

Speech Prosody Mining and the Unsupervised Learning of Mandarin Tones

Shuo Zhang

Department of Linguistics, Georgetown University

Contents

1	Introduction	3
2	Research Questions	4
3	How important is tone recognition?	5
4	Why is tone learning hard?	7
4.1	Sources of variability: Evidence from speech perception and production	8
4.1.1	Local context	10
4.1.2	Broad context	12
4.2	Prosodic modeling	13
4.2.1	Phonology-based models	14
4.2.2	Phonetic-based models	15
4.2.3	PENTA and quantitative Target Approximation (qTA)	17
4.3	Evaluation data set	20
5	Supervised learning of Mandarin tones	21
5.1	Feature representation and selection	21
5.2	Context-dependent modeling	23
5.3	Tone nucleus region modeling	26
6	Unsupervised learning of Mandarin tones	27
6.1	Unsupervised and semi-supervised learning of Mandarin tones	28
6.2	Semi-supervised learning of Mandarin tones	30

7	Speech prosody mining using time-series mining techniques	31
7.1	Overview of time-series mining	32
7.2	Time-series and F_0 feature representation	33
7.2.1	Prosodic modeling feature representation	33
7.2.2	Time-series symbolic representation	35
7.2.3	Time-series normalization	37
7.3	Distance measure	38
7.4	F0 pattern mining using time-series mining techniques	40
8	Chapter outlines	40

Abstract

Tone recognition is an important and difficult task in Mandarin speech recognition that current commercial systems largely leave out. From an information theoretic point of view, tones carry at least as much as information as vowels. Recent years have seen increasing interests in the tone recognition research from the machine learning community, most of which aimed at building robust supervised machine learning system with improved accuracy, while limited effort has been made in unsupervised approaches. In my doctoral thesis work, I explore the speech prosody mining and unsupervised learning of Mandarin tones. In this document, I define the problem of unsupervised learning of Mandarin tone and identify the major challenges in tone recognition by both supervised and unsupervised learning systems. I then outline the main relevant topical areas of research and review the previous works on each of those areas while expanding the exploration of the research questions. Finally, I briefly describe the tentative plans for my thesis work and outline the intended chapters.

1 Introduction

Tone recognition is an important and difficult task in Mandarin speech recognition that current commercial systems largely leave out [7]. From an information theoretic point of view, tones carry at least as much as information as vowels [1]. Recent years have seen increasing interests in the tone recognition research from the machine learning community [2-5].

Several major challenges make tone recognition in spontaneous speech a difficult task. First, although traditionally being the single most important cue of tone recognition, the F0 contour based features alone are often considered to be insufficient, and must be coupled with other types of features such as intensity, voice quality, etc [1]. However, the improvement by incorporating these features is usually small. Second, it has been shown that the distortion of F0 contours from the canonical templates of tone categories (i.e., the surface tone contour variability problem) mainly comes from two sources : (1) local context, i.e., coarticulation caused by the adjacent tone environments of the current syllable, due to the physiological and phonological constraints; (2) Higher level broad range context, or sometimes known as expressive functions[6] outside of prosodic domain, such as focus, sentence modality (e.g., question vs. declaration), and sentence length. Therefore, how to identify and exploit this context information in order to restore the underlying tone target can be crucial to substantial improvement in utilizing F0 based features in

tone recognition. Third, to this date, most systems of tone recognition employ supervised learning, which relies on the high cost of manually annotating and labeling a large amount of audio data.

There has been limited effort on the unsupervised and semi-supervised learning of tones. [7,8] explored the clustering of Mandarin tones and English pitch accent using spectral clustering with both F0 and intensity features. The result is in general worse but in many cases comparable to the best supervised algorithms. However, many aspects of this unsupervised learning task can be further investigated, including (but not limit to) the most effective use of feature representation, distance measures, and clustering algorithms.

In the meantime, unsupervised learning of tones using time-series mining techniques has shown great potential for improving the efficiency and accuracy of this task. [9] showed that by viewing tone F0 contours as a collection of time series, were able to employ techniques from time-series mining to speed up the computation of distance measures as well as requiring less storage space. The Symbolic Aggregate approXimation (SAX) representation of time series allows for efficient computation based on string based algorithms, while achieving better clustering accuracy on a read speech data set than using numeric representation with Euclidean distance (and comparable with numeric feature using Dynamic Time Warping distance measure, and the First Derivative (D1) feature with Euclidean distance). However, this approach is yet to be evaluated on a comparable data set as [7,8], as well as spontaneous speech data sets.

2 Research Questions

This Problem Statement is based on the following goals and research questions of my thesis project.

This project has two main goals:

(1) In **speech prosody mining**, use time-series mining and pattern discovery to better understand the F0-contour prosodic patterns in a corpus of spontaneous speech and the sources of variation for F0 tone variation. This part is carried out with annotated metadata of syllable boundaries, tones, and other non-acoustic information.

(2) In **unsupervised learning of tones**, use various context-dependent modeling and unsupervised learning to improve the accuracy of tone recognition.

In this project, we focus on the problem of unsupervised learning of Mandarin tones with the following research questions:

(1) Better understanding the tone recognition problem. Are human listeners always able to recognize tones in or out of context, based on acoustic signal alone with reasonable high accuracy? If not, from an information theory perspective, how much information does tone carry comparing to the segmental, lexical and other non-tonal contextual information in human speech understanding? (in other words, how much can Mandarin speakers understand from mono-tone Mandarin speech? This has implications on the question of how well should we expect a tone recognizer to perform based on acoustic information alone)

(2) Context-dependent modeling in tone recognition.

(2a) What can we learn about the sources of variation for tone F0 contours in a large speech corpus, by using time-series mining based pattern discovery algorithms?

(2b) How can we incorporate local context-dependent (i.e., coarticulation depending on the adjacent tone contexts) information to modify the F0 contour time series that makes the unsupervised or semi-supervised learning more effective?

(2c) How can we incorporate broader context information (e.g., focus, modality, etc.) to improve tone recognition accuracy in an unsupervised framework?

(3) Features. Using these techniques, can we obtain comparable or better performance using F0 contour based features alone?

(4) Unsupervised learning. How can we combine F0 based features with other types of features (e.g., spectral) to improve the performance in an unsupervised or semi-supervised framework?

3 How important is tone recognition?

Until very recently, the state-of-the-art speech recognition of tone languages typically discard tone information altogether. There has been many discussions on why this is so [], which is mainly attributed to the high error rate of tone recognition that supersedes the benefit of including tone information. Recent development in tone recognition, however, reveals the importance of tone information in speech recognition from an information theoretic point of view. In [Surendran 2004], the functional load of tones is found to be at least as high as vowels while lower than the consonants. Here, the functional load of tones (FL) is defined as the information we lose if the phonological system loses the contrast posed by tones:

$$FL(tone) = \frac{H(M_u) - H(M_u^{-tone})}{H(M_u)} \quad (1)$$

where M_u is a sequence of units of type U in Mandarin, U being either a syllable or word in the entropy function $H(\cdot)$ calculation, M_u^{-tone} refers to such a sequence without tone contrast, and $H(\cdot)$ is the standard entropy function of the system (i.e., a well defined *language*). FL is defined as the functional load of tone. Intuitively, it characterizes how much information the system lose if the tone contrast is lost, therefore how important is a contrast within a phonological system. Table 1 shows the functional load of tones versus other segmental contrast in Mandarin.

Table 1: Functional Load of tones vs other contrasts in Mandarin

x	$FL_{syll}(x)$	$FL_{word}(x)$
Consonants	0.235	0.081
Tones	0.108	0.021
Vowels	0.091	0.022
Stops	0.029	0.006
Fricatives	0.021	0.005
Place	0.065	0.014
Manner	0.034	0.006
Aspiration	0.002	0.0003

This analysis captures the importance of tones in Mandarin speech recognition systems. However, it should also be interpreted carefully. First, it has the implicit assumption of taking a pure phonological point of view while ignoring all other information that can be used to distinguish words and syllables such as n-grams language models, which uses non-acoustic information during HMM based speech recognition.

Second, on the other hand, it also calls for an analysis of how important tone is for Mandarin speech recognition in humans. It is often an implicit assumption in tone recognition literature that humans are always able to perform with reasonably high accuracy in tone recognition even when contextual tonal, segmental, and other information are unavailable. However, speech experiments have revealed that native Mandarin speakers perform below chance in isolated syllable tone recognition when additional information usually available in speech is removed [], depending on the specific experimental conditions. This is also supported by additional experimental evidence that humans are

able to perform with greater than 90% accuracy in speech recognition with tone information removed in Mandarin (i.e., monotone F0 is imposed synthetically) [].

Therefore, it is also worthwhile to investigate from an information theory point of view just how important is tone to Mandarin word recognition in human speakers (which from the above discussion, would reveal that it is much less important than [Surendran 2004] have suggested). The implication of such an analysis would be that, we cannot expect a machine to perform perfectly on isolated syllable tone recognition (or context-independent) when humans cannot do it in the first place. Meanwhile, it suggests the importance to develop context-dependent tone recognition algorithms that also uses contextual information beyond the information present in the acoustic signal of isolated syllables.

4 Why is tone learning hard?

In the canonical forms of Mandarin tone system, four tone categories are present: high-level (High tone), rising (Rising tone), low-dipping (Low tone) and falling (Falling tone). Following convention, these are also referred to as tone 1, 2, 3, and 4 in this paper. Figure 1 shows the canonical contours of four Mandarin tones. In running speech, however, the contour shapes often become much distorted from canonical shapes, making it difficult to identify correct tone categories. Therefore, the challenge of tone recognition is to identify sources of variability, and exploit these knowledge to restore the underlying tone category. In continuous speech, there are many local and broader factors that play a role in producing the final tone contour shapes. These factors are usually rather convoluted and confounded and is therefore hard to identify. It requires carefully designed control experiments to reveal the effect and function of each contributing factor.

As an example, Figure 2 shows how a tone-controlled production experiment reveals variability of tone contour shapes in statement vs. question sentences with varying focus position [data from Liu et al 2006] . Looking at this simple example, which is far less complex than real spontaneous speech data, gives us much clue to why is tone learning hard. First, we can see clearly that even though this is all Rising tone, most tones in this sentence have a rather flat shape, resembling the High tone. This is contrasted with Figure 3, where a sentence is spoken with all high tones. Second, the variability of the tone shape is dependent on the sentence focus and modality conditions. These must also be taken into account in tone recognition with real data, where this

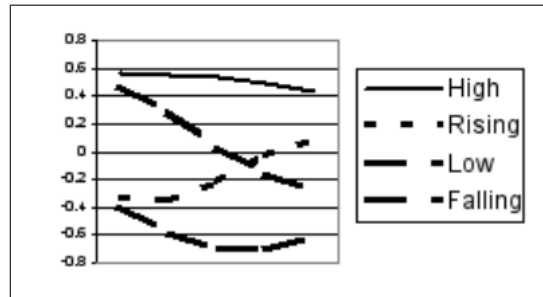


Figure 1: Mandarin 4 tones canonical contour shapes (adapted from [Levow 2006])

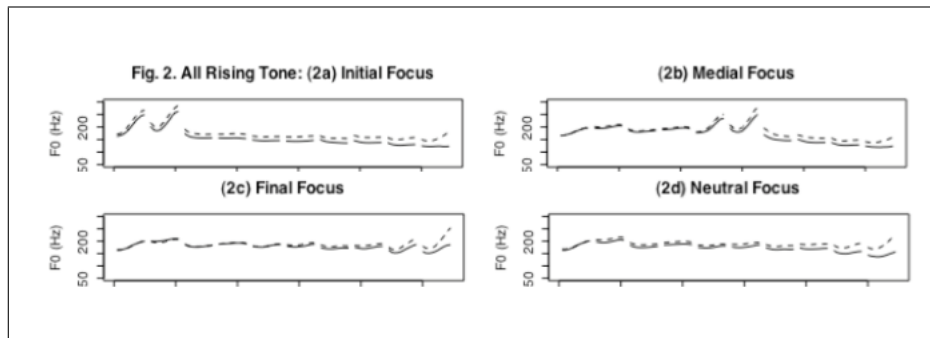


Figure 2: Experimentally controlled production of all Rising tones in statement vs question sentences with varying focus position (adapted from [Liu et al 2006])

type of uniform tone combination is highly unlikely.

In this section we discuss previous works on the sources of variability from both speech production and prosodic modeling perspective.

4.1 Sources of variability: Evidence from speech perception and production

[Gauthier et al 2007] identified two major sources for the extensive overlap between the tone contour shapes in running speech. The first is the difference in the pitch range of individual speakers and the second is the variability introduced by tonal context in connected speech [Shen 1990, Xu 1997]. In tone recognition literature, the speaker difference is usually removed by normalization. The tonal context, as mentioned above, is where the majority of the research effort has been devoted in identifying and isolating sources of variability.

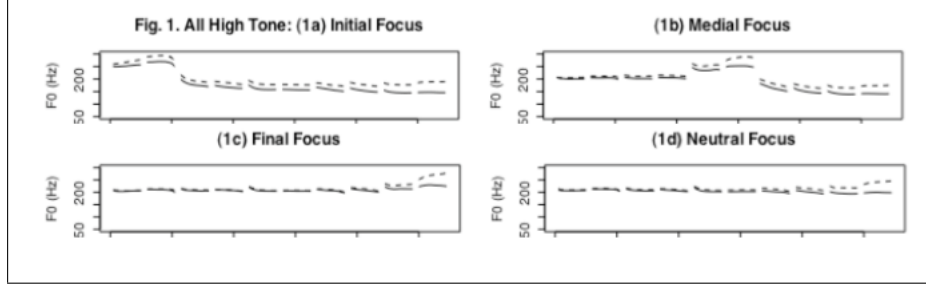


Figure 3: Experimentally controlled production of all High tones in statement vs question sentences with varying focus position (adapted from [Liu et al 2006])

Recent works from speech production and perception experiments have made substantial progress on identifying sources of variability in tone production by carrying out experiments with carefully controlled conditions and designed data set[]. These experimental works also became important references for the tone recognition research community, as it is important to also exploit these knowledge in the machine learning of tones and recover underlying tone target. These are referred to as context-dependent (CI) models.

It is generally assumed that the prosodic contours in Mandarin is shaped by at least two different levels of factors: tones, which work at the syllable level, and intonation, which works at a higher level (phrase, sentence, etc). As mentioned above, the great difficulty to isolate these layers of factors from the observed prosodic contour in speech data calls for carefully designed production studies that vary one set of factors while keeping the other set fixed. Recent works of such nature have identified the two levels of sources for tone variability in continuous speech production: First, a *local context* refers to the distortion to tones from its canonical shapes due to the co-articulation with adjacent tones. Second, a *broad context* refers to the further modification of tones on the intonation level of a larger prosodic unit than syllable. Examples of broad context including the topic, focus, and modality of a sentence. Often these broad contexts originate from domains outside of the phonological and articulatory constraints of speech production, and is motivated by communicative functions in the syntactic, semantic and pragmatic domain []. The next two sections review works on both the local and broader context in tone production.

4.1.1 Local context

The local context of tone production is concerned with the behavior of tone contour F0 trajectories with regard to its immediate environment, i.e., adjacent syllables. Experiments were conducted with two principles in mind: first, to examine the effect of neighboring tones, speech production tasks are designed to reflect all different combinations of tones as target words. Second, to minimize the effect of higher level intonation, participants are asked to read the target words embedded in carrier sentences with careful speech with neutral focus. Several important results emerge from these studies of the local context.

(1) **Local "conflicting" context distorts tone contour shapes in incompatible environments.** [Xu 1994] studied how the amount of deviation of a tone from its canonical form due to coarticulation varied depending on the nature of the tonal context. By grouping tone environments into "compatible" vs "conflicting" contexts, it was observed that in a context where adjacent tonal values agree (a "compatible" context, such as when the High tone "HH" ending with "H" is followed by a High-Falling tone "HL" beginning also with "H")¹, the deviation was relatively small. In a context where adjacent tonal values disagree (a "conflicting" context, such as when the said High tone is followed by a Low-Rising tone "LH" starting with "L"), the deviation was much greater, sometimes even to the extent of changing the direction of a dynamic tone.²

The author also examined the perception of co-articulated tones when tones are presented out of context in isolation. The results suggest that identification of tones in the compatible context was highly accurate with or without the original tonal context. Tonal identification for the conflicting context remained accurate only when the tones were presented with the original tonal context. Without the original context, i.e., in isolation, correct tone identification dropped below chance. As discussed above, this result provides important counter-evidence to the "myth" in tone recognition literature that human listeners are always able to recognize tones with high accuracy, even when they're distorted and in isolation.

(2) **Carry-over effects is greater than anticipatory effects in tone co-articulation.** [Xu 1997] examines acoustic variations of tones in Mandarin

¹The "H"(High),"L"(Low) refers to phonological compositional representation of tones, where for example, a rising tone is represented as "LH" with two pitch targets: low and high.

²The authors refer to tones with moving targets as "dynamic tone" (such as a low-rising tone), vs. static tone, where the tone has stable pitch targets (such as a high tone).

under the influence of different tonal contexts. In particular, variations in the four Mandarin tones due to anticipatory and carry-over effects are analyzed by examining the time course of f_0 contours of bi-tonal sequences. Using balanced nonsense sequences produced in different carriers with balanced tonal structures, this study establishes a baseline for local contextual tonal variation in Mandarin. It is found that anticipatory and carry-over tonal influences differ both in magnitude and in nature. Carry-over effects are mostly assimilatory: the starting f_0 of a tone is assimilated to the offset value of a previous tone. Anticipatory effects, on the other hand, are mostly dissimilatory: a low onset value of a tone raises the maximum f_0 value of a preceding tone. While the magnitude of the carry-over effect is large, anticipatory effects are relatively small. This conclusion has been cited many times subsequently by tone recognition researchers, whose own data analysis also showed support for this asymmetry over and over again []. There are also many machine learning approaches that take this effect into consideration in tone modeling.

(3) **Physiological constraints of tone co-articulation.** [Xu and Sun 2002] studied the maximum speed of pitch change in human speech production, which contributes to the understanding of the underlying mechanism for tone co-articulation and its implication for tone modeling. An experimental paradigm was adopted in which subjects (native speakers of English and Mandarin) produced rapid successions of pitch shifts by imitating synthesized model pitch undulation patterns. The speed of pitch change was measured both in terms of response time and excursion time needed to complete the entire pitch shift. Results show that excursion time is nearly twice as long as response time. Also, it is found that the maximum speed of pitch change varies quite linearly with excursion size, and that it is different for pitch rises and falls. Comparisons of present data with data on speed of pitch change from studies of real speech found them to be largely comparable. This suggests that the maximum speed of pitch change is often approached in speech, and that the role of physiological constraints in determining the shape and alignment of F_0 contours in speech is probably greater than has been appreciated.

(4) **F_0 peak delay.** Fundamental frequency (F_0) peak delay refers to the phenomenon that an F_0 peak sometimes occurs after the syllable it is associated with either lexically or prosodically. Previous studies suggested that peak delay was only found in Rising tones in Mandarin []. [Xu 2001] investigated peak delay and its relationship with tone, tonal context, and speech rate. Analysis of F_0 contours and peak alignment revealed that at normal speech rate, peak delay occurred regularly in both the R and weakened H tones but only occasionally in the H tone; at slow speech rate, peak delay continued to occur regularly in the R tone but only occasionally in weakened

H and rarely in the H tone; at fast speech rate, peak delay occurred not only regularly in the R and weakened H tones, but also frequently in the H tone. Results of F0 contour alignment analysis indicate that peak delay occurs when there is a sharp F0 rise near the end of a syllable, regardless of the cause of the rise. The author concluded that much of the variability in the shape and alignment of F0 contours in Mandarin is attributable to the interaction of underlying pitch targets with tonal contexts and articulatory constraints, rather than due to actual misalignment between underlying pitch units and segmental units.

4.1.2 Broad context

In this section we review previous works regarding the effect of broad context on tones in speech production and perception.

(1)**Focus in the temporal domain.** It is well known that focus affects the pitch of what is being focused. More importantly, [Xu et al 2004] showed that focus also extensively affects the pitch ranges of non-focused regions in a sentence, suggesting a wide effect of focus on the temporal domain. Results from this study show that in a declarative sentence, focus is realized not only by expanding the pitch range of the focused item, but also by compressing the pitch range of post-focus items, and possibly requiring that the pitch range of pre-focus items remain neutral. The authors proposed that the domain of a single, narrow focus consists of three temporal zones (pre-focus, on-focus, post-focus), with distinct pitch range adjustment for each. This proposal has received positive support when it was later applied to a machine learning model that improves tone recognition accuracy by incorporating focus information [Surendran 2005]. [Liu et al 2006] also used focus as an effective input feature to a decision-tree based classifier to predict the modality of a sentence from the prosodic domain (question vs. statement).

In [Xu 1999], the author further examined how lexical tone and focus contribute to the formation and alignment of F0 contours using short Mandarin sentences consisted of five syllables with varying tones on the middle three syllables. The sentences were produced with four different focus patterns: focus on the first, second, or last word, or with no narrow focus. The results indicate that while the lexical tone of a syllable is the most important determining factor for the local F0 contour of the syllable, focus extensively modulates the global shape of the F0 curve, which in turn affects the height and shape of local contours. Moreover, despite extensive variations in shape and height, the F0 contour of a tone remains closely aligned with the associated syllable.

(2)**Long-term F0 variations.** Many tone recognition algorithms incor-

porate features that encode the relative pitch heights of the tones [], in conjunction with the features that encodes the contour shapes. However, the pitch heights of tones in longer temporal units of spontaneous speech (e.g., a sentence) must be normalized not only based on individual speaker differences, but also to compensate for the long-term F0 pitch variations due to factors such as down-drift and other pragmatic functions such as focus and topic.

[Wang and Xu 2011] reports an experimental investigation of the prosodic encoding of topic and focus in Mandarin by examining disyllabic subject nouns elicited in four discourse contexts. They also looked at how prosodic effects of topic and focus differ from each other and how they interact with sentence length, downstep and newness to determine sentence-initial F0 variation. Sixty short discourses were recorded with variable focus, topic level, newness, downstep, and sentence length conditions by six native speakers. The important conclusions include:

- (a) The difference between topic and focus lies in that focus both raises on-focus F0 and lowers post-focus F0, while topic raises the F0 register at the beginning of the sentence while allowing F0 to drop gradually afterwards.

- (b) topic has higher pitch register in isolated and discourse-initial sentences than in non-initial sentences,

- (c) longer sentences have higher sentence-initial F0 than shorter sentences, but the differences are small in magnitude and are independent of topic and focus,

- (d) the effect of downstep is independent of topic and focus, but is large in magnitude and accounts for a significant amount of the F0 declination in a sentence,

- (e) newness has no F0 manifestation independent of other factors, but a newly mentioned word is slightly longer than a previously mentioned word, and

- (f) the effects of topic, focus, downstep and sentence length are largely cumulative.

4.2 Prosodic modeling

Research in prosodic modeling provides evidence and often validation to the results obtained from speech production experimental works in tone variability. The goal of prosodic modeling is to analyze and synthesize speech prosody that is as natural as possible, while requiring less resources in storage, computation, and supervision. These models are often used to generate F0 contours for speech synthesis. As such, typically prosodic models are evaluated by means of its capability to (1) re-synthesize speech melodies that closely approximate

the original data (training data) ; (2) predicatively synthesize / generate speech melodies of unseen data given annotated sentence and syllable conditions (i.e., input features). Most systems include an evaluation of (1) but only few aim to do (2).

Meanwhile, on a different level, prosodic modeling is also associated with the theoretical aspect of the representation of speech intonation in the suprasegmental domain of phonology, and the underlying mechanism for the production and perception of speech prosody. While the current project is not concerned with either the generation of F0 contours for speech synthesis or the phonological and psycholinguistic theory of speech production and perception, there are a few aspects in prosodic modeling that can shed light on the machine learning of tones. Specifically, the representation of intonation is directly relevant to the choice of the type of feature vectors we use in tone recognition. Moreover, the modeling of F0 generation process can also provide clues as to how tone variability occurs in real-time spontaneous speech.

In the next two sections, I follow the convention in speech prosodic modeling and divide the review of relevant works into two parts: phonology-based and phonetic-based models. One of the phonetic-based models of particular interest to Mandarin is the pitch Target Approximation and PENTA model [Prom-on et al 2009], which I will discuss separately. This model was developed specifically with Mandarin tones in mind, while it has been also generalized to be a language independent model.

4.2.1 Phonology-based models

The goal of a phonological model is to study the universal organization and underlying structure of intonation. Complex intonation patterns are compressed into a set of highly succinct and abstract vocabulary with wide coverage. These (categorical) symbols are regarded as the primitive entities in representing intonation. The development of such an inventory is often based on phonetic analysis of F0 curves either from a production perspective or from a perception perspective.

One of the dominant views on intonational phonology is the Autosegmental-Metrical (AM) model [Goldsmith 1990; Ladd 1996]. Ladd (1996) states four principles of the AM approach to intonation: Linearity of tonal structure; Distinction between pitch accent and stress; Analysis of pitch accents in terms of level tones; Local sources for global trends.

The most influential example is Pierrehumberts model for American English [Pierrehumbert 1980]. The original model has been extended by Beckman and Pierrehumbert [1986], and has evolved into a standard for transcribing in-

tonation of American English - Tone and Break Indices (ToBI) [Silverman et al 1992]. The main idea of Pierrehumberts model, or AM approach in general [Ladd 1996], is that English intonation can be represented by two types of tones: High and Low. The model describes intonation with multiple levels. The largest prosodic unit is called intonational phrase, which can contain several intermediate phrases. A phrase is represented as a sequence of high (H) or low (L) tones.

The tone inventory include: *Pitch Accents* can be either single tones (H^* , L^*) or bitonal (H^*+L , $H+L^*$, L^*+H , $L+H^*$). The $*$ symbol represents the alignment of a tone with the accented syllable. *Boundary tones*, denoted by the $\%$ symbol, are single High or Low tones aligned with the edges of a phrase ($H\%$, $L\%$). Thus, they indicate the onset and offset pitch of an intonational phrase, respectively. *Phrase accents*, indicated by the $-$ symbol, are used to represent the pitch movement between a pitch accent and a boundary tone ($H-$, $L-$).

Pierrehumberts model has a linear structure in describing intonation in that intonation is solely determined by a local component, which is in contrast to the superpositional approach which treats intonation resulting from the addition of several components, such as a local pitch accent and a global phrase contour. The mapping from phonology onto acoustics and physiology is a dynamic interpretative process [Pierrehumbert and Beckman 1988]. In this process, a set of phonetic realization rules are applied to convert abstract tonal representation into F0 contours by considering the metrical prominence of the syllables and the temporal alignment of the tones with the accented syllables [Pierrehumbert 1981]. This direct phonology-F0 mapping implies that a distinct layer of phonetic model is not necessary.

4.2.2 Phonetic-based models

Phonetic models use a set of continuous parameters to describe intonation patterns observable in an F0 contour (Taylor, 2000). An important goal is that the model should be capable of reconstructing F0 contours faithfully when appropriate parameters are given. However, to make it functional, a phonetic model should also be linguistically meaningful. In fact, using certain functions, such as polynomial equations, to accurately represent F0 contour is not a difficult task. What is more challenging is developing a model whose parameters are predictable from available linguistic information.

Fujisaki Model. [Fujisaki et al 1983, 1988] developed an intonation model for Japanese, which has also been applied to other languages (e.g., German (Mixdorff, 2000)). The model additively superimposes a basic F0

value (F_{min}), a phrase component, and an accent component, on a logarithmic scale. The control mechanisms of the two components are realized as critically damped second-order systems responding to impulse commands in the case of the phrase component, and rectangular commands in the case of the accent component. The Fujisaki model attempts to relate the functions to the human speech production apparatus. It is a typical superpositional approach in that it assumes different intonation components to be superimposed on top of each other. This is different from the AM approach described above, which is predominantly sequential. One of the critical steps towards application of the Fujisaki model is the extraction of model parameters from F0 contours. Fully automatic extraction of model parameters has been developed (Mixdorff, 2000), which would facilitate researchers working with the model on large corpora.

quantitative Target Approximation. A pitch target approximation model for generating F0 contours in Mandarin Chinese was proposed by [Xu and Wang 1997, 2001] and quantified in [Xu et al 1999] with the quantitative Target Approximation model (qTA). In this model, the surface F0 contour is viewed as the result of asymptotic approximation to an underlying pitch target, which can be a static target ([high] or [low]) or a dynamic target ([rise] or [fall]). A pitch target is defined as the smallest unit that is articulatorily operable. The host unit of a pitch target is assumed to be the syllable (for Mandarin, at least). This model is primarily based on evidence from Mandarin, where four basic pitch targets ([high], [low], [rise] and [fall]) are associated with the lexical tones in Mandarin: H (High), L (Low), R (Rising) and F (Falling), respectively. Rather than simply simulating the surface F0 contours, this model aims at some deeper speech production mechanism, which is similar to the Fujisaki model in this regard. It also emphasizes the role of articulatory constraints in intonation modeling. Due to the relevance of the qTA model to the current project, I discuss the relevant literature in more detail in the next section.

Non-parametric models. Both phonological and phonetic frameworks seek to model F0 contours effectively with a set of more abstract representations. However, F0 values themselves are no doubt good indicators of high-level linguistic information [Sun 2002]. As discussed above, the goal in developing parametric models is to find a better representation of F0 contours. However, if inappropriate forms are used, the predictions can be significantly different from the original F0 contours [Sun 2002]. Therefore, some researchers use the original F0 contours directly or F0 with some trivial modifications as the output targets. Such systems are referred to as non-parametric models [Black and Hunt 1996; Ross and Ostendorf 1999;], and often times they can achieve

very competitive results. In this thesis I also experiment with such F0 based features in time-series mining, and explore how feature representations based on F0 transformations perform effectively in clustering tasks.

4.2.3 PENTA and quantitative Target Approximation (qTA)

[Prom-on et al 2009] reports the development of a quantitative target approximation qTA model for generating F0 contours of speech. The model simulates the production of tone and intonation as a process of syllable-synchronized sequential target approximation [Xu 2005]. As a speech production model, the qTA and the associated Parallel Encoding and Target Approximation (PENTA) model has generated much debate on its architecture, representation, and predictive powers [see]. In the current context, however, we're only interested in evaluating how good is the qTA model in terms of its capability to numerically predict speech prosody contours as well as its representation of tone targets as input features to our unsupervised learning framework. The qTA model is evaluated by extracting function-specific model parameters from natural speech via supervised learning automatic analysis by synthesis and comparing the F0 contours generated with the extracted parameters to those of natural utterances through numerical evaluation and perceptual testing. In this model, each tone is produced with a pitch target in mind, defined by a linear equation with a slope and a intercept parameters, m and b :

$$x(t) = mx + b \quad (2)$$

However, the realization of this target is often constrained and deviated by the characteristic factors of the human vocal folds, such as the continuity of pitch change (no sudden change in the derivatives of the curves across syllable boundary) and the limitation of the maximum speed of pitch change [Xu and Sun 02]. As a result, actual F0 contours of tones are characterized by a third-order critically damped system:

$$f_0(t) = x(t) + (c_1 + c_2t + c_3t^2)e^{\lambda t} \quad (3)$$

Intuitively this can be seen as casting a noise component on top of the linear pitch target. Solving this equation, we have in total three parameters to represent a tone contour with the qTA model: slope and height of the pitch target, namely, m and b , and λ , which represents how fast the pitch change is approaching the target slope and height. Figure 4 shows a number of different combinations of the parameters to demonstrate how the actual F0 behaves in accordance with the underlying pitch targets.

Problem Statement

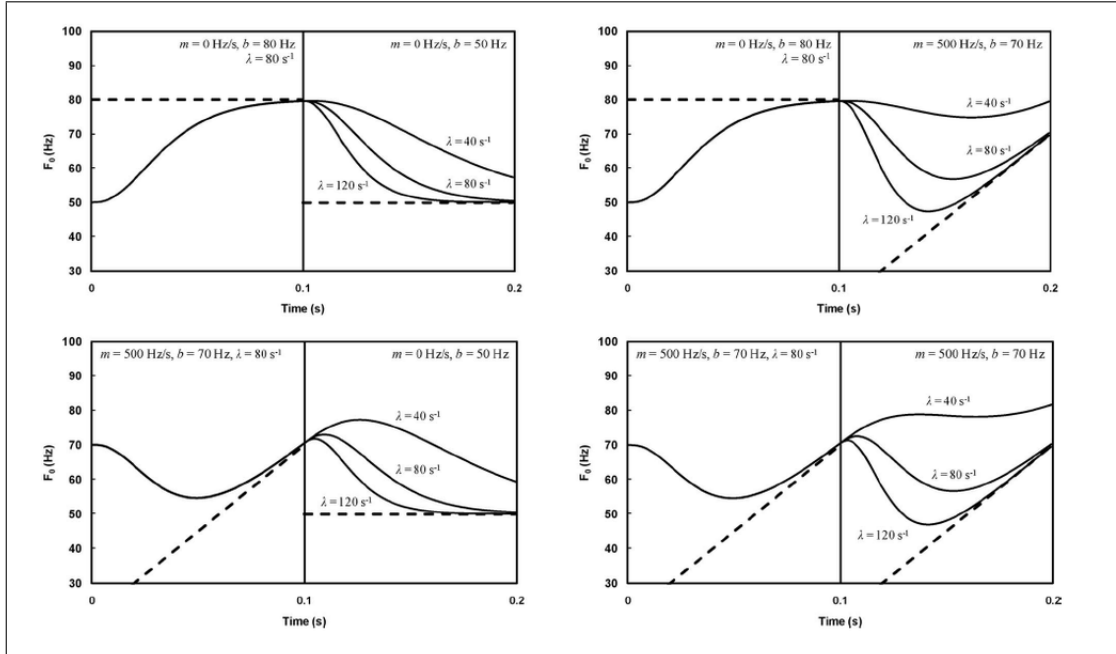


Figure 4: Examples of F₀ contours generated by the qTA model with varying values of m , b , and λ . The dashed lines indicate the underlying pitch targets which are linear functions of m and b . The vertical lines show the syllable boundaries through which the articulatory state propagates.

The computational tools for extracting parameters and performing re-synthesis based on qTA model have been developed in Praat Scripting Language (PSL). In the early version of this tool (PENTATrainer1³), parameter optimization is achieved by an exhaustive search through the parameter space via error minimization algorithms [Prom-on et al 2009]. The local parameter sets learned from this process are then summarized into categorical ones by averaging across individual occurrences of the same functional categories [Prom-on et al 2009]. While the synthesis results did closely approximate the original F0 contours, there are few disadvantages of this strategy. First, the estimated parameters are optimal for the local syllable but not necessarily for the functional categories. Second, the estimation of λ is often not satisfactory because it may be stuck at a local minimum and fails to converge to global minimum [Xu and Prom-on 2014].

To address this problem and upgrade the qTA model to include more features reflecting function-related variability, [Xu and Prom-on 2014] developed PENTATrainer 2 using stochastic learning from real speech data with annotations of metadata information about the sentences (referred to as layered pseudo-hierarchical functional annotation scheme, which requires the manual labeling of only the temporal domains of the functional units). More specifically, each syllable is annotated with its syllable boundaries, tone categories of the current and adjacent tones, focus / stress status, and associated sentential modality (see Figure 5. Overall, this version is characterized by its use of a functional annotation scheme, and training from the annotated data to obtain the parameter values, making it more explicitly a standard machine learning approach that encodes various types of input features. In this respect, it is unlike prosodic modeling approaches typical in previous literature. It also has the ability to predicatively generate synthesized speech melody on unseen data, given that the test data are also annotated in this set of input feature labels. The authors argue that this set of annotation is much less labor intensive than traditional frameworks [].

[Xu, Lee, Prom-on and Liu in press] comment on the advantage of such a learning model using the example of how to learn tone sandhi in this framework (below). This is an expected effect of the model if we consider it from a machine learning perspective:

...tone sandhi (Chen 2000). For example, the Mandarin Tone 3 is changed to T2 when followed by another Tone 3. With PENTATrainer2 this rule can be operationalized as the result of an interaction between two functions: lexical tonal contrast and boundary-

³Downloaded from <http://www.homepages.ucl.ac.uk/~uclyyix/PENTATrainer1/>.

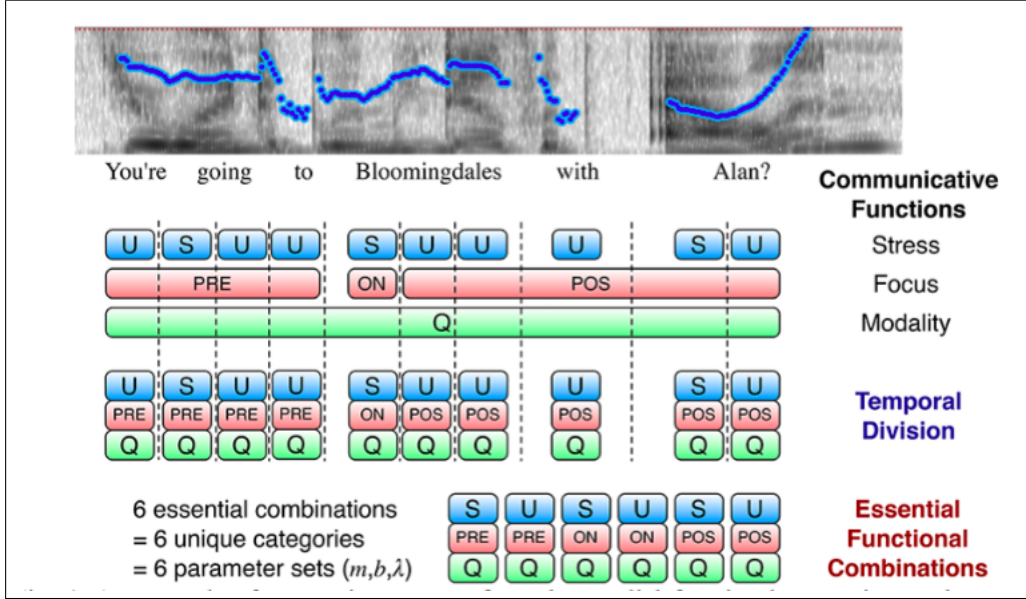


Figure 5: PENTATrainer 2 annotation scheme

marking. That is, the pitch target to be implemented in articulation is jointly determined by the morphemic tone of the current syllable, the morphemic tone of the next syllable, and by the strength of the boundary between the two syllables. Such functional interaction may allow T3 to develop a pitch target variant that happens to be similar to that of another tone, e.g., T2. But the two do not need to be identical, since the functional combinations are not the same. As found in Xu and Prom-on (2014), the best modelling result was obtained when the sandhi T3 was allowed to learn its own target, rather than when it was forced to use the T2 target. This result is consistent with the empirical finding of subtle yet consistent differences between the original and sandhi-derived T2 in Mandarin (Peng 2000, Xu 1997). Thus the obligatoriness of associating a unique target to each functional combination may have led to the development of tone sandhi in the first place. But further research along this line is needed.”

4.3 Evaluation data set

[Liu et al 2006] This paper uses the B-Spline coefficients plus some acoustic features to train decision tress to do question/statement classification from

prosody. This is interesting since it uses parameter based features. Depending on the specifics of B-Spline, it may shed light on our quest for using parameter/coefficients based features in classification or (unsupervised) clustering, which may differ. One interesting point, already discussed in Xu and Promon 2014, is that the issue of data set. For instance, in this paper the author analyzed global prosody by using a data set where tone effects are removed, in which all tones in a sentence are first tone. This kind of design is indeed very helpful in understanding the behavior of the tone distortions, and its good to think about (1) how I can manage the data in our research and (2) how this issue would matter in our un-supervised framework and a broader data mining framework (not the linguistic framework).

5 Supervised learning of Mandarin tones

Supervised learning is the predominant approach in Mandarin tone learning research. A successful unsupervised learning framework must learn from the supervised learning literature. Until recently, tone information is not utilized in commercial speech recognition systems. Recent years have seen the renewed interest in improving tone recognition using context-dependent models in supervised learning frameworks. These improvements include better understanding of the features used in tone recognition, and the effort to utilize contextual information to recover underlying tone targets. In Section 5.3, we also review works that attempts to recover underlying tone targets by identifying the most relevant "nucleus" region in the syllable F0 contour.

5.1 Feature representation and selection

Intuitively, F0 contour is the most and only relevant feature of tones. However, the problem of using F0 features, as discussed above, lies in its variability in running speech. To improve tone recognition accuracy beyond the constraints of using F0, there has been many efforts to identify and develop other useful features that can be extracted from the speech signal.

[Surendran 2007] concentrated on solving this problem by carefully and incrementally testing and (creatively) identifying effective features in all dimensions. In this dissertation, the author conducted hundreds of experiments on a variety of datasets of broadcast speech to determine the effectiveness of a set of 68 features involving pitch, duration, and overall intensity. The results suggest that (1) modifying the pitch and intensity of a syllable based on its neighbors was useful. In particular, pitch normalization by subtracting

the mean pitch of the preceding syllable. (2) Among the twenty voice quality measures used in tone recognition, energy in various frequency bands was the most useful. (3) Further experiments determined a set of 60 band energy features that greatly aided the recognition of low and neutral tones. However, the recall for low tones remained below fifty percent. (4) Context knowing the tones of surrounding syllables did not help as much one would have expected, suggesting our features are already capturing a lot of contextual information. (5) Stronger syllables (such as focus) were easier to recognize in lab speech, but the effect was much less for broadcast speech.

In another doctoral dissertation, [Yu 2011] investigated a similar problem by asking how is tone learned (by machines) and acquired (by humans) from the speech signal through the available information in various phonological spaces (such as F0 and voice quality). The author concentrated her investigation on Cantonese but also included a variety of tone language data such as Mandarin. The results show evidence from human perceptual experiments and computational modeling: (i) motivating a temporal domain from the speech signal for tonal maps beyond the span of a single syllable, and (ii) demonstrating that voice source parameters beyond f0 must be included for characterizing phonetic spaces for tonal maps in a wide range of languages. To understand the sampling requirement of tones from both experimental and computational approaches, the author concluded that human listeners identify tones degraded to be coarsely sampled at a comparable level of accuracy to that for intact tones in Cantonese, and classification by machine with acoustic parameter spaces defined only over a few real values shows a near partition of the phonetic space in the sample of languages studied.

The correlates of intensity to tones have been demonstrated in early works of speech experiments. In a series of intriguingly designed tone perception judgment experiments, [Xu and Whalen 1992] tests the information presents in the amplitude contours and in brief segments of Mandarin tones. In the first two experiments, the researchers used a signal-correlated noise (by adding samples with flipped signs to the original samples such that the amplitude is unchanged but the F0 and formant structure information is removed) to obtain the amplitude contours of the tones without retaining information of F0 and formants. The results showed that Mandarin speakers are able to identify tones with high accuracy using only amplitude contour information (although later it was shown that amplitude values correlate highly with absolute F0 values for tone 2,3, and 4). In experiment 3 and 4, the authors extracted brief tone segments of variable length using a hamming window, and tested which the accuracy of tone identification at each position along the tone contour (e.g., onset at 0ms,20ms,40ms,etc., from the beginning of the tone contour).

The result suggests (somewhat unsurprisingly) that tone 2 and tone 4 are identified with better accuracy when movements of the segments are similar to their respective movement trajectories (i.e., rising for tone 2 and falling for tone 4). For tone 1 and tone 3, the listeners identified more accurately when there are little pitch movement, using differences of absolute F0 (lower sounding pitch judged as tone 3). This indicates the pitch register effect of the tone perception, which is largely unexplored in previous research.

5.2 Context-dependent modeling

As mentioned above, recent works have largely distinguished local from broad context in the context-dependent investigation of F0 variability. In practice, while tone recognition literature conceptually identify those two types of contexts, in general the two are combined to work together inside a single machine learning framework. Therefore I discuss the context-dependent frameworks in tone recognition without separating these two types into different sections.

[Wang and Seneff 2000] investigates the improvement of tone recognition using contextual information in a Mandarin spoken digit recognition application. The authors focus on two aspects of the contextual intonational effect (one broad and one local context): first, F0 down drift during the course of an utterance (sentence); Second, the distortion of F0 frequency height and slope according to different tone combination contexts (i.e., preceding and following tones). To address the first problem, a linear model was built for sentential down drift and a value is subtracted from the observed F0 values. To take into account the contextual information for a context-independent (CI) system, the proposed algorithm adjusts the observed F0 contours based on a comparison of tone combination statistics between the CI and a context-dependent (CD) system with the focus on two parameters: F0 frequency and slope of the contour. A CI model is then trained from the CD adjusted F0 values. This system has an overall error reduction rate of 26.1% from the base system.

Similarly, [Levow 2005] incorporates both local (left and right) context and broader context features in tone recognition with linear kernel SVM, while paying attention to the individual feature sets. The local context is encoded in two types of features: difference and expanded features. Both features have sub-features that encode left or right contexts.

The first set of features (difference features) correspond to differences between the current syllable and its preceding and following syllables. They include difference between pitch maxima, pitch means, pitch at the midpoint of the syllable, pitch slopes, intensity maxima, and intensity means. The second set of features, known as extended syllable features, are simply the last

pitch values from the end of the preceding syllable and the first from the beginning of the following syllable, as well as the pitch maxima and means of these adjacent syllables.

The context dependent features are found to consistently outperform context independent features. The result of additional contrastive experiments suggest that left tone context is much more important than right context in tone category identification. In fact, the right context is shown to decrease the performance of the classifier. This result is consistent with speech perception/production experimental results from previous work [Xu 2001] that the tone co-articulation effect is asymmetric.

The broader context feature mainly includes the F0 compensation for down drift (similar to [Wang and Seneff 2000]). The author used the median slope per syllable across the corpus as phrase-based falling contour compensation. As [Wang and Seneff 2000] did, [Levow 2005] find the re-estimation based on individual phrase slope overfit to the specific tone configuration reduced accuracy. In the phrase based feature representation, each pitch value is thus replaced with an estimate of the pitch value with downdrifting removed, by adding back the estimated pitch drop to pitch values later in the phrase. The result showed improvements in classification accuracy. The author commented that since the phrase segmentation employed here was very simple, it is expected that more nuanced approach with finer grained phrase boundary and possibly phrase accent detection would likely yield greater benefit.

[Wang and Levow 2011] proposed a tone recognition approach that employs linear chain Conditional Random Fields (CRF) to model tone variation due to intonation effects. Three linear chain CRFs are built, aimed at modeling intonation effects at phrase, sentence-and story-level boundaries. All linear-chain CRFs are found to outperform the baseline unigram model, and the biggest improvement is found in recognizing 3rd tones (4%) in overall accuracy. In particular, Phrase Bigram CRFs show a 39% improvement in recognizing 3rd tones located at initial boundaries. This improvement shows that the position specific modeling of initial tones in bigram CRFs captures the intonation effects better than the baseline unigram model.

Looking specifically at broader context, [Surendran, Levow and Xu 2005] exploits the focus conditions to improve tone recognition. Pre-,post-, in-focus, and no-focus conditions are distinguished. The experiment with known focus labels found that pre-focus and no-focus behave similarly in terms of tone recognition error rate, with post-focus having the largest error rates. Meanwhile, on-focus syllables are the easiest to recognize with minimal error rates. Overall, by training and testing SVMs conditioned upon different focus condition groups, the classification error rate reduced 42.9% comparing to the

Table 2: Tone recognition using focus (adapted from [Surendran et al 2005])

Condition	Error Rates
Combined: not using focus (baseline)	15.16%
No-focus syllables	7.74%
Pre-focus syllables	7.74%
In-focus syllables	0.80%
Post-focus syllables	18.37%
Combined: Conditional on correct focus	8.66%

baseline, where no focus-group is identified. However, in this experiment, the focus labels had to be manually annotated and it was available on both training and testing sets. This is a unrealistic scenario in real applications. Next, the researchers conducted experiments to incrementally reduce the requirement on manual labeling. The second experiment assumes focus labels are only known during training, and uses F0 and intensity based features to predict the focus conditions on testing data set. The third experiment assumes that correct focus label is not available at all. In this experiment, the focus labels for both training and testing data are predicted from the confidence rating on the tone recognition algorithm without using focus information. It is observed and hypothesized that the tones classified with the highest confidence score is the location of the focus. In both of these subsequent experiments using predicted focus label, it is interesting to observe that even though the prediction errors were high on the focus label (more than 30%), the error reduction of the ultimate tone recognizer is still comparable with the first experiment where the correct focus label is known (error rate below 10%). The authors attribute this to the similar behavior of the tone classifier on pre- and no-focus conditions, where most of the confusion happens in the focus label prediction. Table 2 shows the summary of results in this study assuming known focus.

[Liu et al 2006] uses the B-Spline coefficients ⁴ plus some acoustic features to train decision trees to do question/statement classification from intonation contours in Mandarin (referred to as modality of the sentence in the literature), using a highly controlled experimental dataset. For 10-syllable utterances, the highest correct classification rate (85%) is achieved when normalized (to remove the effects of speaker, tone, and focus) final F0s of the 7th and the last

⁴Informally, B-Spline is a type of piecewise polynomial that functionally approximate the intonation curve.

syllables are included in the tree construction. The results confirm the previous finding that the difference between statement and question intonations in Mandarin is manifested by an increasing departure from a common starting point toward the end of the sentence. Meanwhile, this paper also raises the question regarding the validity of using compact model coefficients to represent F0 contours in supervised and unsupervised learning (as other works such as [Zhang 2015] have found that in polynomial and qTA coefficients do not perform well in unsupervised learning for Mandarin tone clustering), and how to effectively use such representations.

5.3 Tone nucleus region modeling

Other tone recognition researchers have sought to identify the most relevant regions in a syllable for the F0 contour of a tone. Such regions is hypothesized to better approximate the true underlying tone target. This is in part motivated by tone production models such as PENTA[] and StemML[], which assume that carryover coarticulation dominates tone realization and thus the true tone is more closely approximated in the latter half of the syllable. [Sun 2002] used the pitch at the midpoint of the syllable and fit the pitch contour from the midpoint to the end of the syllable for pitch accent recognition, effectively a temporal segmentation. Subsequently, [Zhang and Hirose 2004] proposed a model which successfully identifies tone nucleus regions for canonical tone production. The tone region is segmented by k- means clustering of pitch contour units; the nucleus itself is identified based on features including segmental time and energy.

[Zhang, Hirose, and Nakamura 2005] presents three ways of modeling the contextual distortion of the F0 contours, including Tone Nucleus modeling, anchoring based F0 normalization, and Hypo and Hyper articulation F0 model. The result of improvement over the tone recognition accuracy is significant.

[Wang and Levow 2006] proposes a strategy to identify nucleus regions of tones using the amplitude and pitch plot segmentation with computational geometry techniques. Given a syllable segmentation, this approach employs amplitude and pitch information to generate an improved sub-syllable segmentation and feature representation. This sub-syllable segmentation is derived from the convex hull of the amplitude-pitch plot. This approach achieves a 15% improvement using the said segmentation strategy over a simple time-only segmentation.

6 Unsupervised learning of Mandarin tones

There has been limited effort on the unsupervised learning of tones. It is well known that most supervised learning framework must rely on a large amount of manual annotation effort that is costly in time and money [Levow 2006b]. This annotation bottleneck as well as a theoretical interest in the learning of tone motivates the use of unsupervised or semi-supervised approaches to tone recognition whereby the reliance on this often scarce resource can be reduced [Levow 2006a]. Another motivation to explore unsupervised learning is to understand the process of language acquisition, as child learners must identify these linguistic categories without explicit instruction by observation of natural language interaction [Levow 2006b]. As such, the goal of such unsupervised learning frameworks is to improve the accuracy of tone learning algorithms with minimum supervision and human labeled data.

Another dimension of unsupervised learning of tones is the data mining of speech prosody in a large spoken or intonation corpus [Raskinis and Kazlauskiene 2013]. The goal of this endeavor, in contrast, lies in the improved understanding of speech intonation and tone contour patterns through unsupervised data mining and pattern discovery algorithms using a large quantity of real speech data. Such patterns will also shed light on the nature of variability of tone in spontaneous speech and provide better features and context-dependent models for tone recognition modeling. Previous works [Zhang 2015] have shown that this type of data mining task can be more efficiently carried out by viewing intonation contour as a time series and employing techniques from the subfield of time-series data mining.

In the next two sections I first review previous works on unsupervised learning of Mandarin tones from the standard speech / tone recognition literature. Then I review the relevant works, concepts and key issues in time-series mining and how it might relate to speech intonation mining. Next, I discuss the important distance measures and algorithms used in both general unsupervised learning and time-series mining. Finally, I explore issues associated with the design and selection of the evaluation data set.

6.1 Unsupervised and semi-supervised learning of Mandarin tones

Some preliminary unsupervised work by [Gauthier et al 2005, 2007] employs self-organizing map⁵ [] and measures of f0 velocity for tone learning. [Gauthier et al 2007] used a raw 30-point pitch vector and the first derivatives (D1) of the F0 values as feature vectors on some 2000 observations of tone contours. In particular, they found that the D1 feature vectors yielded an almost perfect result in classifying unseen stimuli, an improvement over using raw 30-point F0 values. This shows the internal structure and intrinsic pattern that can be exploited in tone learning. Meanwhile, since this study did not use spontaneous speech data set, it is yet to be seen how it performs on more challenging data⁶.

[Levow 2006a, 2006b] concentrated on the problem of unsupervised learning and semi-supervised learning in Mandarin tone recognition. [Levow 2006a] employed asymmetric k-lines clustering [], a spectral clustering algorithm, as the primary unsupervised learning approach. Rather than assuming that all clusters are uniform and spherical, this approach enhances clustering effectiveness when clusters may not be spherical and may vary in size and shape. The author argues that this flexibility yields a good match to the structure of Mandarin tone data where both shape and size of clusters vary across tones. A comparison is made between k-means clustering, symmetric k-lines clustering [18], and Laplacian Eigenmaps [17] with k-lines clustering.

This algorithm is evaluated on a clean read data set and a spontaneous broadcast news data set. Results show that in all cases, accuracy based on the asymmetric clustering is significantly better than most common class assignment and in some cases approaches 96% of labelled classification accuracy. The best results are achieved on the clean focused syllables, reaching 87% accuracy. For combined in-focus and pre-focus syllables, this rate drops to 77%. These rates contrast with 99-93% accuracies in supervised classification using linear SVM classifiers with several thousand labelled training examples. On broadcast news audio, accuracy for Mandarin reaches 57% (much greater than the 25% chance level), comparing with the 72% accuracy achieved using supervised linear SVMs with 600 labeled training examples. Table 3 summarizes

⁵A self-organizing map (SOM) is a type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map (from Wikipedia).

⁶Chris Kirov and me have performed an initial trial experiment on a small sample of spontaneous speech data, which did not show good results. However, it is yet to be evaluated on a more extensive real data set.

Table 3: Tone recognition with unsupervised and supervised learning (adapted and modified from [Levow 2006a, 2006b])

Condition	Unsup.	Supervised	Semi-supervised
Lab, In-focus	87%	99%	94%
Lab, Pre and In-focus	77%	93%	n/a
Broadcast News	78%	81.3%	70%

the results using the unsupervised vs. supervised approaches.

Contrastive experiments of different clustering algorithms showed that the asymmetric k-lines clustering approach consistently outperforms the corresponding symmetric clustering learner, as well as Laplacian Eigenmaps with binary weights for English pitch accent classification (shown in Figure 6). To the author’s surprise, k-means clustering outperforms all of the other approaches when producing 3-14 clusters.⁷ Accuracy for the optimal choice of clusters and parameters is comparable for asymmetric k-lines clustering and k-means, and somewhat better than all other techniques considered. The author attributes this similar performance to careful feature selection process for tone and pitch accent modeling. It was reported that for the four tone classification task in Mandarin using two stage clustering, the best clustering using asymmetric k-lines strongly outperforms k-means, at 87% and 74.75% accuracy respectively.

The feature set used in this study is well informed by previous work on tone learning and tone production. It adopted the common practice in supervised tone learning in using multiple types of features beyond the F0, and features that reflect the context-dependent nature of tone contour shapes. The basic features include F0 features (five equidistant points sampled from the F0 contour of the syllable nucleus, and the mean F0 feature) and intensity features (both are normalized per speaker and log scaled). The final region of each syllable is identified as the nucleus region. To account for co-articulation effects, nucleus region slope features are computed according to qTA’s assumptions []. These are further log-scaled and normalized to compensate for greater speeds of pitch fall than pitch rise[15]. Figure 7 shows the well-separatedness of the four tones in the read speech data set in terms of slope vs. height of its pitch target. Overall, while this yields valuable evidence for feature experimenta-

⁷In fact, it has been shown in time-series mining literature [] that k-means and Euclidean distance are extremely powerful techniques as the the data gets bigger, despite their simplicity.

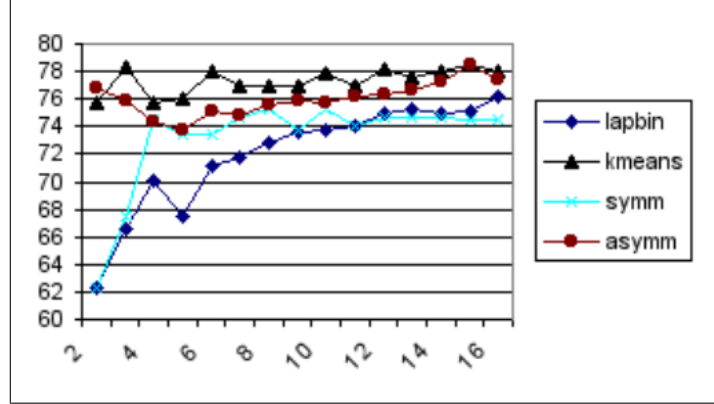


Figure 6: Comparison of different clustering algorithms with varying number of clusters and clustering accuracy

tion, it is doubtful that this pattern can be generalized to the spontaneous speech data.

6.2 Semi-supervised learning of Mandarin tones

[Levow 2006b] further explored semi-supervised learning for this task, using the Manifold Regularization framework [Belkin et al., 2004]. This framework postulates an underlying intrinsic distribution on a low dimensional manifold for data with an observed, ambient distribution that may be in a higher dimensional space with pairwise distances preserved. This paper uses Laplacian Support Vector Machines, A semi-supervised classification algorithm, which allows training and classification based on both labeled and unlabeled training examples. For each Mandarin data set, for each class, the model uses a small set (40) of labeled training instances in conjunction with 60 unlabeled instances, and tests on 40 instances.

The semi-supervised classifier achieved comparable results with the unsupervised algorithm (see Table 3). Interestingly, the semi-supervised classifier also reliably outperforms an SVM classifier with an RBF kernel trained on the same labeled training instances.

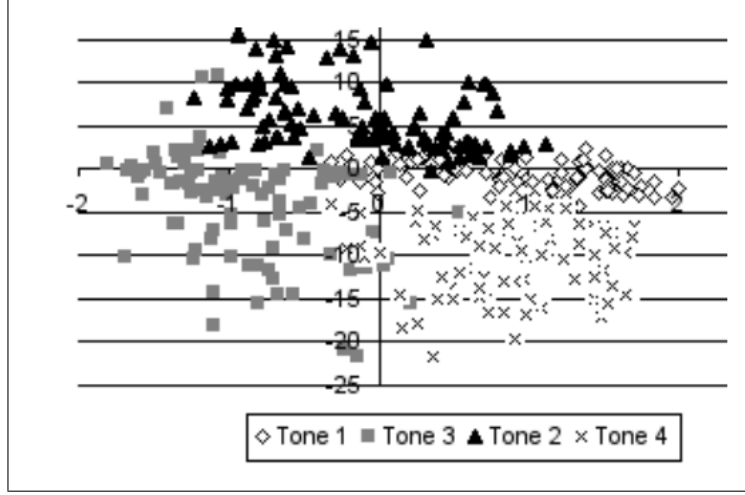


Figure 7: Cluster separation of pitch target slope (x) vs height (y) in Mandarin read speech data set

7 Speech prosody mining using time-series mining techniques

Previous works in corpus based intonation research [Raskinis and Kazlauskiene 2013, Zhang 2015] have shown the challenges of data mining and clustering of speech intonation data in a large spoken / intonation corpus. At the core of this task is the expensive computation of a large amount of high-dimensional pairwise distance (especially with the widely preferred Dynamic Time Warping or DTW distance measure for time-series data) to obtain the distance matrix, and to find the most effective low-dimension feature representation for the F0 time-series data that faithfully preserve the true distances among objects with increased efficiency for storage.

[Zhang 2015] showed the potential of employing time-series data mining techniques to solve these problems, including techniques to speed up DTW distance computation [] and to transform feature representations for both dimensionality reduction and efficiency of computation and storage []. In fact, F0 contour data can be naturally viewed as time-series data with F0 on the y-axis and time on the x-axis. Time-series mining has been successfully applied in a number of fields that deal with various kinds of time-series, including (the most relevant for our current purposes) Music Information Retrieval (MIR), where F0 contour data from music signal are indexed and searched, and meaningful patterns discovered [].

7.1 Overview of time-series mining

Formally, a time series $T = t_1, \dots, t_p$ is an ordered set of p real-valued variables, where t_i is the time index. Time-series mining deals specifically with the data mining tasks with time-series data. [Lin et al 2007] outlined the main tasks that time-series mining research is concerned with:

(1) Indexing: Given a query time series Q , and some similarity/dissimilarity measure $D(Q, C)$, find the most similar time series in database DB [2, 11, 20, 30, 63].

(2) Clustering: Find natural groupings of the time series in database DB under some similarity/dissimilarity measure $D(Q, C)$ [29, 36].

(3) Classification: Given an unlabeled time series Q , assign it to one of two or more predefined classes [24].

(4) Summarization: Given a time series Q containing n datapoints where n is an extremely large number, create a (possibly graphic) approximation of Q which retains its essential features but fits on a single page, computer screen, executive summary etc [43].

(5) Anomaly Detection: Given a time series Q , and some model of normal behavior, find all sections of Q which contain anomalies or surprising/interesting/unexpected/novel behavior [14, 34, 54].

Due to the typical large size of data mining tasks and the high dimensionality of time-series, a generic time-series mining framework is as follows: [20] (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 (hopefully very few) accesses to the original data on disk to confirm the solution obtained in Step 2, or to modify the solution so it agrees with the solution we would have obtained on the original data.

However, in practice, as will be discussed below, 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. The distance measure also needs to effectively capture the true (and meaningful) distances among objects, that also allows reasonably efficient computation (tractable) often by using special techniques to prune off impossible candidates. In the next sections I discuss some of the most relevant key issues in these areas.

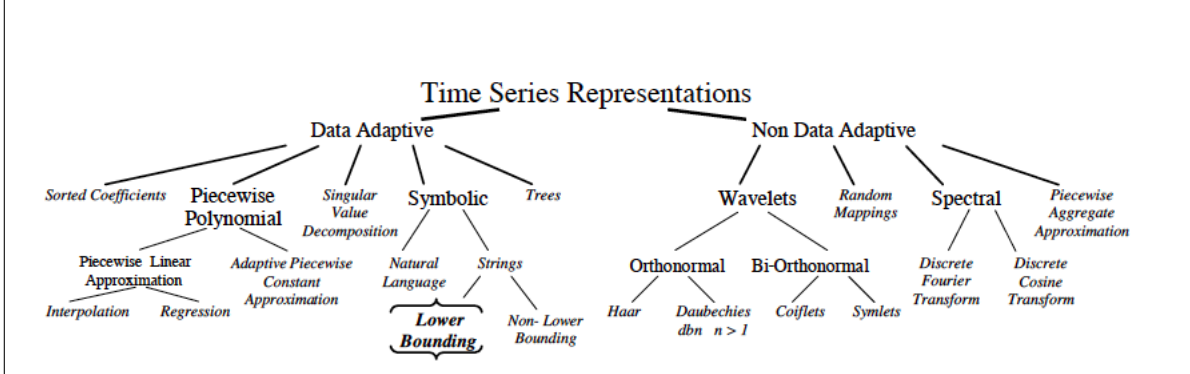


Figure 8: Types of Time-series Data Representations (adapted from Lin 2010)

7.2 Time-series and F_0 feature representation

In Section 4.2.1, I have reviewed prosodic models and their relevance to feature representation of F_0 contours in speech. Meanwhile, a great number of representations have been proposed for time-series mining, including both real-valued (such as DFT, DWT) and symbolic representations (Such as SAX). Figure 8 illustrates the most commonly used representations [Lin 2010].

In the next sections I consider both types of feature representations and discuss their strength and limitations.

7.2.1 Prosodic modeling feature representation

(1) **Polynomial Regression.** The most straightforward way to model F_0 contour curves is to use polynomial functions to fit the F_0 contour of each utterance. A F_0 contour can thus be represented by the coefficient vector $[c_1, c_2, \dots, c_{n+1}]$ of a n -th order polynomial. This has been done in a number of studies for tone and intonation [Hirst et al., 2000; Liu et al., 2006, and many others]. Alternatively, one could use a spline function, a piece-wise polynomial to model different sections of a complex contour. This approach greatly reduces the dimensionality of the original F_0 contour. However, [Xu 2011] points out that the critical question about polynomial representations is that whether they are linguistically meaningful, and whether they can be used in predictive modeling, i.e., serving as categorical parameters that can be generalized to other instances of the same category. This has not been evaluated in a predictive synthesis context. [Zhang 2015] experimented with the third degree polynomial representation (a 4d vector) of F_0 contours in a clustering task with a clean read data set of Mandarin tones. The conclusion suggests that the clustering accuracy is low.

(2) **quantitative Target Approximation.** Given that qTA model (see section 4.2.3) has been shown to perform well in producing curves that closely resemble real tone contours in connected speech, an important question to be asked is: do qTA parameters perform well to reflect the similarities between tones in perception? In other words, we want to make sure that qTA parameters have the property where perceptually similar tone contour shapes also have similar parameter values. In [Zhang 2015]’s experiments the results suggest a negative answer to this question.

(3) **Raw Time-series F0 Vector.** [Gauthier et al 2007] showed that unsupervised classification using Self-Organizing Map (Neural Network) yielded a nearly 80% correct result when time-series are represented with a 30-point raw F0 vector.

In [Zhang 2015], in order to find the best method of working with F0 vectors, several transformations are made from the raw F0 vectors, including un-normalized and normalized. For each of these, in order to avoid the logarithmic behavior of Hertz unit, three versions are created, using Hertz, Bark, and Cent scale representations, giving rise to a total of 6 types of feature vectors. The conversion from Hertz to Bark and Cent are computed as below:

$$F_{CENT} = 1200 * \log_2 \frac{F_{HZ}}{F_{REF}} \quad (4)$$

where the F_{REF} is the reference frequency, corresponding to the minimum F0 in the computation of the pitch track (set to 55Hz).

$$F_{BK} = 7 * \log[F_{HZ}/650 + ([1 + (F_{HZ}/650)^2]^{1/2})] \quad (5)$$

In a series of clustering experiments, [Zhang 2015] found that (1) normalized contours yields higher accuracy than raw F0 contours; (2) the use of Hertz, Bark, or Cent scales did not have significant differences in the results; (3) the F0 vector achieves much higher accuracy with the DTW distance measure than Euclidean distance measure.

(4) **First Derivative Vector (D1).** The discrete first derivative feature (D1) is obtained simply by taking the first derivative of the original signal of the tone contour, and downsampled to 30 point:

$$D1 = 0.5 * (F_0(t + 1) - F_0(t - 1)) \quad (6)$$

for all timestamps t .

Intuitively, the D1 feature captures the movement of the pitch trajectory at each timestamp. It also serves as a normalization strategy where the differences in pitch height among different speakers are removed. Otherwise, the D1

does not reduce the dimensionality of the time series, nor does it create new abstract features as a combination of other features. As a first glance, it is not a fundamentally different transformation over the F0 feature. Surprisingly, [Gauthier et al 2007] showed a near-perfect performance using the D1 feature in a classification task with Self-Organizing Map. In the same experiment, the F0 feature performs around 20% lower than the D1 feature. In [Zhang 2015], results suggest that D1 feature also achieves superior performance in unsupervised clustering tasks even when paired with Euclidean distance, in contrast to the F0 vector.

7.2.2 Time-series symbolic representation

[Lin et al 2007] points out the limitations in data mining algorithms for real-valued time-series representations (such as DFT and DWT). For example, in anomaly detection we cannot meaningfully define the probability of observing any particular set of wavelet coefficients, since the probability of observing any real number is zero [38]. Such limitations have led researchers to consider using a symbolic representation of time series.

There has been many symbolic representations proposed for time-series [3, 27]. However, none of the techniques allows a distance measure that lower bounds a distance measure defined on the original time series. This constitutes a problem for the generic time-series mining framework discussed above, since the approximate solution to problem created in main memory may be arbitrarily dissimilar to the true solution that would have been obtained on the original data. A symbolic approach that allows lower bounding of the true distance would not only satisfy the requirement of the generic framework, but also enables us to use a variety of algorithms and data structures which are only defined for discrete data, including hashing, Markov models, and suffix trees [Lin et al 2007]. Symbolic Aggregation approXimation (or SAX) [Lin 2010] 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 related MINDIST distance function for SAX is discussed in Section 7.3.

The SAX representation transforms the pitch contour into a symbolic representation using Piecewise Aggregate Approximation technique (PAA, Figure 12), with a user-designated length (w =desired length of the feature vector) and 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 (Figure 10). Here, the Gaussian distribution is used to obtain the break-points for vertical pitch space so that each region (represented by a symbol)

is equiprobable (probability of that symbol is given by the integration of the area under the Gaussian curve to as defined by the break points). This ensures that the probability of a segment being assigned any symbol is the same. Figure 10 shows an example [Lin et al 2007] of SAX transformation of a time-series of length 128. It is discretized by first obtaining a PAA approximation and then using predetermined breakpoints to map the PAA coefficients into SAX symbols.

SAX has been evaluated in various classic time-series mining tasks and shown to work well in numerous applications to mine data from a variety of fields such as bioinformatics, finance, telemedicine, audio and image signal processing, and network traffic analysis. In particular, it has been shown to preserve meaningful information from the original data and produce competitive results for classifying and clustering time series.

SAX has been less explored in the domain of audio signal F0 pattern mining. [Valero, Salamon and Gomez 2015] experimented with SAX representation for the computation of F0 contour similarity from music audio signals in a Query-By-Humming (QBH) task⁸ in the context of Music Information Retrieval (MIR). However, results suggest that SAX does not perform well for musical time-series data in the context of QBH. The authors attribute this to the fact that SAX does not consider any particularities the origin domain of the time series may have - in this case, the musical notes. Thus, in the case of QBH, SAX may be abstracting away key musically-related information from the melodic contours required for properly performing the alignment [of F0 contour pairs].

[Zhang 2014, 2015a, 2015b] showed the effectiveness of SAX in mining speech or speech-like F0 contours - tones, where fine, noisy details of the F0 contours do not carry substantial crucial information comparing to its global shape. In [Zhang 2014], musical F0 contours from Beijing opera singing are converted to SAX representations in order to compare its similarity to linguistic tones of its lyrics. In a manually constructed data set consisted of balanced F0 contour shapes from all four tones, four types of evaluation measures showed that SAX representation faithfully preserves the original clusters. Similarly, [Zhang 2015b] experimented with the SAX parameters by iterating through different combinations of w and a to find the best correlation with the perceived pitch relationships in pairwise contour analysis. In [Zhang 2015a], SAX is shown to perform significantly better than raw F0-based contour features when subjected to the K-means clustering algorithm (see details in Section

⁸In 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.

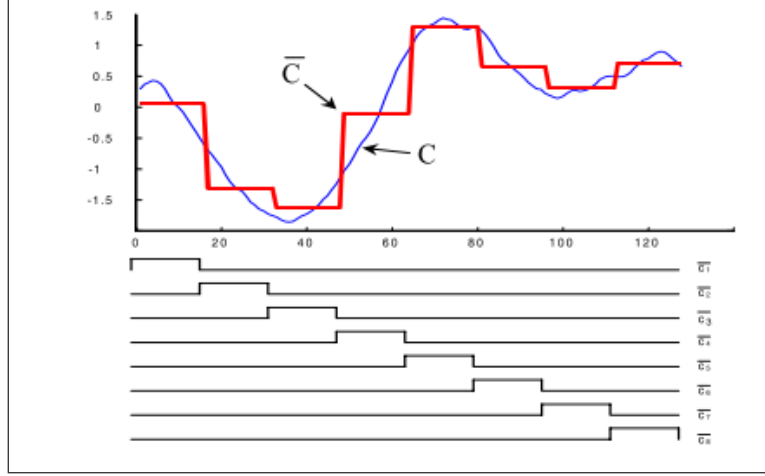


Figure 9: Piecewise Aggregate Approximation. The PAA representation can be visualized as an attempt to model a time series with a linear combination of box basis functions. In this case, a sequence of length 128 is reduced to 8 dimensions (adpated from [Lin et al 2007])

7.4). This result is similar to the experiment on the Space Shuttle telemetry data set considered by [Lin et al 2003], where the SAX outperformed original data due to the possible smoothing effect of SAX’s dimensionality reduction. In this context, contrary to the MIR melodic retrieval, the ”abstracting” of SAX becomes a strength of the algorithm.

7.2.3 Time-series normalization

Many literature reviews [?] 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 (in fact this is a general strategy that is also used in mining non-time-series data):

$$z = (x_i - \mu)/\sigma \quad (7)$$

where μ is the mean of the contour and σ is the standard deviation.

However, [Zhang 2015] observed that speech tone time-series data has a special property so that the z-score transformation would seriously distort the shapes of the tones in the normalized corpus. Essentially this is caused by the presence of many flat or near flat contours (such as in level tones). 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

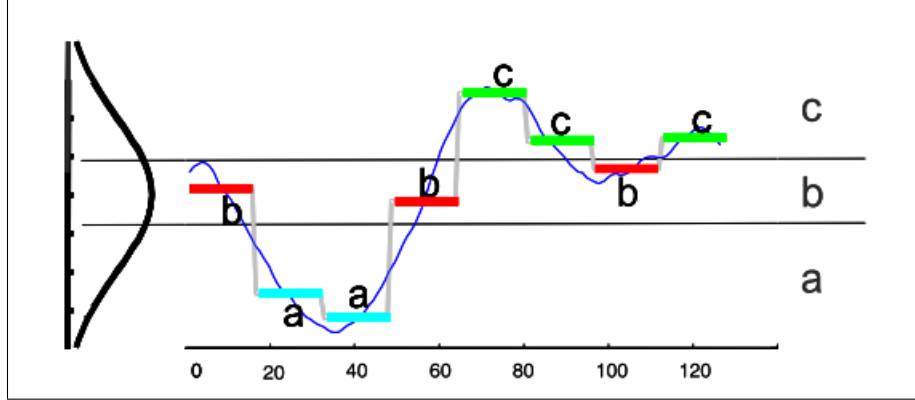


Figure 10: Symbolic Aggregate Approximation, with original length $n = 128$, number of segments $w = 8$ and alphabet size $a = 3$, with output word **baabc-cbc** (adpated from [Lin et al 2007])

in the values of the flat or near flat contours (since each point in this contour is identical and also identical to the mean), and turn such contours into a significantly un-flat contour.

Since normalization is required for a meaningful comparison of time-series patterns, we need to design an alternative strategy for time-series normalization in this task. There are many strategies that fulfills this purpose. After trial and error experimentation, one of the best working measure is the Subtract-Mean normalization strategy (see Equation 8). It effectively retains the original shapes in a normalized corpus of tones.

$$z = (x_i - \mu) \quad (8)$$

This issue also exists in SAX representation since normalized time-series SAX significantly outperforms un-normalized ones [Zhang 2015]. Therefore it is a built-in requirement of SAX to first normalize the time-series using z-score transformation. Fortunately, SAX also has a built-in remedy for this problem: 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.

7.3 Distance measure

(1) **Euclidean distance.** The Euclidean distance is one the most widely used and computationally economic distance measures [Lin 2010], defined as follows on the n -dimensional Euclidean space for a time-series of length n :

$$d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (9)$$

(2) **DTW distance.** In time series analysis, dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed (Figure 11). DTW distance between two time series is computed with dynamic programming, by recursively solving the optimal alignment between two sequences in subproblems, and return the shortest distance between the two (i.e., best alignment) time series. In particular, the optimal path is the path that minimizes the warping cost:

$$DTW(Q, C) = \min \left\{ \sqrt{\sum_{k=1}^K (w_k)^2} \right\} \quad (10)$$

where w_k is the matrix element $(i, j)_k$ that also belongs to k -th element of a warping path W , a contiguous set of matrix elements that represent a mapping between Q and C .

This warping path can be found using standard dynamic programming to evaluate the following recurrence.

$$\gamma(i, j) = d(q_i, c_j) + \min \left\{ \gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1) \right\} \quad (11)$$

where $d(i, j)$ is the distance found in the current cell, and $\gamma(i, j)$ is the cumulative distance of $d(i, j)$ and the minimum cumulative distances from the three adjacent cells.

To make the computation tractable and to prevent pathological warping, many constraints have been proposed to impose upon the possible warping window. One such window is shown in green in Figure 11.

(3) **MINDIST distance function.** The MINDIST function is a distance measure defined for the SAX representation of the time series, which, crucially, have been proved to lower bound the true distances of original data [Lin et al 2003]. It returns the minimum distance computed between two strings by building on the PAA representation distance function and substitute the distance computation with a subroutine of `dist()` function:

$$MINDIST(Q, C) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^w dist(q_i, c_i)^2} \quad (12)$$

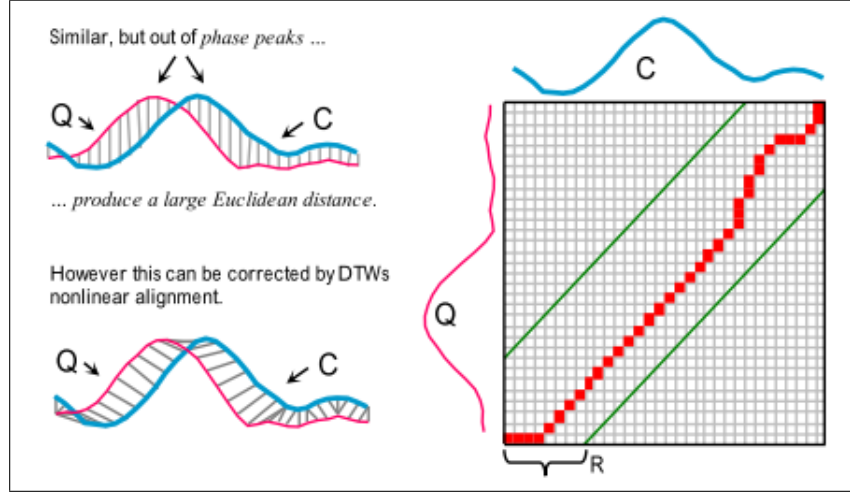


Figure 11: Euclidean distance vs. Dynamic Time Warping: example (adapted from [Rakthanmanon et al 2012])

The `dist()` function can be implemented by search over a lookup table (for details see [Lin et al 2003]). The lower bounding property is an important heuristic for pruning sub-optimal candidate when finding the minimum distance: to avoid expensive computational cost of computing the true distances over a large number of time series, we can first use a SAX approximation to the original time series and compute the MINDIST distance matrix. Since the MINDIST is proved to lower bound the true distance, then we can prune off those whose MINDIST distance is greater than the best so far, since there is no way its true distance would be less than the best minimum distance so far.

7.4 F0 pattern mining using time-series mining techniques

[Koegh] shows that there are many data mining tasks that can be more efficiently solved as a time-series mining task.

[Sankalp et al 2014] applied exact motif (pattern) discovery to the F0 melodic patterns large collection of Indian music recordings every possible by computing similarity between every possible subsequence pair obtained within an audio recording. show the possible applications of the search tool and network visualization.

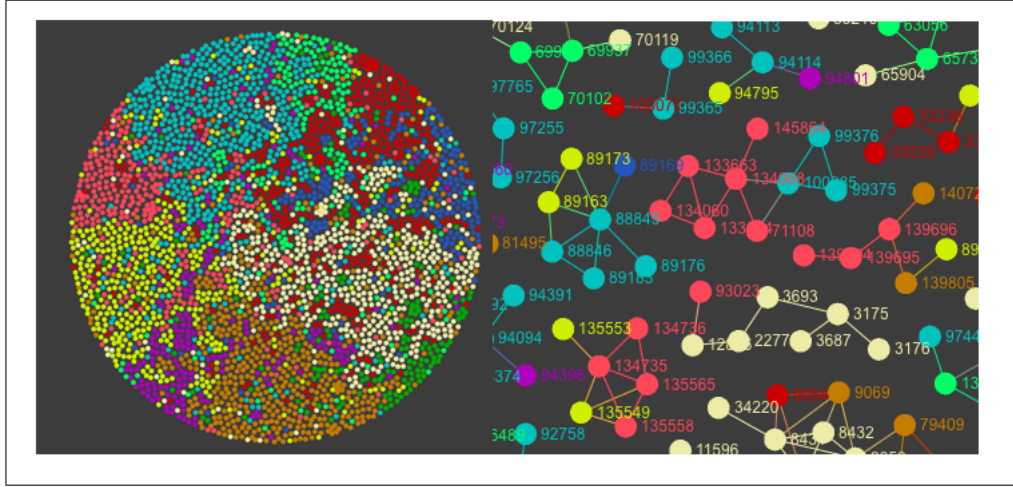


Figure 12: Network visualization of melodic patterns mined from a large collection of Indian music. Left: Global structure. Right: Individual similarity groupings. Each node represents a subsequence (pattern) of melodic contours about 2s long.

8 Chapter outlines