Applying functional partition in the investigation of lexical tonal-pattern categories in an under-resourced Chinese dialect *

Junru Wu[1,2]，Yiya Chen[2]，Vincent J van Heuven[2,3]，Niels O Schiller[2]
(1. East China Normal University, Dept. of Chinese Language and Literature, Lab of Language Cognition and Evolution, Shanghai, 200241； 2. Leiden University Centre for Linguistics, Leiden, 2300 RA; 3. Dept. Hungarian and Applied Linguistics, University of Pannonia, Veszprém)

Abstract: The present study applied functional partition to investigate disyllabic lexical tonal-pattern categories in an under-resourced Chinese dialect, Jinan Mandarin. A Two-Stage partitioning procedure was introduced to process a multi-speaker corpus that contains irregular lexical variants in a semi-automatic way. In the first stage, a program provides suggestions for the phonetician to decide the lexical tonal variants for the recordings of each word, based on the result of a functional k-means partitioning algorithm and tonal information from an available pronunciation dictionary of a related Chinese dialect, i.e. Standard Chinese. The second stage iterates a functional version of k-means partitioning with Silhouette-based criteria to abstract an optimal number of tonal patterns from the whole corpus, which also allows the phoneticians to adjust the results of the automatic procedure in a controlled way and so redo partitioning for a subset of clusters. The procedure yielded eleven disyllabic tonal patterns for Jinan Mandarin, representing the tonal system used by contemporary Jinan Mandarin speakers from a wide range of age groups. The procedure used in this paper is different from previous linguistic descriptions, which were based on more elderly speakers' pronunciations. This method incorporates phoneticians' linguistic knowledge and preliminary linguistic resources into the procedure of partitioning. It can improve the efficiency and objectivity in the investigation of lexical tonal-pattern categories when building pronunciation dictionaries for under-resourced languages.

Keywords：pattern recognition; phonetics; tone; pronunciation dictionary; k-means partition

Pronunciation dictionaries are usually expensive to build, especially for under-resourced languages and dialects [1]. Sometimes, linguistic descriptions and dictionaries are available. However, these resources usually only cover the canonical or stable lexical variants used by elderly speakers, while under-resourced languages and dialects usually have rich lexical variants, due to the lack of standardization.

As for tonal dialects of Mandarin Chinese, many of which are widely used but not standardized, lexical variants usually come with different tonal-patterns. For instance, as shown in Figure 1, the word 'simple' allows for two different tonal variants in Jinan Mandarin (JM), while the word 'very' allows for only one [2].

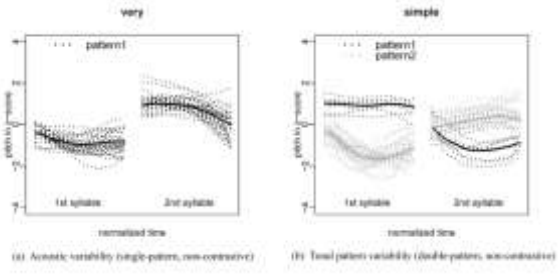

Figure 1. Pitch contour distributions from a mono-pattern word (left) and a dual-pattern word (right) [2].

To further model such dialects, whether for linguistic or engineering purposes, the following questions need to be answered: Which tonal variant(s) does a given word have? Which tonal patterns does the language system have?

These questions are basic. The results can be used in building linguistic theories or baseline dictionaries,

***基金项目**：晨光计划Chenguang Program
作者简介：吴君如 Junru Wu （1985），女 （汉），连江，讲师
jrwu@zhwx.ecnu.edu.cn

which can then be used for the evaluation of NLP models. However, to achieve answers to these questions, laborious manual labeling is required and the results suffer from subjectivity and human errors. If we can introduce some automaticity into the procedure, the workload can be reduced and the accuracy can be improved. Based on the above consideration, a Two-Stage semi-automatic partitioning procedure is proposed in this paper.

1 Two-Stage Semi-Automatic Partition

We propose a Two-Stage semi-automatic partitioning procedure to retrieve the word-wise tonal variant(s) and the basic tonal patterns from a multi-speaker disyllabic corpus.

The core algorithm of this Two-Stage Semi-Automatic Partition is functional k-means partition [3], which partitions the observed curves into a given number (k) of clusters. K-means partition is chosen over the other types of partitioning methods for the following reason: the centroid-based nature of k-means partition fits the nature of phoneme perception. Psycholinguists found that there are "prototypes" in phonological categories, and it is more difficult to discriminate sounds that are closer to the prototypes in acoustic distribution than those that are closer to non-prototypes [4, 5]. K-means partition also assumes "prototypes" within each cluster, and the adscription of items depends on their distance from the closest prototypes [3]. Compared with the assumptions of other approaches - such as the dichotic hierarchy assumed by the hierarchical clustering, the within-cluster normalcy assumed by the distribution-based clustering, and the sparse areas assumed by density-base clustering [6] - the prototypes assumed by k-means partition are more reasonable.

In the current proposal, a functional version of k-means partition is used, which means every pitch contour is treated as one curve, and the algorithm partitions the curves into a given number of clusters [7]. Depending on the stage of investigation, the number of clusters is either given to the model directly or selected from a range based on Silhouette width [8, 9]. The partitioning is performed in two stages, yielding lexical tonal variants and general tonal patterns, respectively.

In the first stage, a phonetician utilizes the program to decide the lexical tonal variants for each word. The word-wise procedure is as follows: 1) plotting all the normalized pitch contours for this word; 2) dividing the curves into a chosen number of clusters; 3) the phonetician typing in a label for each cluster; 4) the phonetician verifying the label of each curve (optional). In this process, the phonetician can choose to see referential labels from a related and more resourceful dialect or a historical system. This stage yields tonal classifications and variant probabilities for each word. It can also extract a preliminary and subjective classification of tonal patterns according to the labels given by the phonetician.

The second stage then chooses an optimal partitioning solution of tonal patterns for the tonal system derived from the lexical tonal variants. Different from the preliminary classification decided by the phonetician, whether two lexical tonal variants belong to the same tonal pattern is decided automatically in this stage by the program, which takes the distribution of all variants into consideration. The results from the previous word-wise stage are fed into the model in the second stage. The procedure is as follows: 1) automatically calculating one prototypical curve for each lexical tonal variant using a depth-based criterion [7, 10], which yields a collection of prototypical curves; 2) excluding the lexical tonal variants with extremely small probabilities, which may in fact be production errors (optional); 3) calculating one preliminary prototype for each cluster, based on a provided preliminary classification; 4) using the preliminary prototypes as the initial center curves to calculate k-means partitions for the prototypical curves; 5) removing the center of the least distinguishable cluster (the cluster with the smallest Silhouette width [8]) and redoing the k-means partition; 6) iterating step 5 until there are only two clusters left, and keeping a record of all the solutions generated in steps 4 and 5; 7) calculating the mean and standard deviation (SD) of the Silhouette values for each partition, subtracting the SD from the mean as the goodness value of the solution, and choosing the solution with the highest goodness value as the optimal partitioning solution.

Since the optimal partitioning solution in this stage is only the best that k-means partition can achieve, there is still space for improvement. One potential problem of k-means partition is that the clusters are expected to be of similar sizes [3]. The real tonal

system can involve closely overlapping tonal patterns, which can be distinguished from other tonal patterns. However, with k-means partition such overlapping tonal patterns would be put in the same cluster within the optimal partitioning solution.

To improve the partition, an additional procedure is introduced, which rearranges a subset of the clusters while keeping the rest of the clusters the same as it was in the given partition. The phonetician, after viewing the plots of the given partition, picks out two clusters that need to be rearranged together, and the number of clusters is designated by the phonetician. The new clusters then replace the original two clusters in the given partition, yielding an adjusted partition. This procedure can start from the optimal solution and be repeated until the adjusted partitioning solution fits the intuition of the phonetician.

2 Experiment

The Two-Stage Semi-automatic Partition is tested with a small multi-speaker corpus of Jinan Mandarin (JM) disyllabic words.

2.1 Corpus Preparation

Forty-two JM native speakers read 400 disyllabic Chinese words in JM. The written words were selected from a corpus of Chinese film subtitles [11], including a list of 200 high-frequency words and a list of 200 low-frequency words. Tonal combinations reported in a published linguistic dictionary for JM are represented as evenly as possible in this corpus [12]. The list was presented in a different randomized order for each JM speakers in a self-paced way.

Praat [13] is used to extract pitch contours from the rhymes. A trained phonetician manually marked the rhymes. Also, in this process, recordings with speech and recording errors were excluded. The pitch contours were converted to semitones with 100Hz as the base and then transformed into z-scores based on the speakers' means and standard deviations [14, 15]. The normalized pitch contours were then interpolated to 20 points per-syllable to remove the difference in duration. A density-based local approach was adopted to eliminate the possible outliers [16]. Local Outlier Factors (LOF) were calculated for each speaker's pitch contours. Any pitch contours with a LOF greater than 1.5 [16] and belonging to the 2.5% with the highest integral density were eliminated from the corpus.

2.2 Word-Wise Partitioning and Verification

In the first stage, word-wise partitioning and verification is performed using the kmeans.fd function of the fda.usc package [7] in R [17] to look for the lexical tonal variants for each word.

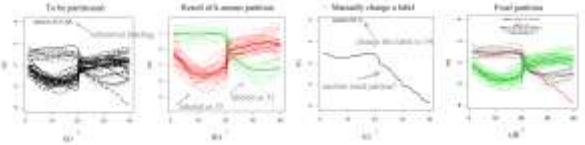

Figure 2. (a) all the pitch contour curves for the word "simple", (b) the result of k-means partition, (c) the curve whose label was changed, (d) final partitioning solution for this word.

Here the procedure for the word "simple" is demonstrated as an example. The pitch contours of all the exemplars of this word were first plotted, as shown in Figure 2a, in which the tonal categories of Standard Chinese (SC) were displayed for reference. With the number of clusters (number of lexical tonal variants) designated as two, k-means partition provided the optimal partitioning solution, as shown in Figure 2b. According to the referential labeling and the tone sandhi rules described by Qian et al.[12], the first cluster was labeled as "35" and the second cluster was labeled as "31". Then we verified the label of each curve and found that the one produced by Speaker 06 probably belongs to another tonal pattern (with a falling contour in the second syllable, as shown in Figure 2c), and so we assigned a different label "34" to this curve. The final partitioning solution for "simple" is shown in Figure 2(d).

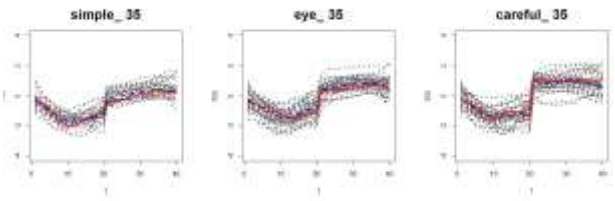

Figure 3. Pitch contours for the lexical tonal variants "simple_35", "eye_35", and "careful_35".

Note that, in this step, the phonetician's labeling assumed a preliminary classification. For instance, the lexical variant "simple_35", "eye_35", and "careful_35" were all labeled with "35" as shown in Figure 3, which means the phonetician assumed that these variants carry the same tonal pattern. This is the preliminary classification (largely subjective, so not an objective partition).

2. 3 Partitioning for Basic Tonal Patterns

2.3.1 Calculating Prototypical Pitch Contours for Lexical Variants

One prototypical pitch contour was calculated for each lexical tonal variant in this step, using the depth.mode function from the fda.usc R package [7]. There are two ways to decide the prototypical curve, choosing the deepest curve (as a real prototype) [7] or calculating a trimmed mean curve (as an abstract prototype)[10], as shown with an example in Figure 3. In the present experiment, the collection of abstract prototypes was used in the analysis.

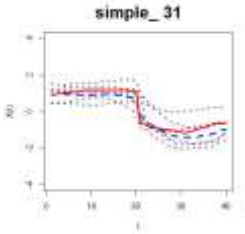

Figure 4. All the curves for the lexical tonal variant "simple_31" (grey dotted curves), the real prototype (red solid curve), and the abstract prototype (blue dashed curve).

2.3.2 Optimizing the General Partitioning solution

In this step, each lexical tonal variant was represented with one prototypical curve. The same collection of these prototypical curves was then partitioned with different parameters according to the following procedure.

The first round of partitioning was fitted with given initial centers [7]. In the experiment, these initial centers were calculated as follows. As mentioned in 3.2, the prototypical curve for each lexical tonal variant labeled with the same tonal pattern was assumed to belong to the same tonal pattern. Here the deepest prototypical curve for each tonal pattern assumed by the phonetician was calculated. The collection of these prototypical curves was taken as the initial centers for the first round of partitioning [7]. The first solution assumed the same number of tonal patterns as given in the preliminary classification, and it adjusted the position of the centers and the corresponding clusters.

Then Silhouette width was calculated for each of the clusters in the first partition. The cluster with the smallest Silhouette width was the least distinguishable cluster [8] and could be inaccurate. Thus, the center corresponding to this cluster was removed in the next round of partitioning. Also, in every coming round of partitioning, the cluster that was least distinguishable in the previous round was removed, until there were only two clusters left. This procedure is illustrated in Figure 5.

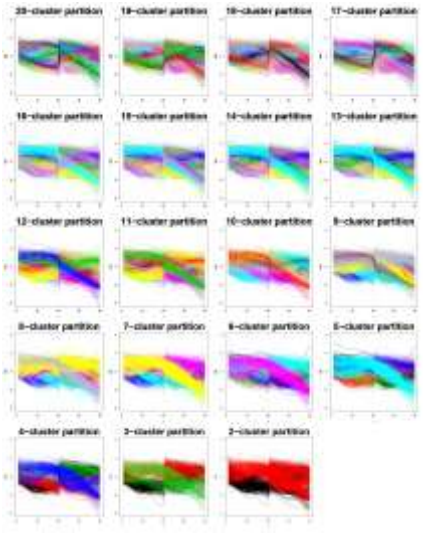

Figure 5. From the first to the last solutions.

A record of Silhouette widths was kept for all the clusters, as well as their mean and standard deviation (SD), in every round of partitioning. On the one hand, the greater the Silhouette width is, the more distinguishable the cluster is, which also applies to the mean Silhouette widths of the whole partition. On the other hand, when comparing the solution where all the clusters are similarly distinguishable against the solution where only some clusters are very distinguishable (and others very messy), we prefer the former. This means that the smaller the SD of Silhouette widths is, the better the solution is. Thus, the goodness of a solution is defined as the Silhouette SD subtracted from the Silhouette mean, taking both criteria into consideration. Accordingly, the optimal solution is chosen from all the candidates (as shown in Figure 6).

Figure 6. The optimal partitioning solution

2.3.3 Adjusting the General Partitioning Solution

Note that some clusters in the optimal partitioning solution, for instance Cluster 7 as shown in Figure 6, appeared to involve different tonal patterns, highlighting that there were sub-clusters that needed further investigation. The phonetician in this study picked out Cluster 7 together with its most similar cluster (Clus-

ter 6) and partitioned them again into four new clusters, Cluster 6, 7, 9, and 10, as demonstrated in Figure 7. The phonetician repeated this procedure until the adjusted solution fit her intuition. Note that, in this process, the adscription of curve was never manually changed. Thus, the adjusted partitioning solution still conformed to the logic of k-means partition, only now with sub-clusters surfacing.

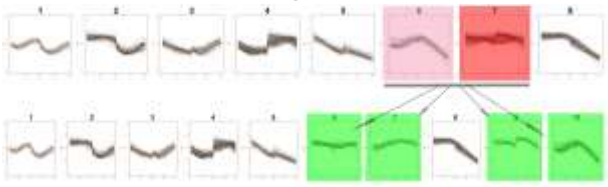

Figure 7. Adjusting Cluster 7 and 6 from the optimized general partitioning solution (upper panel) into four new clusters (Cluster 6, 7, 9, and 10 in the lower panel)

3 Results

3.1 Word-wise Partitioning

As shown in Figure 8, Lexical tonal variants are frequent in JM, but many lexical tonal variants have a low probability.

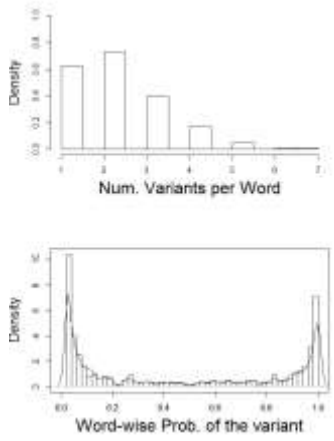

Figure 8. Density plots for the number of variants per word (upper panel) and for the probability of variants.

The phonetician labeled 20 preliminary disyllabic tonal patterns as the preliminary classification. Obviously, the disyllabic tonal patterns are related to the citation tones of the morphemes which composed these disyllabic words. The coding contains two parts, the citation tone of the first syllable (1, 2, 3, or 4) and the citation tone of the second syllable (1, 2, 3, 4, or 5 = neutral tone). As expected, the labeling is more complex than the published linguistic dictionary for JM [12] and the SC tonal categories for reference. Many words had two variants, one ending with a neutral tone and one with non-neutral tones, such as the "35" and "31" variants of "simple" in Figure 4. Since exemplars with extreme values were excluded in corpus preparation, the deepest curve and the trimmed mean curve were usually similar, except that latter was smoother.

3.2 Optimized & Adjusted General Partitioning Results

Figures 9 & 10 show the optimized and adjusted general partitioning solution (with low-probability lexical variants removed). The clusters plotted in separate panels are clearly distinguishable. They represent the disyllabic tonal patterns of JM, optimally eight but these can be further classified into eleven. A prototypical curve can be found for each cluster (either trimmed means or deepest curve), each representing the shape of one tonal pattern.

The general partitioning results indicate tonal merging. Compared with the preliminary classification by the phonetician, the general partitioning results seem to ignore the difference of citation tones in the first syllable. For instance, curves from the presumed tonal classes "31" and "21" were partitioned into the same cluster (as shown in Figure 10 Cluster 2), where these two presumed tonal classes are indeed visually indistinguishable. Similar merging was also found between other presumed tonal classes in "3" and "2" (such as in "31-21", "32-22","33-23", and "34-24"), and between "1" and "4" (such as in "12-42", "13-43", "14-44"). The neutral tone showed a regressive dissimilating sandhi effect on the previous syllable, and its disyllabic tonal pattern sometimes converged with unrelated tonal combinations. For instance, as shown in Figure 10, the presumed tonal class "35" primarily portioned into the same clusters with "13-43" (Clusters 3) or "12-42" (Cluster 4), but showed a very different tonal pattern compared with those of the other tonal classes beginning with the citation tone "3" (e.g. in Clusters 2, 6, 8, and 9). Also, the highest tonal patterns (shown as Cluster 7 in Figure 9 and as Clusters 6, 7, and 9 in Figure 10) were very similar, and only surfaced after adjustments. Nevertheless, the sub-clusters seemed to reflect the difference of monosyllabic citation tones that relate to the disyllabic tonal classes. The Clusters 6, 7, and 9 within the adjusted general partitioning solution are primarily associated with the tonal classes "33-23", "25", and "22-32" respectively.

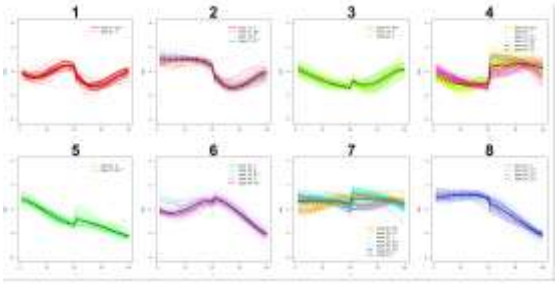

Figure 9. Optimized general partitioning solution color- and line- coded according to the preliminary classification

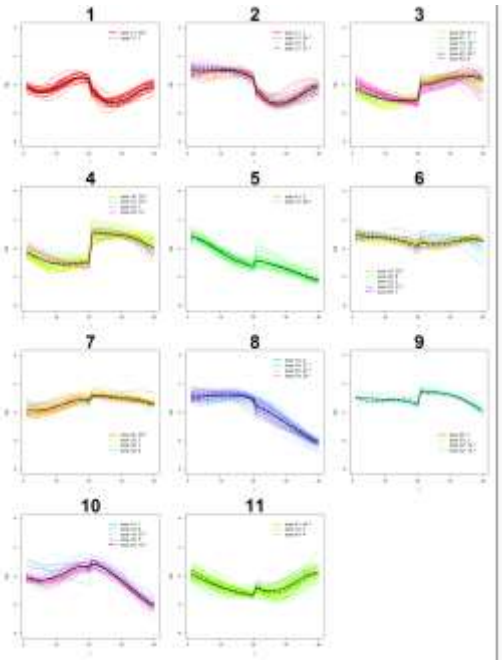

Figure 10. Adjusted general partitioning solution color- and line- coded according to the preliminary classification

4 Discussion and Conclusion

In this paper, we proposed a Two-Stage semi-automatic partitioning procedure to extract lexical tonal variants and tonal patterns from a multi-speaker corpus.

This procedure integrates the phonetician's linguistic knowledge with the objective procedure of partitioning. All the steps conform to the logic of k-means partition [3] and perceptual magnet theory [4, 5], while manual labeling is limited to the lexical level.

The phonetician's workload is reduced in different ways. First, resources from related dialects can be introduced as references in the labeling procedure, reducing the intensity of intellectual challenge. Second, the automatic partitioning and model selection procedures liberate the phonetician from the most difficult and subjective decisions. He or she only needs to mark any curves that "may" come from different tonal categories with different labels, and the algorithms will automatically find out the most appropriate number of tonal patterns and the ascription of each lexical variant. Third, even when part of the optimal partitioning solution is counter-intuitive, the manual adjustments are still limited to pointing out the clusters to be refined, instead of manually correcting the labeling one-by-one.

This procedure also has limitations. First, it can only be applied on corpuses with multiple renditions of the same words. Second, the exemplars processed together must contain the same number of syllables. For instance, the JM corpus only includes disyllabic words. Third, the duration and metrical differences between different tonal patterns are ignored, although they can be important for tonal perception.

In sum, this Two-Stage semi-automatic partitioning procedure, although with limitations, can improve the efficiency and objectivity in the investigation of lexical tonal-pattern variants and basic tonal patterns of an under-resourced language.

References

[1] Besacier L, Barnard E, Karpov A, et al. Automatic speech recognition for under-resourced languages: A survey[J]. Speech Communication, 2014, 56(0): 85-100.

[2] Wu J, Chen Y, Van Heuven V J, et al. Tonal variability in lexical access[J]. Language, Cognition and Neuroscience, 2014, 29(10): 1317-1324.

[3] Hartigan J A, Wong M A. Algorithm AS 136: A K-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100 -108.

[4] Iverson P, Kuhl P K. Tests of the perceptual magnet effect for American English/r/and/l[J]. The Journal of the Acoustical Society of America, 1994, 95: 2976.

[5] Iverson P, Kuhl P K. Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling[J]. The Journal of the Acoustical Society of America, 1995, 97(1): 553-562.

[6] Estivill-Castro V. Why so many clustering algorithms: a position paper[J]. ACM SIGKDD Explorations Newsletter, 2002, 4(1): 65-75.

[7] Febrero-Bande M, Fuente M O D L. Statistical Computing in Functional Data Analysis: The R Package fda.usc[J]. Journal of Statistical Software, 2012, 51(4): 1-28.

[8] Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational and applied mathematics, 1987, 20: 53-65.

[9] Maechler M, Rousseeuw P, Struyf A, et al. cluster: Cluster Analysis Basics and Extensions. R package version 1.15.2. [M]. 2014.

[10] Fraiman R, Muniz G. Trimmed means for functional data[J]. Test, 2001, 10(2): 419-440.

[11] Cai Q, Brysbaert M. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles[J]. PLoS ONE, 2010, 5(6): e10729.

[12] Qian Z-Y. jinan fangyan cidian [Jinan dialect dictionary] [M]//LI R. Xiandai hanyu fangyan da cidian [Dictionary of Modern Chinese Dia-

lects]. Nanjing; Jiangsu Education Press. 1997.
[13] Boersma P, Van Heuven V. Praat, a system for doing phonetics by computer[J]. Glot International, 2001, 5(9/10): 341-345.
[14] Lobanov B M. Classification of Russian vowels spoken by different speakers[J]. The Journal of the Acoustical Society of America, 1971, 49(2B): 606-608.
[15] Chen Y. How does phonology guide phonetics in segment–f0 interaction?[J]. Journal of Phonetics, 2011, 39(4): 612-625.
[16] Breunig M M, Kriegel H-P, Ng R T, et al. LOF: identifying density-based local outliers[C]. ACM Sigmod Record: ACM, 2000:93-104.
[17] R_Core_Team. R: a language and environment for statistical computing [Computer program]. R Foundation for Statistical Computing, Vienna, Austria, version 2.15 [M]. 2013.

This article was published as
Junru Wu, Yiya Chen, Vincent J. van Heuven & Niels O. Schiller (2018). Applying functional partition in the investigation of lexical tonal-pattern categories in an under-resourced Chinese dialect. In J. Tao, Z. Fang, C. Bao, D. Wang & Y. Li (eds.) *Man-machine speech communication*. Singapore: Springer, 24-35.
DOI 10.1007/978-981-10-8111-8_3