Mining linguistic tone patterns with symbolic representation

Shuo Zhang
Department of Linguistics, Georgetown University
ssz6@georgetown.edu

Abstract

This paper conceptualizes speech prosody data mining and its potential application in data-driven phonology/phonetics research. We first conceptualize Speech Prosody Mining (SPM) in a time-series data mining framework. Specifically, we propose using efficient symbolic representations for speech prosody time-series similarity computation. We experiment with both symbolic and numeric representations and distance measures in a series of time-series classification and clustering experiments on a dataset of Mandarin tones. Evaluation results show that symbolic representation performs comparably with other representations at a reduced cost, which enables us to efficiently mine large speech prosody corpora while opening up to possibilities of using a wide range of algorithms that require discrete valued data. We discuss the potential of SPM using time-series mining techniques in future works.

1 Introduction

Current investigations on the phonology of intonation and tones (or pitch accent) typically employ data-driven approaches by building research on top of manual annotations of a large amount of speech prosody data (for example, (Morén and Zsiga, 2006; Zsiga and Zec, 2013), and many others). Meanwhile, researchers are also limited by the amount of resources invested in such expensive endeavor of manual annotations. Given this paradox, we believe that this type of data driven approach in phonology-phonetics interface can benefit from tools that can efficiently index, query, classify, cluster, summarize, and discover meaningful prosodic patterns from a large speech prosody corpus.

The data mining of f_0[1] (pitch) contour patterns from audio data has recently gained success in the domain of Music Information Retrieval (aka MIR, see (Gulati and Serra, 2014; Gulati et al., 2015; Ganguli, 2015) for examples). In contrast, the data mining of speech prosody f_0 data (here on referred to as Speech Prosody Mining (SPM)[2]) is a less explored research topic (Raskinis and Kazlauskiene, 2013). Fundamentally, SPM in a large prosody corpus aims at discovering meaningful patterns in the f_0 data using efficient time-series data mining techniques adapted to the speech prosody domain. Such knowledge has many potential applications in prosody-related tasks, including speech prosody modeling and speech recognition. Moreover, a Speech Prosody Query and Retrieval (SPQR) tool can be also of great utility to researchers in speech science and theoretical phonology/phonetics (tone and intonation).

Due to the nature of speech prosody data, SPM in a large prosody corpus faces classic time-series data mining challenges such as high dimensionality, high feature correlation, and high time complexity in operations such as pair-wise distance computation. Many of these challenges have been addressed in the time-series data mining literature by proposing heuristics that make use of cheaper and more efficient approximate representations of time-series (e.g., symbolic representations). However, a central question to be addressed in SPM is how to adapt these generic techniques to develop the most efficient methods for computing similar-

[1] In this paper we use the terms *fundamental frequency*, f_0, and *pitch* somewhat interchangably.

[2] As the previous research in this specific area is sparse, we have coined this term for the current paper as we conceptualize the time-series data mining based framework for the pattern discovery, similarity computation and content retrieval from speech prosody databases.

Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 1–9,
Berlin, Germany, August 11, 2016.

ity for the speech prosody time-series data (that also preserves the most meaningful information within this domain).

In this paper, we first conceptualize SPM in a time-series mining framework. We outline the various components of SPM surrounding the central question of efficiently computing similarity for speech prosody time-series data. In particular, we propose using Symbolic Aggregate approXimation (SAX) representation for time-series in various SPM tasks. We then evaluate the use of SAX against several alternative representations for f_0 time-series in a series of classic data mining tasks using a data set of Mandarin tones(Gauthier et al., 2007). Finally we discuss potential challenges and SPM applications to be addressed in future works.

2 SPM framework: Computing similarity for speech prosody time-series

2.1 Overview

Formally, a time series T = t_1,...,t_p is an ordered set of p real-valued variables, where i is the time index. Speech prosody consists of time-ordered fundamental frequency f_0 data points computed at a specified hop size, which can be naturally viewed as a time-series.

Due to the typical large size of data mining tasks and the high dimensionality of time-series, it is often impossible to fit the data into main memory for computation. A generic time-series mining framework is proposed as follows (Faloutsos et al., 1994): (1)Create an *approximation* of the data, which will fit in main memory, yet retains the essential features of interest; (2) Approximately solve the task at hand in main memory; (3)Make few accesses to the original data on disk to confirm/modify the solution obtained in Step 2. In practice, however, the success of this generic framework depends on the efficient time-series representation and distance measure in the approximated space that allows the lower bounding of true distances in the original space, along with the appropriate and tractable distance computation(Lin et al., 2007).

2.2 Subsequence generation

There are two ways to generate time-series subsequences (such as sequences of syllabic tones). In the first mode ("long sequence"), we store time-series as a long sequence (e.g., one audio recording file lasting an hour), and we extract subsequences (such as tones or tone n-grams) on the fly while performing data mining tasks by sliding a moving window across the entire time-series. In this mode, the two parameters to be specified are: (1) the length of the desired subsequence; (2) hop size, i.e., the distance between two consecutive windowing (such as n samples, where $n \geq 1$). Most time-series mining applications work in this way (e.g., financial, weather, DNA, EKG, etc). The second mode is to store pre-extracted time-series as individual subsequences and perform experiments on these subsequences directly. The first mode usually generates much more numerous (partially overlapping) subsequences due to the small hop size used in practice.

In speech prosody, however, it is meaningful to store tone or intonation time-series as subsequences. For example, it is generally acknowledged that tones are associated with each syllable (or some sub-syllabic structure like the mora(Morén and Zsiga, 2006)). Intuitively, we are only interested in tone time-series that have beginning and ending points at syllable boundaries. This is true for tone syllable n-grams as well. On the other hand, in a motif discovery[3] context, it is conceivable that f_0 patterns that begin or end in the middle of the syllable could still be of interest (due to misalignment of tone and syllable, such as peak delay (Xu, 2001)). In that case, using the long sequence mode, we might discover novel, previously unknown patterns and knowledges about interesting phenomena in a large prosody corpus.

2.3 F_0 time-series representations and Symbolic Aggregate approXimation

A great deal of effort has been devoted to developing efficient approximate representations for time-series data(Debarr et al., 2012). The limitations of real-valued approaches [4] have led researchers to consider using a symbolic representation of time series. A symbolic approach that allows lower bounding of the true distance would not only satisfy the requirement of the generic framework outlined in Section 2.1, but also enables us to use a

[3] As will be discussed later, motif discovery is a classic time-series mining task that searches for all previously unspecified recurring time-series subsequence patterns in the data set in an exhaustive manner.

[4] Since in many contexts the probability of observing any real number is zero, this may limit the types of algorithms that can work with these representations.

variety of algorithms and data structures that are only defined for discrete data, including hashing, Markov models, and suffix trees (Lin et al., 2007).

Symbolic Aggregation approXmation (or SAX (Lin et al., 2003)) is the first symbolic representation for time series that allows for dimensionality reduction and indexing with a lower-bounding distance measure at the same time. The SAX algorithm first transforms the original time-series into a representation with a lower time resolution, using Piecewise Aggregate Approximation technique (PAA, (Keogh et al., 2001)) as an intermediate step. Then it quantizes the pitch space using an alphabet, therefore transforms the entire time-series into a string. It has two parameters: a *word size*(*w*=desired length of the symbolic feature vector) and an *alphabet size* (*a*), the latter being used to divide the pitch space of the contour into *a* equiprobable parts assuming a Gaussian distribution of F0 values (the breakpoints are obtained from a lookup statistical table). After we obtain the breakpoints, each segment of the time-series can be assigned a letter based on the alphabet bin it is in. Figure 1 shows an example of SAX transformation of a time-series of length 128 into the string 'baabccbc'.

In the current paper, we consider several other representations of f_0 time-series data for evaluation against the SAX representation:

(1) *Non-parametric f_0 representation*. f_0 contour units can be directly represented by downsampled or transformed f_0 data points (originally in Hertz, or transformed to Cent or Bark scale[5]). (Gauthier et al., 2007) showed that unsupervised classification using Self-Organizing Map yielded a 80% correct result with 30-point f_0 vectors. In the same study, the First Derivative Vector (D1) is shown to be a more effective feature than f_0.

(2) *Prosody modeling phonetic model representation.* In speech prosody modeling, the most straightforward phonetic model representation of pitch contours is to use polynomial functions(Hirst et al., 2000) to fit the f_0 contour of each utterance unit (such as a tone). A f_0 contour can thus be represented by the coefficient vector [c_1, c_2,..., c_{n+1}] of a *n*-th order polynomial. An alternative and linguistically more meaningful model is the quantitative Target Approximation(qTA)(Prom-on et al., 2009). qTA models tone/intonation production as a process of syllable-wise pitch target approximation, where the target is defined by a linear equation with slope *m* and intercept *b*. The actual realization of the f_0 contour is constrained by articulatory factors, characterized by a third-order critically damped system with parameter λ.

[5]Cent and Bark scales are transformations of Hertz values in order to more closely reflect the human perception of pitch differences.

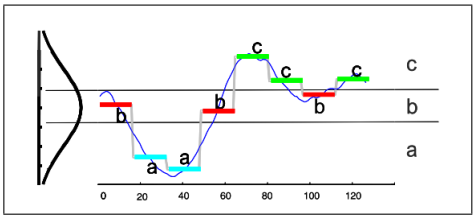

Figure 1: Symbolic Aggregate Approximation, with original length n = 128, number of segments w = 8 and alphabet size a = 3, with output word **baabccbc** (adpated from Lin et al 2007)

Euclidean	DTW(LB_Keogh)	MINDIST
(30)norm-f_0	(30)norm-f_0	(10,20)SAX
(30)norm-f_0-bk	(30)norm-f_0-bk	
(30)norm-f_0-ct	(30)norm-f_0-ct	
(30)D1	(30)D1	
(4)polynomial		
(3)qTA		

Table 1: Experiments on distance measures and time-series (TS) representations. Each column shows various TS representations with a distance measure in the top row, with the dimensionality or dimensionality range of the TS-representation in the preceding parenthesis (norm=normalized, bk=Bark, ct=Cent, D1=first derivative)

2.4 Distance computation

The first two of the following three distance measures work on numeric representation of time-series data, while the third works on symbolic data.

(1) *Euclidean Distance* is an effective and economic distance measure in many contexts despite its simplicity. (Mueen et al., 2009) shows that the difference between Euclidean and DTW distance (discussed next) becomes smaller as the data gets bigger.

(2) *Dynamic time warping* (DTW, see (Rakthanmanon et al., 2012)) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed in a non-linear fashion. The optimal (shortest) DTW distance, or the best alignment between two time series is obtained with dynamic programming (similar to edit distance for string alignment). It is described

in the literature as being extremely effective for computing distances for time-series data (Xi et al., 2006) and has become indispensable for many domains. In practice, due to its high cost, various lower-bounding techniques are proposed to speed up DTW distance computation in a large database. The LB_Keogh lower bounding technique (Keogh, 2002), for example, speeds up the computation by first taking an approximated distance between the time-series that is both fast to compute and lower bounds the true distance. Only when this approximated distance is better than the best-so-far do we need to actually compute the DTW distance between the pair. This makes DTW essentially an *O(n)* algorithm. However, in general DTW distance matrix computation in a big data set remains a more expensive operation in practice. By using symbolic representation, our goal is to find a more efficient representation and distance measure that perform comparably to DTW.

(3) *MINDIST distance function* returns the minimum distance between two SAX strings. It lower bounds the true distances of original data (Lin et al., 2003). It can be computed efficiently by summing up the letter-wise distances in the SAX string (letter-wise distance can be obtained from a lookup table using the Gaussian-equiprobable breakpoints mentioned before). For a given value of the alphabet size a, the table needs only be calculated once, then stored for fast lookup.

Table 1 gives an overview of all the time-series representations and distance measures evaluated in this paper.

2.5 Pitch normalization

Many literature reviews(Lin, 2005) in time-series mining assert that time series must be normalized using the z-score transformation normalization strategy so that each contour has a standard deviation of 1 and mean of 0. However, we observe that z-score transformation distorts the shapes of tone or intonation contours with a relatively flat shape. Since z-score transformation expresses each data point in a time series by its relative value to the mean in terms of standard deviation, it would magnify the differences in the values of the flat or near flat contours, and turn such contours into a significantly un-flat contour. To solve this problem, we propose the Subtract-Mean normalization strategy:

$$z_0 = (x_i - \mu) \tag{1}$$

SAX has a requirement to first normalize the time-series using the standard-deviation normalized z-score transformation. In our implementation, when the standard deviation of a subsequence time-series is less than a pre-set threshold (a very small number), all of its segments will be assigned the same symbol.

3 Related work

There has been limited amount of previous works on f_0 pattern data mining, including MIR f_0 melodic pattern mining, and corpus based speech intonation research.

(Gulati and Serra, 2014) mined a database of melodic contour patterns in Indian Art Music. Due to its astronomical size (over 14 trillion pairs of distance computation), various techniques are used to speed up the computation, including lower bounding in DTW distance computation (Keogh, 2002) and early abandoning techniques for distance computation (Rakthanmanon et al., 2012). The goal is to discover highly similar melodic time-series patterns that frequently occur in the collection (motifs) in an unsupervised manner, and the result is evaluated using the query-by-content paradigm (in which a seed query pattern is provided and top-K similar patterns are returned). The meaningfulness of the discovered pattern is then assessed by experts of Indian music.

(Gulati et al., 2015) experimented with 560 different combinations of procedures and parameter values (including sampling rate of the melody representation, pitch quantization levels, normalization techniques and distance measures) in a large-scale evaluation for the similarity computation in MIR domain. The results showed that melodic fragment similarity is particularly sensitive to distance measures and normalization techniques, while sampling rate plays a less important role.

(Valero-Mas et al., 2015) experimented with SAX representation for the computation of F0 contour similarity from music audio signals in a Query-By-Humming (QBH) task[6] in MIR. Results suggest that SAX does not perform well for music

[6]In a query by humming task, a user hums a subsection of the melody of a desired song to search for the song from a database of music recordings.

TSR-DIST	SAX-MINDIST				BK-EU	HERTZ-EU	HERTZ-DTW	D1	qTA	polynomial
K	1	3	5	7	1	1	1	1	1	1
dimension	20	20	20	20	30	30	30	30	3	4
accuracy	0.81	0.87	0.87	0.89	0.92	0.92	0.93	0.93	0.73	0.70
CR	0.66	0.66	0.66	0.66	1	1	1	1	0.1	0.13

Table 2: K-Nearest Neighbor tone classification results, with 10-fold cross validation (CR=compression rate, TSR=time-series representation, DIST=distance measure, EU=Euclidean Distance, SAX parameters (w,a)=(20,17), test_size=1600, training_size=320)

time-series data in the context of QBH. The authors attribute this to the fact that the SAX representation loses information important in music melodic retrieval through its dimensionality reduction process.

To the best of our knowledge, the only work that attempted at data mining of speech prosody is (Raskinis and Kazlauskiene, 2013)'s work on clustering intonation patterns in Lithuanian (although it did not explicitly employ any time-series data mining techniques). While this work examined the use of a number of distance measures (mainly Euclidean vs. DTW), it is a early-stage research and no clear conclusion was reached regarding either the effectiveness of the discussed methods or the significance of the patterns discovered.

4 Case study: mining Mandarin tones with symbolic representation

In this section we report our initial experimentation using SAX and other types of time-series representations and distance measures discussed above. We evaluate on a data set of pre-extracted subsequences of Mandarin tone time-series. The current experiment aims at a focused evaluation of these techniques on a controlled and relatively small data set in order to study their behaviors when dealing with speech prosody time-series data. Therefore, we have deliberately chosen a smaller and relatively clean dataset from which we can clearly see the consequences when we vary the parameters. The ultimate goal is of course to use this framework to mine large databases of tone and intonation corpora.

4.1 Experimental setup

Our evaluation data set of Mandarin tones is drawn from the (Gauthier et al., 2007) data used for unsupervised learning of Mandarin tones with the Self-Organizing Map. This data set contains lab speech (480*4=1920 tones, three speakers each produced 160 instances of each of the four tone

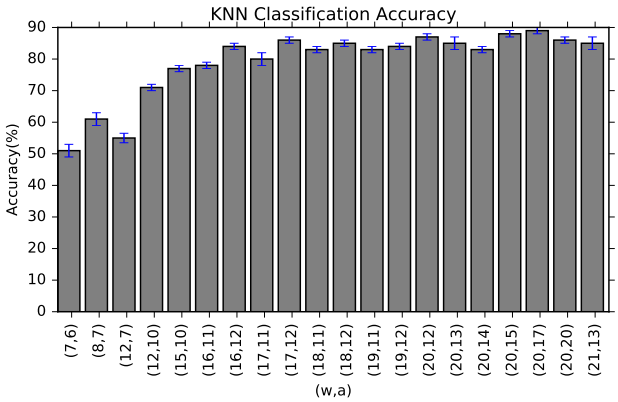

Figure 2: Classification accuracy(%) by SAX parameters w and a

categories), where all possible tone combinations are permuted with the original intention to study the variability of tone shapes in running speech. The tonal environments introduce noisy deviation of tone shapes from tone templates, making tone recognition a mildly challenging task. The target tones are spoken in a carrier sentence and later extracted as syllable-length subsequences.

Table 1 shows the time-series representations and distance measures to be evaluated in this paper. Following conventions in time-series data mining literature, we evaluate these combinations using time-series classification and clustering. For classification, we use k-nearest neighbor (KNN) and Decision Tree, both of which are widely used in time-series classification[7]. We report only accuracy on the classification experiments considering the balanced nature of the data set. All classification experiments are done with 10-fold cross validation with randomized split of data.

Following the convention of using a smaller training size in time-series data mining literature (and considering the optimal time complexity for splitting data size in KNN), the classification ex-

[7]In practice, KNN is an expensive algorithm that scale more poorly than decision trees.

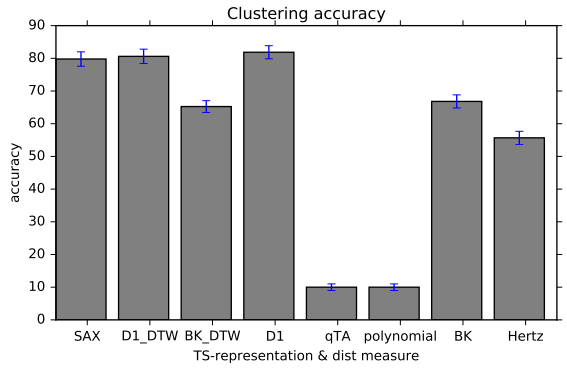

Figure 3: Average clustering accuracy for 1920 Mandarin tones (%) from 5 iterations. For numeric representations, Euclidean distance is used by default unless otherwise noted

periments are carried out using 1600 samples for testing and 320 samples for training (with total size of the data set being 1920 samples of tone contour time-series). To optimize SAX parameters (with MINDIST distance measure), for w, we search from 6 up to $2n/3$ (n is the length of the time series); for a, we search each value between 6 and 20. It is observed that low value for a results in poor classification results (since the MINDIST is a sum of pairwise letter distance between two strings). Figure 2 shows how classification accuracy varies depending on w and a.

For clustering experiments we use the k-means clustering algorithm, where accuracy is computed against true tone labels for each instance. [8]

4.2 Results

First we report time-series classification results on the Mandarin dataset using K-Nearest Neighbor (KNN) and Decision Trees (J48 as implemented in Weka). These classification results are presented in Table 2 and Table 3, respectively. First, we observe that the original f_0 (Hertz) representation performs comparably with normalized-Bark and First Derivative (D1) vectors, using Euclidean distance and DTW distance[9]. All of these achieved above 90% accuracy and F1 score using $K = 1$. The DTW distance showed slight advantage over Euclidean distance. All of the numeric representations performed comparably when K varies, so only results for $K = 1$ are shown. Second, we note that the SAX representation achieved reasonable but lower score (with lower dimensionality, compression rate being 0.66). In particular, it performs worse on $K = 1$, and the performance improves significantly when we increase K to 3, 5, and 7. Third, the qTA and polynomial representations achieved significantly lower classification accuracy, at the advantage of having very low dimensionality (compression rate around 0.1). These trends also showed up in the Decision Tree classification, which has comparable classification accuracy with KNN (with lower cost). Overall, we note that SAX shows slight disadvantage in the time-series classification accuracy, while being able to achieve a 0.66 compression rate for time and space complexity.

The true utility of the SPM framework lies

TSR	SAX	BK	Hertz	D1	qTA	poly
accuracy	0.88	0.93	0.92	0.93	0.83	0.79
CR	0.66	1	1	1	0.1	0.13

Table 3: Decision tree classification results (with 10-fold cross validation), CR=compression rate

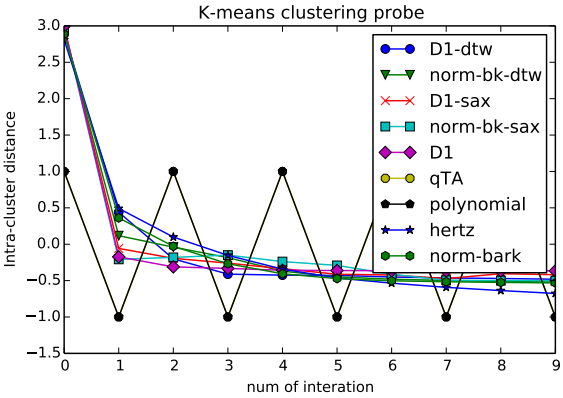

Figure 4: Kmeans clustering objective function by number of iteration. Intra-cluster distance (y axis) is normalized, distance measure noted except for Euclidean. Only polynomial and qTA (overlapped) showed periodic oscillation without a trend to converge.

[8]The clustering accuracy measure is defined by comparing the assigned cluster labels to the true tone labels of each time series, obtaining a confusion matrix showing the true labels against the predicted labels of the cluster assignments, where predicted labels is the most predominant label (i.e., the tone category with the most number of instances among all tones assigned that label) within that cluster. If the the predominant label is not unique across clusters, then a low value of 0.1 is arbitrarily assigned to the accuracy to represent the accuracy is undecidable.

[9]The difference between Bark and Cent features is small, so we only report results for Bark.

in detecting patterns in an unsupervised setting. Comparing to classification, SAX shows more distinct advantage in the clustering experiments. In the following discussion, we note that we are able to use a bigger compression rate for SAX in the clustering experiments (i.e., smaller word size), at $w = 13$, which gives a compression rate of approximately 0.43.

The clustering accuracy is summarized in Figure 3. We establish baseline accuracy of 56% with normalized f_0 representation, indicating the difficulty of this task (although this is still well above chance level of 25% for four tones). Clustering results suggest that (1) The D1 feature significantly outperforms the f_0-based features with Euclidean distance; (2) The DTW distance computed with LB_Keogh heuristic shows its clear advantage with f_0 features, although its utility is quite limited in this dataset, comparing to others; (3) it is noteworthy that SAX is in a much lower dimension yet performs comparably with D1; (4) polynomial and qTA model coefficient based features perform below chance in this task, indicating distances in the original space are not preserved in the ultra-low dimensional parameter space. To probe into these behaviors, we plot the k-means objective function against the number of iteration in Figure 4. In particular, the polynomial and qTA parameters show periodic oscillation of intra-cluster distances, lacking a trend to converge. SAX shows quick convergence, ranking among the most effective.

Overall, in this unsupervised setting, it is noteworthy that DTW is not showing its advantage in computing time-series similarity for the current tone dataset as seen in literature in other domains (see previous discussion). We are yet to evaluate DTW on a harder and bigger dataset for its utility in speech prosody tasks (as comparing to SAX).

Finally, we plot distance matrixes in Figure 5, which may give a hint as to why SAX is a more effective representation than the f_0-Hertz vectors in clustering: In Figure 5 we can clearly see that the lower dimension SAX-MINDIST distance reflects the intrinsic structure of the clusters with lower distances along the diagonal, whereas the distances are all somewhat equalized in the f_0 distance matrix. Overall, SAX and its low dimensionality property may act as a noise-reduction mechanism for revealing the internal structure of the data, making it more suitable for speech tasks comparing to MIR tasks.

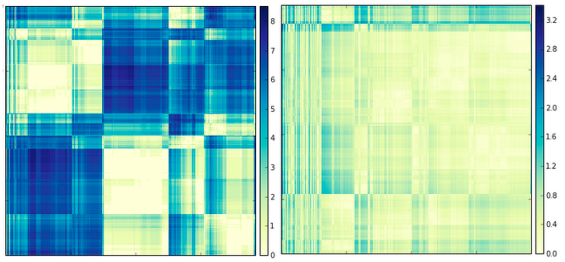

Figure 5: SAX-MINDIST(left) and f_0-Euclidean (right) Distance matrix of 1920 Mandarin tones sorted by tone category. Tones are ordered by tone categories along the x- and y-axis. Origin at top left corner (i.e., on both axis data points are ordered by 480 instances of tone 1, tone 2, tone 3, and tone 4 when moving away from origin).

5 Discussion

In the above experiments we showed the potential of how SPM could benefit from time-series mining techniques, such as low-dimension symbolic representation of time-series that can exploit computational gains from the data compression as well as the availability of efficient string matching algorithms(Ganguli, 2015) (whose utility in SPM is our future research task).

We observed one paradox in our evaluation of SAX, between KNN classification and k-means clustering: the latter is able to achieve better performance (comparing to other time-series representations within the same experiments) with a greater compression rate (0.4) of SAX, whereas the former performs relatively worse with a higher compression rate (0.7) while requiring a larger value of k (k=7 is with SAX performs comparably with k=1 for other representations). We attribute this difference to the nature of the two algorithms, KNN classification and k-means clustering. It is possible that in SAX-MINDIST space, data points have lower distances to cluster centers, but higher distances to its nearest k neighbors within the same tone category (that is, comparing to Euclidean distance).

Meanwhile, a property of SAX is that each segment used to represent the original time-series must be of the same length. This might not be an ideal situation in many applications (exemplified in (Ganguli, 2015)) where variable lengths segments are desired. The complexity of converting

to such an variable-length representation may be greater than the original SAX, as one must design some criteria to decide whether the next frame of audio samples belong to the current segment or it should be the start of a new segment. One intuitive strategy is inactivity detection (i.e., flat period can be represented with a single symbol). Moreover, the utility and implications of symbolic representations (equal- or variable-length) for tones and intonation is also of great theoretical interest to phonology [10].

6 Future works

There are many research questions and applications to be addressed in future works of SPM. Our next step following the current experiments is to evaluate the various time-series representations and distance measures on a larger dataset of spontaneous speech (e.g., newscast speech) in order to find the most efficient methods of computing similarity for speech prosody time-series. Such methods will be useful for all SPM tasks in a large database.

Another useful SPM task is to perform time-series motif discovery in speech prosody data. Motif discovery aims at discovering previously unspecified patterns that frequently occur in the time-series database. Typically we perform motif discovery by generating time-series subsequences on-the-fly (original time-series stored as one long sequence), and then iteratively updating the most frequently occurring patterns. In this way we consider potential motifs in an almost exhaustive manner, with substantial overlaps between consecutive patterns extracted.

Motif discovery has great potential utility for discovering patterns in intonation and tone research. For example, to better understand the nature of contextual variability of Mandarin tone contour shapes in spontaneous speech[11], we might be interested in performing motif discovery with window length being equal to one syllable or syllable n-grams, which considers syllables along with its neighboring tones. Alternatively, we may use variable window length and discover motifs of any length. Of course, a challenge that follows is how to assess the meaningfulness of the motifs discovered.

A direct application of SPM is a Speech Prosody Query and Retrieval (SPQR) tool that can assist phonologists, phoneticians, and speech prosody researchers in their daily research tasks (that might be done manually otherwise). The basic functionality of this tool is for the user to query a speech prosody corpus using a seed pattern (to be defined using a meaningful prosodic unit, such as a syllabic tone, tone n-gram, an intonation phrase), and retrieve the top k similar patterns in the corpus. The seed pattern can be selected using example top K-motifs extracted from the corpus, or using a user-supplied pattern (numeric, symbolic data points, or audio files). The researcher can further assess the meaningfulness of the patterns discovered by means of intrinsic (i.e., within the phonology/phonetics domain) or extrinsic evaluation (i.e., combined with annotations in other domains such as syntax, semantics, and information structure in discourse). Extended functionalities of the SPQR tool will showcase the motif discovery and other applications of SPM. The application can be implemented with a GUI web interface and use pre-computed time-series similarity indexes for faster retrieval[12].

[10] Personally communication with scholars in phonology.

[11] The variability problem refers to the fact that while there exists a canonical contour shape for each tone category, in spontaneous speech the shapes of tones are distorted greatly due to many contextual factors. This is a fundamental challenge in tone recognition.

[12] A similar application for querying melodic patterns in Indian music (developed by Sankalp Gulati at Music Technology Group, Universitat Pompeu Fabra) is available here: http://dunya.compmusic.upf.edu/motifdiscovery/.

References

David Debarr, Jessica Lin, Sheri Williamson, and Kirk Borne. 2012. Pattern recognition in time series. *Advances in Machine Learning and Data Mining for Astronomy*.

C Faloutsos, M Ranganathan, and Y Manolopulos. 1994. Fast subsequence matching in time-series database. *SIGMOD Record*, 23:419–429.

Rastog A Pandit V Kantan P Rao P. Ganguli, K. 2015. Efficient melodic query based audio search for hindustani vocal compositions. *Proceedings of ISMIR 2015*.

Bruno Gauthier, Rushen Shi, and Yi Xu. 2007. Learning phonetic categories by tracking movements. *Cognition*, 103(1):80–106, apr.

Sankalp Gulati and Joan Serra. 2014. Mining Melodic Patterns in Large Audio Collections of Indian Art Music. *Proceedings of International Conference on Signal Image Technology & Internet Based Systems*

(SITIS) - Multimedia Information Retrieval and Applications, Marrakech, Morocco 2014.

Sankalp Gulati, Joan Serr, and Xavier Serra. 2015. An Evaluation Of Methodologies For Melodic Similarity In Audio Recordings Of Indian Art Music. *Proceedings of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Brisbane, Australia 2015*, pages 678–682.

Daniel Hirst, Albert Di Cristo, and Robert Espesser, 2000. *Prosody: Theory and Experiment: Studies Presented to Gösta Bruce*, chapter Levels of Representation and Levels of Analysis for the Description of Intonation Systems, pages 51–87. Springer Netherlands, Dordrecht.

E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. 2001. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*, 3(3):263–286.

E Keogh. 2002. Exact indexing of dynamic time warping. *28th International Conference on Very Large Data Bases*, pages 406–417.

Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. 2003. A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery - DMKD '03*, page 2.

Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Min Knowl Disc*, 15:107–144.

Jessica Lin. 2005. Mining time-series data. *Data Mining and Knowledge Discovery Handbook*.

Bruce Morén and Elizabeth Zsiga. 2006. The lexical and post-lexical phonology of thai tones. *Natural Language & Linguistic Theory*, 24(1):113–178.

Abdullah Mueen, Eamonn Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover. 2009. Exact Discovery of Time Series Motifs. *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 473–484.

Santitham Prom-on, Yi Xu, and Bundit Thipakorn. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*, 125(1):405–24, jan.

Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. 2012. Searching and mining trillions of time series subsequences under dynamic time warping. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 262–270.

G Raskinis and A Kazlauskiene. 2013. From speech corpus to intonation corpus: clustering phrase pitch contours of lithuanian. *Proceedings of the 19th Nordic Conference of Computational Linguistics*.

Jose J Valero-Mas, Justin Salamon, and Emilia Gómez. 2015. Analyzing the influence of pitch quantization and note segmentation on singing voice alignment in the context of audio-based Query-by-Humming. *Sound and Music Computing Conference*.

Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. 2006. Fast time series classification using numerosity reduction. In *Proceedings of the Twenty-Third International Conference on Machine Learning*, pages 1033–1040.

Y Xu. 2001. Fundamental frequency peak delay in Mandarin. *Phonetica*, 58(1-2):26–52.

Elizabeth Zsiga and Draga Zec. 2013. Contextual evidence for the representation of pitch accents in standard serbian. *Language & Speech,*, 56(1):69–104.