuous speech contains rich tonal phenomena such as tone sandhi (Chen et al., 2016; Shen, 1990) and tone coarticulation (Yuan et al., 2023). Therefore, developing transcription and analysis methods for continuous speech is important.

One feasible approach is to improve existing Connectionist Temporal Classification (CTC) methods. CTC (Graves et al., 2006) stands as a pivotal and widely recognized loss function designed for handling sequences without aligned input and target labels, such as in Automatic Speech Recognition (ASR) (Amodei et al., 2016) and Optical Character Recognition (OCR) (Liu et al., 2015). Nonetheless, applying CTC methods directly to tonal transcriptions proves to be inappropriate due to potential problems including data scarcity, the inherent similarities between tones, and the noise introduced by manual transcription. However, adapting CTC concepts for tone transcription presents a promising direction for future research, though it requires more experiments and data support.

10.4 Potential Risk

When using `ToneLab` to train models, it's important to ensure data privacy and security to avoid unauthorized access. Large volumes of speech data can be leaked, potentially exposing participants' speech characteristics and violating their privacy.

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Martijn Bartelds and Martijn Wieling. 2022. Quantifying language variation acoustically with few resources. *arXiv preprint arXiv:2205.02694*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pages 1–5. IEEE.

Xingyu Cai, Jiahong Yuan, Yuchen Bian, Guangxu Xun, Jiaji Huang, and Kenneth Church. 2021. W-ctc: a connectionist temporal classification loss with wild cards. In *International Conference on Learning Representations*.

Yuen-Ren Chao. 1930. A system of tone letters. *Le maître phonétique*.

Charles Chen, Razvan C Bunescu, Li Xu, and Chang Liu. 2016. Tone classification in mandarin chinese using convolutional neural networks. In *Interspeech*, pages 2150–2154.

Matthew Y Chen. 1991. An overview of tone sandhi phenomena across chinese dialects. *Journal of Chinese Linguistics Monograph Series*, (3):111–156.

W South Coblin. 2005. *Comparative phonology of the Huáng-Xiào dialects*. Institute of Linguistics, Academia Sinica.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 369–376. ACM.

Ken Hale. 1992. Endangered languages: On endangered languages and the safeguarding of diversity. *language*, 68(1):1–42.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Wilbert Heeringa, John Nerbonne, and Peter Kleiweg. 2012. Validating dialect comparison methods. *Journal of Classification*.

Dah-an Ho. 2003. The characteristics of mandarin dialects. *The Sino-Tibetan languages*, pages 126–130.

Douglas Honorof and Douglas Whalen. 2005. Perception of pitch location within a speakerś f0 range. *J. Acoust. Soc. Am.*, 117:2193–2200.

De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. 2016. Connectionist temporal modeling for weakly supervised action labeling. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 137–153. Springer.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Guillaume Jacques and Alexis Michaud. 2011. Approaching the historical phonology of three highly eroded sino-tibetan languages: Naxi, na and laze. *Diachronica*, 28(4):468–498.

Randy J LaPolla. 2006. Development of the sino-tibetan language family. *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*, page 225.

Randy J LaPolla. 2010. Language contact and language change in the history of the sinitic languages. *Procedia-Social and Behavioral Sciences*, 2(5):6858–6868.

Randy J LaPolla. 2013. Eastern asia: Sino-tibetan linguistic history. *The global prehistory of human migration*, pages 204–208.

Randy LaPolla FAHA. 2001. The role of migration and language contact in the development of the sino-tibetan language family.

Aleksandr Laptev, Somshubra Majumdar, and Boris Ginsburg. 2021. Ctc variations through new wfst topologies. *arXiv preprint arXiv:2110.03098*.

Qi Liu, Lijuan Wang, and Qiang Huo. 2015. A study on effects of implicit and explicit language model information for dblstm-ctc based handwriting recognition. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 461–465. IEEE.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Yajie Miao, Mohammad Gowayyed, and Florian Metze. 2015. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. In *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)*, pages 167–174. IEEE.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Niko Moritz, Takaaki Hori, and Jonathan Le Roux. 2021. Semi-supervised speech recognition via graph-based temporal classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552. IEEE.

Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. 2017. graph2vec: Learning distributed representations of graphs. *CoRR*, abs/1707.05005.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

Vineel Pratap, Awni Hannun, Gabriel Synnaeve, and Ronan Collobert. 2022. Star temporal classification: Sequence modeling with partially labeled data. *Advances in Neural Information Processing Systems*, 35:13392–13403.

Jelena Prokić and John Nerbonne. 2008. Recognising groups among dialects. *International journal of humanities and arts computing*, 2(1-2):153–172.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Neville Ryant, Jiahong Yuan, and Mark Liberman. 2014. Mandarin tone classification without pitch tracking. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4868–4872.

Catherine Ryu, Mandarin Tone Perception & Production Team, and Michigan State University Libraries. Accessed 1 January 2022. Tone perfect: Multimodal database for mandarin chinese. https://tone.lib.msu.edu/. Accessed 1 January 2022.

Xiaonan Susan Shen. 1990. Tonal coarticulation in mandarin. *Journal of Phonetics*, 18(2):281–295.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

Zhiqiang Tang. 2023. *Entering Tone in Jiang-Huai Mandarin: Experimental Research on Acoustic-Physiological-Perceptual Aspects (PhD thesis)*. Nanjing Normal University, Nanjing.

G. Thurgood and R.J. LaPolla. 2003. *The Sino-Tibetan Languages*. Routledge language family series. Routledge.

Warren S Torgerson. 1952. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

William SY Wang and Chaofen Sun. 2015. *The Oxford handbook of Chinese linguistics*. Oxford University Press.

Martijn Wieling, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics*, 40(2):307–314.

Curtis Wigington, Brian Price, and Scott Cohen. 2019. Multi-label connectionist temporal classification. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 979–986. IEEE.

Hongyu Xiang and Zhijian Ou. 2019. Crf-based single-stage acoustic modeling with ctc topology. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5676–5680. IEEE.

Jiahong Yuan, Xingyu Cai, and Kenneth Church. 2023. Improved contextualized speech representations for tonal analysis. In *Proceedings of Interspeech 2023*, pages 4513–4517.

Jiahong Yuan, Neville Ryant, Xingyu Cai, Kenneth Church, and Mark Liberman. 2021. Automatic recognition of suprasegmentals in speech. *Preprint*, arXiv:2108.01122.

Xiaoyu Zeng. 2018. A case study of dialect contact of early mandarin. *Lingua*, 208:31–43.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124.

A Detailed Dialect Information

Table 6 provides detailed information on the province, city, cluster, sub-slices, East Longitude, and North Latitude for the 31 dialect regions. The positions of the dialect regions in Figure 4 and Figure 5 are determined by their actual East Longitude and North Latitude.

B Full Results of the Dialect Group Clustering

Table 7 presents the results of seven clustering algorithms—single link (`sl`), complete link (`cl`), group average (`ga`), weighted average (`wa`), unweighted centroid (`uc`), weighted centroid (`wc`), and minimum variance (`mv`)—applied to the `Tone2Vec` and `Baseline` methods.

Discussion Table 7 indicates that the choice of clustering algorithm significantly affects accuracy, with a difference of 22.58% between the best and worst clustering algorithms for `Tone2Vec` and 12.91% for `Baseline`. Among the seven clustering algorithms, `Tone2Vec` outperformed `Baseline` in five methods, while `Baseline` outperformed in two. Considering the influence of different clustering algorithms, these results demonstrate that `Tone2Vec` provides better tone representations than `Baseline`, especially with the highest accuracy of 83.87%, which is significantly higher than the best performance of `Baseline` at 70.97%.

C Discussion of Tone2Vec and Automatic Models

C.1 The Design of Proposed Tone2Vec

Transforming tone transcriptions l into representations $g(l) \in \mathbb{R}^D$ can be regarded as mapping each element of the set $\mathcal{S}$ to a metric space $(g(\mathcal{S}), d) = (g(l_1), ..., g(l_n), d) \subset (\mathbb{R}^D, d)$. This mapping process quantifies the dissimilarity between tonal transcriptions l_i and l_j through the metric $d(g(l_i), g(l_j))$. Given the challenges associated with direct selection of the mapping, we advocate for the construction of a similarity mapping $D(*, l_j)$ for $j = 1, 2, ..., n$, to effectively discern transcription similarities and establish a basis in the space. This can be rigorously defined in Definition 1.

Definition 1. *Let $g : \mathcal{S} \to V$ be a mapping from the set of tone transcriptions $\mathcal{S}$ into a metric space V equipped with metric d. Suppose there exists a metric $D : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ such that for any $s_i, s_j \in \mathcal{S}$, $D(s_i, s_j) = d(g(s_i), g(s_j))$. Define a mapping $\hat{g} : \mathcal{S} \to \mathbb{R}^n$ by*

$$\hat{g}(s_i) = (D(s_i, s_1), D(s_i, s_2), \ldots, D(s_i, s_n)) \tag{8}$$

where $\hat{g}(s_i) \in \mathbb{R}^n$ and $n = |\mathcal{S}|$. Then, the l_1-norm distance between $\hat{g}(s_i)$ and $\hat{g}(s_j)$ is given by $|\hat{g}(s_i) - \hat{g}(s_j)|_1 = 2 \cdot |d(g(s_i) - g(s_j))| + \sum_{k \neq i,j} |d(g(s_i), g(s_k)) - d(g(s_j), g(s_k))|$

The selection of metric D centers on capturing the nuances of pitch variations inherent in tones. In this paper, we map each transcription l to a simulated smooth pitch variation curve $f_l(x)$.

C.2 Approximated MAE Loss

Here, we map each transcription l to a continuous variation curve $\hat{p}_l(x)$ instead of the simulated smooth pitch variation curve $p_l(x)$. For transcriptions with two units, the same linear curve as `Tone2Vec` is employed to represent pitch variations, while for those with three units, such as `(312)`, we use a piecewise linear function curve, which is two connected linear segments, to interpolate the points $(1, 3)$, $(2, 1)$, and $(3, 2)$. The divergence between any pair of tone transcriptions can be quantitatively assessed by calculating the area between their pitch variation curves. The corresponding result is the loss function $\hat{D}(\mathcal{M}(x_j), y_j)$ used in our training process. This loss function, based on piecewise linear fitting of pitch variance, is simpler compared to `Tone2Vec`'s calculation (requiring only the difference at corresponding positions).

C.3 Evaluation of Automatic Tone Transcription: The Variance Metric

In the evaluation of tone transcription, to eliminate the unpredictability of absolute pitch in individual speech, we use relative pitch as the evaluation criterion. Thus, the evaluation metric `Variance` has been proposed.

First, we normalize any transcription l within the range [0, 1], denoted as $f_1(l)$. Specifically, we map the highest pitch value to 1, the lowest to 0, and evenly distribute the intermediate values. The examples below illustrate our process:

- Transcription `(412)`:

$$\text{max: } 4, \text{ min: } 1 \to \left(\frac{4-1}{4-1}, \frac{1-1}{4-1}, \frac{2-1}{4-1}\right) = (1, 0, 0.333)$$

Point	Province	City	Cluster	Sub-slices	East Longitude (°E)	North Latitude (°N)
1	Jiangxi	Jiujiang	Huangxiao	-	115.408	29.617
2	Jiangxi	Jiujiang	Huangxiao	-	116.012	29.735
3	Anhui	Tongling	Huangxiao	-	117.442	30.883
4	Anhui	Anqing	Huangxiao	-	117.020	30.300
5	Shaanxi	Shangluo	Huangxiao	-	109.160	33.429
6	Hubei	Huanggang	Huangxiao	Luotian	115.433	30.925
7	Hubei	Xiaogan	Huangxiao	Xiaogan	113.533	30.925
8	Hubei	Xiaogan	Huangxiao	Yunmeng	113.759	31.027
9	Hubei	Xiaogan	Huangxiao	Xiaogan	113.817	31.733
10	Hubei	Huanggang	Huangxiao	E'dong	114.581	31.303
11	Hubei	Huanggang	Huangxiao	-	115.917	30.008
12	Hubei	Xiaogan	Huangxiao	-	113.633	31.275
13	Anhui	Chuzhou	Hongchao	Yangzhou	118.933	32.700
14	Anhui	Chuzhou	Hongchao	-	118.312	32.301
15	Anhui	Wuhu	Hongchao	-	118.408	31.258
16	Anhui	Chizhou	Hongchao	Rongjiu	118.208	30.575
17	Anhui	Xuancheng	Hongchao	-	119.350	30.908
18	Anhui	Wuwei	Hongchao	-	117.908	31.217
19	Anhui	Chizhou	Hongchao	-	117.467	30.525
20	Anhui	Anqing	Hongchao	Anqing	116.908	30.958
21	Anhui	Huainan	Hongchao	-	116.975	32.608
22	Anhui	Xuancheng	Hongchao	-	119.117	31.133
23	Anhui	Wuhu	Hongchao	-	118.508	31.175
24	Anhui	Lu'an	Hongchao	Hongchao	116.633	31.675
25	Jiangsu	Yancheng	Hongchao	-	120.205	33.396
26	Jiangsu	Zhenjiang	Hongchao	-	119.430	32.195
27	Jiangsu	Nanjing	Hongchao	-	118.460	32.020
28	Jiangsu	Yangzhou	Hongchao	-	119.421	33.231
29	Jiangsu	Yangzhou	Hongchao	-	119.430	32.380
30	Jiangsu	Huai'an	Hongchao	-	119.375	33.883
31	Jiangsu	Huai'an	Hongchao	-	119.032	33.559

Table 6: Detailed Dialect Information from Hongchao and Huangxiao Clusters.

- Transcription (25):

$$\text{max: } 5, \text{ min: } 2 \rightarrow \left(\frac{2-1}{5-2}, \frac{5-1}{5-2}\right) = (0, 1)$$

For any two transcriptions, l_1 and l_2, we obtain their relative pitches $f_1(l_1)$ and $f_1(l_2)$. We use $\hat{D}(\sigma(f_1(l_1)), \sigma(f_1(l_2)))$ to measure the difference in relative pitch, resulting in the `Variance` metric, where σ is the sigmoid function.

Method	sl	cl	ga	wa	uc	wc	mv
Tone2Vec	64.52	70.97	70.97	64.52	70.97	61.29	**83.87**
Baseline	58.06	67.74	67.74	**70.97**	61.29	67.74	61.29

Table 7: Accuracy of Tone2Vec and Baseline methods with all seven clustering algorithms in Dialect Group Clustering. The underlined values represent the higher accuracy for each clustering algorithm. Bold numbers represent the best performance for each method.