REPRESENTATION AND MACHINE LEARNING OF MANDARIN TONE CATEGORIES IN CONNECTED SPEECH

Shuo Zhang
Department of Linguistics, Georgetown Univerisity
Computational Phonology
May 2014

ABSTRACT

Bayesian models of categorical perception in speech benefits from the assumption that categories in speech sounds can be represented by a mixture of multivariate Gaussian distribution. Such assumptions makes mathematical analysis of posterior distribution more accessible. However, in the modeling of tone categories, the problem of dimensionality challenges us to find an efficient lower-dimension time-series representation that obeys the Gaussian assumption. In this paper, I take the first steps in exploring the properties of the tone categories and the appropriate time-series representations in a corpus of Mandarin connected speech audio recording. Different time-series representations based on models such as polynomial, qTA, and SAX are discussed. Machine learning (clustering and classification) experiments are conducted to evaluate the effectiveness of these representations. The results showed that the performance on the current dataset is below ideal in all measures tested comparing to previous work using SOM. Different degrees of separability are observed between - category. I discuss the limitations and validations to the different models considered.

1. INTRODUCTION

The influence of phonetic categories on speech perception has been well studied in the segmental domain. More specifically, categorical perception has been observed primarily for consonants [3], while vowel perception has been shown to be more continuous [4]. More recent work has studied the perceptual magnet effect on a number of vowels in different languages [5–7]. Categorical phenomena has also been shown in non-speech domains of cognition [8]. In the suprasegmental domain, such as the second language acquisition of a tone language, for instance, it is conceivable that we may expect to find the effect of existing intonational/tonal categorical influences across the production and perception of the target language's prosody, even though few studies have explicitly adopted this framework [11].

Bayesian models offer new insights to the understanding of categorical phenomena in speech perception by computing specific, numerical, and empirically testable predictions of the behavior of the perceptual bias. Feldman et al [9, 10] developed a Bayesian model of perceptual magnet, whose predictions are well matched by behavioral data in English vowels using multidimensional scaling analysis [5]. Here I propose two considerations in extending this model to the suprasegmental domain. First, in this model, phonetic categories (e.g., vowels) are modeled as a mixture of multivariate Gaussian distributions of speech sounds, whose dimensions depend on the dimension of the vectors that define the category (e.g., in the case of vowels, F1-F2 bivariate Gaussian). Despite being a simplified representation of distributions of speech sounds in real-life phonetic categories, this model therefore assumes a conveniently known distribution that makes it possible to compute posterior distributions by straightforward mathematical analysis. However, if a phonetic category cannot be represented efficiently by 2 or 3 features, then we face the problem of high dimensionality that makes it difficult to compute distance between vectors. In those cases we must consider ways of dimensionality reduction.

Second, the original paper focused on perceptual magnet using vowel categories. Nonetheless, in principle there is no components to limit the model to the segmental domain and prevent generalization into prosodic domains (i.e., tone and intonation). As discussed above, the attempt to extend the Feldman et al [9] perceptual-magnet model to tone perception faces two challenges: (1) we must prove that tone vectors that belongs to a single category can be described by a Gaussian distribution; (2)we must find a suitable representation of the time-series vectors of tones that addresses the dimensionality problem, in the mean time reflecting the Gaussian-like distribution property of tone categories.

This paper reports the first steps in exploring the properties of the tone categories and the appropriate time-series representations in a corpus of Mandarin connected speech audio recording. The research questions are defined as follows: (1) Can tone contours in this dataset be clustered / classified into the four tone categories in Mandarin? (2) What is the appropriate time-series representation that addresses both the dimensionality problem and the Gaussian assumption? (3) Can we assume the tone categories are represented with a Gaussian distribution?

The rest of the paper is organized as follows. The next section discusses the methodologies proposed in this pa-

SPRING 2014

per. This includes alternate approaches of time-series representation, and methodologies of clustering using different combinations of representations and vectors. Following the discussion, we proceed to report the procedures in experimenting with these methods, and the results of each procedure.

2. MANDARIN TONE INVENTORY AND VARIABILITY IN CONNECTED SPEECH

In the canonical forms of Mandarin tone system, four tone categories are present: high-level (High tone), rising (Rising tone), low-dipping (Low tone) or falling (Falling tone). Following convention, these are also referred to as tone 1, 2, 3, and 4 in this paper. Figure 1 shows the (time-normalized) F0 contours of five tokens and their means of the four Mandarin tones produced in citation form by a male speaker (data from [18]). As can be seen, when produced in isolation by a single speaker, the tones are well separated even when time-normalized. They become much less separated, however, when spoken in connected speech and when uttered by different speakers. Figure 1(b) and (c) show the means and distributions of the same four tones spoken in connected utterances by three male speakers [18]. Gauthier et al [11] identified two major sources for the extensive overlap between the tones. 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 [16, 18]. Similar variability has been found in other tone languages [12].

To the author's knowledge, most previous works on tone F0 contours have not addressed tone separability and perception in connected speech. In the discussion of the difficulty of tone learning in tone 3 and tone 2, for instance, the majority of literature focus on tone perception in isolation [1,2,16], where the tones are more well separated and bear different properties from tones in connected speech (e.g., in connected speech, tone 3 is realized as a short low tone, whereas in isolation it is pronounced as a low-rising tone with the longest duration among all four tones). Prom-on et al [15] has developed a series of models based on PENTA and qTA, focusing on the production of F0 tone contours in connected speech. However, the models primarily focus on the generation of F0 contours in speech production, and do not in general address perception and identification of tone categories. From a tone perception and acquisition perspective, however, tone perception in connected speech occupies an important place in the understanding of the mechanisms of tone identification, and poses special challenge to both human and machine learning of tone categories from the large amount of variances found in real-life connected speech.

As discussed above, a primary goal in the current paper is to compare and evaluate different ways of time-series representation in order to find feature-vector combinations that is (1)low in dimensionality, (2)effective in clustering tones into categories represented by Gaussian distributions. Gauthier et al [11] used a raw 33-point pitch vector and the first derivatives (D1) of the F0 values as feature vectors on

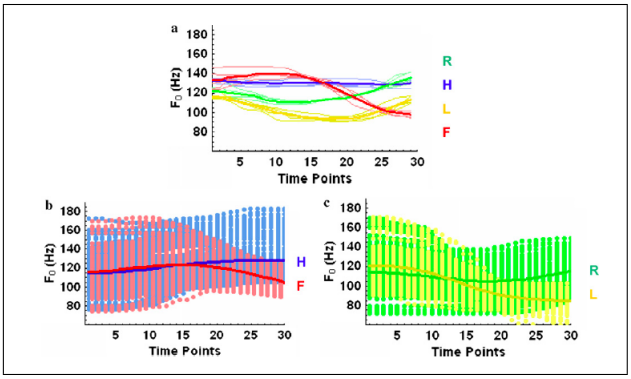

Figure 1. Tones produced (a) in citation form by one speaker and (b, c) in connected speech by three speakers. Thick lines correspond to means while pale background to the distribution of High (blue), Rise (green), Low (yellow) and Fall (red) (data from Xu (1997)).

some 2000 observations of tone contours, and obtained optimal results with Self-Organizing Map(SOM), a type of neural network. In particular, they found that the D1 feature vectors yielded an almost perfect result in clustering and classification tasks. However, no attempt of dimensionality reduction was made. Moreoever, in comparison, the current study is tested on a smaller dataset (around 400 tone contours), which may consitutes a limitation on the results of the study. In the next section, I discuss the four types of time-series representation used in this paper.

3. TIME-SERIES REPRESENTATION

3.1 Polynomial Fitting

Suppose that the surface F0 contour can be represented by a polynomial in the form of:

$$y = a + bx + cx^2 + ... + mx^n \quad (1)$$

In this representation, we consider the representation of a pitch contour vector (with a dimension from 6 to 30) using a vector of polynomial parameters. Specifically, we fit extracted F0 contour pitch vectors of different length to a third-degree polynomial in the following form, represented by four parameters a, b, c, and d:

$$y = a + bx + cx^2 + dx^3 \quad (2)$$

The choice of a third-degree polynomial is motivated by (1) the property of F0 curves of tones, which is asymmetric and non-linear; (2) the small number of dimensions and the efficiency of representating a variety of shapes. The polynomial parameters for a 4-d vector (four parameters of the polynomial) on time-normalized pitch contours, as well as a 5-d vector with the duration of the given contour (four parameters of polynomial plus a duration parameter) group for clustering and classification.

3.2 quantitative Target Approximation

The qTA parameter representation is based on the quantitative Target Approximation model of tone production

(Prom-on et al [15]). In this model, in the process of tone contour production, 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:

$$x(t) = mx + b \tag{3}$$

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 [19]. As a result, actual F0 contours of tones are characterized by a third-order critically damped system:

$$f0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{\lambda} t \tag{4}$$

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 2 shows a number of different combinations of the parameters to demonstrate how the actual F0 behaves in accordance with the underlying pitch targets. In this paper, we experiment with representing each tone contour with the 3-d vector of qTA parameters. Given that qTA model has been shown to perform well in producing curves that closely resemble real tone contours in connected speech, an important question to be asked in this paper is: do qTA parameters perform well to reflect the similarities between tones in perception? In other words, in order to use the qTA parameters to achieve both dimensionality reduction and model-based clustering, we want to make sure that qTA parameters have the property where perceptually similar tone contour shapes also have similar parameter values. However, that is an empirical question not addressed in the proposal of qTA model, and it is yet to be seen whether qTA parameters are suitable for this task. Similarly, we also investigate the behavior of polynomial parameters in this regard.

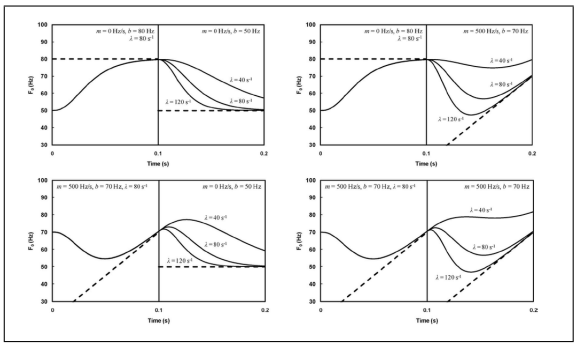

Figure 2. Examples of F0 contours generated by the qTA model with varying values of m, b, and λ. The dashed lines indicate the underlying pitch targets which are linear functions of m and b. The vertical lines show the syllable boundaries through which the articulatory state propagates.

3.3 Symbolic Aggregate approXimation

An important capacity of human cognition is its capacity to abstract away the commonalities from groups of pitch contours with much different fine detail variations. In this study, we experiment with the Symbolic Aggregate approXimation (SAX) representation of pitch contour vectors. SAX is the first symbolic representation for time series that allows for dimensionality reduction and indexing with a lower-bounding distance measure. In classic data mining tasks such as clustering, classification, index, etc., SAX is as good as well-known representations such as DWT and DFT, while requiring less storage space. [14].

Even though SAX representation is mostly used in time series data mining in the fields of bioinformatic (DNA sequencing) and shape mining, it has been applied to the computation of contour similarity from audio signals in a Query By Humming task [17] in the context of music information retrieval (MIR). It transforms the pitch contour into a symbolic representation with a user-designated length (nseg=desired length of the feature vector) and alphabet size (m), the latter being used to divide the pitch space of the contour into m parts assuming a gaussian distribution of F0 values. It is in principle very suitable for the current task as discussed above, as it is able to transform the fast-changing time varying signal of pitch contour into a coarse representation of abstract "shapes", which mimics the human cognition. In the current paper, we also show results from hierarchical clustering using a core dataset identified by the SAX representation to bear maximum simiarity to the tone contour shapes of the tone categories. These results validates that the SAX representation is indeed a good representation of tone contour shapes, analygous to the H-M-L representation used by linguists.

3.4 Raw Time-series Pitch Vector

Gauthier [11] showed that clustering and classification using SOM yielded a nearly 80% correct result when time-series are represented with a 33-point raw vector, despite the large dimension of this vector. In the current paper, we experimented with both different length raw pitch vectors and a equal length 30-point vector representation of pitch contours. The latter is desired also because of the requirement of many machine learning algorithms, including the SAX algorithm that converts equal-length vectors into dimensionality-reduced symbolic representations.

4. DATA COLLECTION AND PREPROCESSING

4.1 Data Collection

The current paper analyzes a connected Mandarin speech dataset consisting of 3 native speakers (2 female). The 'adult' data set is drawn from an existing data set composed of 11 young adult speakers of Mandarin in Beijing performing a semi-spontaneous speech production task in a conversational and casual speech style. The participants are given a list of words and to produce sentences in the

format of: ”wo3 xi3huan1/bu4 xi3huan1 X.”(”I like/don’t like X.”)

The participants pick items in the word list based on whether they like it or not, and produced 98 sentences per person. The current data set is a subset of the original data. A total of 348 tone contour units, evenly distributed across all four tones, are segmented and extracted. The format of this task determines that this dataset is a more restricted representation of the entire universe of adult speech that the L1 child is exposed to, especially for our purposes, considering the vast amount of possible variations that exists in the tone contours of adult speech in more varied positions, coarticulations, and emotional states in many dimensions of the acoustic space.

4.2 Data Preprocessing

The audio data is partly-automatically segmented and labeled by trained phonetician. Several praat scripts are implemented to automate this process. To extract pitch contours and their F0 values, a praat script is implemented for the extraction of pitch contours to automatically produce a csv data file. It generates a pitch object in Praat, in which pitch is computed using autocorrelation algorithm (pitch step 0.01s with standard setting on pitch floor and ceiling) for each frame of the pitch contour unit(i.e., PCU, a unit of one tone contour, usually corresponding to a syllable) that are identified and labeled earlier. Crucially, each PCU is assigned a pitch-con ID so that each of the pitch values consisting of this particular contour is identified collectively by this ID, facilitating the analysis at later stages. Speaker information is encoded into PCU-ID so that each PCU-ID in the entire data set (Adult and Child) is unique. F0 values are normalized so that each pitch contour has a mean F0 value of 0 and sd of 1. The mean pitch height for each pitch contour is also computed and normalized against the mean value of the overall F0 values for a given speaker. Due to the known F0 differences between tones (especially for tone 1 (H) and tone 3(L), which cannot be efficiently identified by their shape alone), mean pitch height is a useful feature for tone identification.

4.3 Model Parameter Extraction

For each of the dimensionality-reducing time series representation, codes are implemented to extract the model parameters from the raw F0 contours. Polynomial parameters are extracted with the numpy polyfit function in python. qTA parameters are extracted using the qTA PENTA Trainer praat script [1] . Symbolic representation of time-series based on SAX algorithm is converted using an implementation in Mablab. [2] In addition to the model parameters, a number of other features are included, including the normalized mean pitch height, previous tone context, and the duration of the contours.

[1] This tool is developed in tandem with the qTA model by [15], downloaded at http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/.

[2] The matlab code is modeled after the SAX demo code available from Jessica Lin at http://www.cs.gmu.edu/ jessica/sax.htm.

5. PROCEDURES AND RESULTS

5.1 Model Based Clustering

5.1.1 Procedure

Model-based clustering is performed on vectors of polynomial parameters, qTA model parameters, and duration, normalized mean F0, as well as previous tone context, using the R package `mclust`(Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation; finite Gaussian mixture modeling fitted via EM algorithm for model-based clustering, classification, and density Estimation, including Bayesian regularization and dimension reduction). This method is particularly targeted at exploring whether the tone contours can cluster into four multivariate Gaussian-like categories.

In this experiment, model based clustering is performed on two conditions, with different combinations of feature vectors: (1)time-normalized model parameters; (2)time-unnormalized model parameters. Table 1 shows the combination of feature vectors on polynomial and qTA-based parameters throughout different clustering experiments in this paper.

model	feature vector
1	polynomial+duration+prev-tone+mean-F0
2	polynomial(time-normalized)+prev-tone+mean-F0
3	1 with 30-point equal-length vectors
4	qTA+duration+prev-tone+mean0F0
5	all parameters (unequal length)
6	all parameters (equal-length 30-point)

Table 1. Feature vector combinations

5.1.2 Results

First, a density plot (see Figure 3) shows that the individual parameters do not form Gaussian-like distributions, with large range parameter values and multiple local maxima. These results indicate that these parameters may not have the desired property that forms a multivariate Gaussian distribution.

Second, model-based clustering showed that no matter the combination of factors, overall, these parameters cannot be clustered into meaningful four categories that correspond to the real tone categories. A representative plot of the result of the clustering is shown in Figure 4. Even though the algorithm did find four clusters (with option of number of clusters between 1 and 9), the tone contours generated by different categories are not well separated. In sum, neither qTA nor polynomial parameters are proved to have the desired property that can reflect degrees of similarity between tones, and are thus not well-suited for the current task.

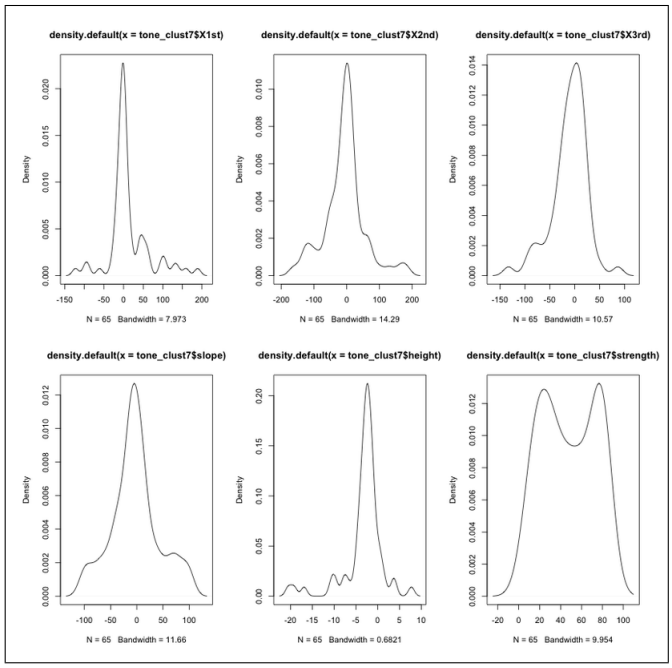

Figure 3. Model parameter distributions

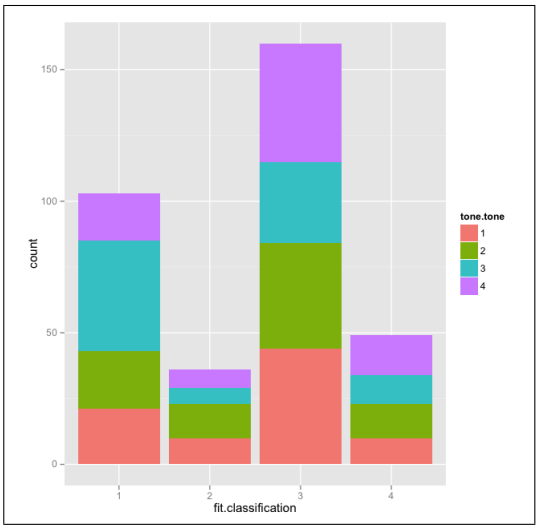

Figure 4. Model-based clustering with model 1 feature vector

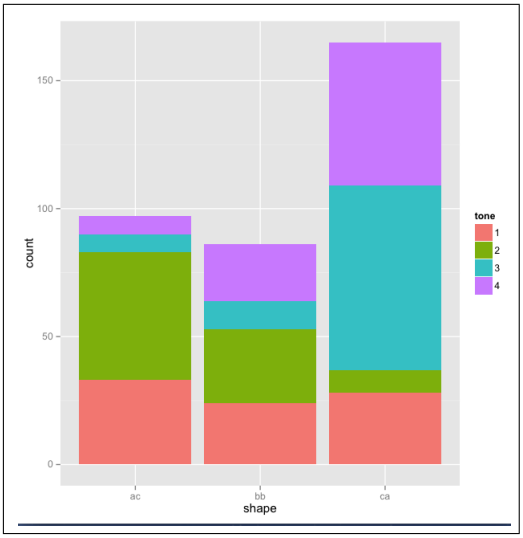

Figure 5. SAX representation of tone contour shapes across tone categories

5.2 SAX-based Clustering and Classification

5.2.1 Procedure

First, we convert raw 30-point pitch vectors into a SAX representation using an implementation discussed above in Matlab. Here we experimented with different parameters of nseg (number of segments used horizontally to represent pitch movements in time) and alphabet size (vertical division of pitch space, which assumes pitch distribution follows a Gaussian distribution). In particular, we experimented with smaller segments (nseg=2 or 3) vs. larger segments (nseg=5). Second, we performed clustering (k-means) of the SAX-based features. Third, we trained a classifier using the SAX-based shapes (e.g., if a pitch space is represented with three symbols, a<b<c, then ac,bb,ca with nseg=2 represent rising, level, and falling shapes of tone contours), along with the duration, mean-F0, and previous tone context feature vectors.

5.2.2 Results

SAX based clustering results showed interesting patterns. First of all, experimentation with different values of nseg and alphabet size shows that, in order to capture the abstract nature of tone contours and to be not affected by the large amount of noise in pitch movements, a limit of nseg <=3 must be placed. This is a reasonable limit considering that linguists use only two or three segments to represent tone contours in any tone language [3]. Alphabet size=3 or 4 are shown to be big enough to capture the levels of division in the pitch space, while small enough not to introduce too fine-grained labeling.

Second, the result of SAX-based clustering, regardless of choice of parameters, showed that all shapes (level, rising, falling, rising-falling, falling-rising, etc), are present in all tone categories. In particular, when we use nseg,alphabet-size=(2,3), we obtained three shapes across all tone categories: ac,bb,ca (as discussed above). An interesting observation to be made here is that (see Figure 5), tone 3 and tone 4 are more well separated than tone 1 and tone 2. These results point to the possibility that SAX-based shapes of pitch contours can be used as a good symbolic(nominal) feature for classification of tones. To test this idea, we built a classifier with the aforementioned feature combinations. The results showed that, despite the separability of tone3 and tone4 shown in Figure 5, the mean F0 and duration are not good enough features to distinguish tone 1 and tone 3, therefore the classification error remains at around 50%. Table 2 shows a comparison of the performance of different classification methods on this set of SAX-based shapes.

[3] In linguistics convention, high tone=H, low tone= L, rising=LH, falling=HL, falling rising=HLH, etc

method	correctly classified instances
Bagging	53.7%
Logistic	50.3%
J48	50.3%
RandomForrest	48.6%

Table 2. Classification results with SAX-based shapes

5.3 Hierarchical Clustering with DTW Distance

5.3.1 Procedure

Hierarchical clustering is performed to explore the similarities among with-in and between category contours without specifying the number of clusters. First, we performed hierarchical clustering on the whole dataset. Second, a core dataset is selected according to the SAX-based shapes so that only shapes within a category that show most similarity to the tone category is chosen (i.e., for tone 1, level:bb, for tone 2, rising: ac, etc.). The goal of this task is to show that given well-behaved data, these contours can be succefully clustered into well-separated groups. It also validates the SAX representation of tone contours. The time-series representation in this experiment is the raw equal-length 30-point pitch vectors.

5.3.2 Results

As expected, the whole data set did not cluster into well-separated, desired four categories on any level of the hierarchy. Smaller clusters that showed tone category aggregation occured on many levels. The clustering with the core subset, however, yielded a most promising result of clustering assignments (see Figure 6) that are indeed well separated. This result offers (1) a validation of the clustering methodology; (2) a validation of SAX-based representation and parameter setting in capturing the shape of pitch contours in a similar manner to linguistic representation of tones in human perception; (3) in reverse, a validation that the variance in the original dataset is indeed too big for the current clustering algorithm to successfully separate.

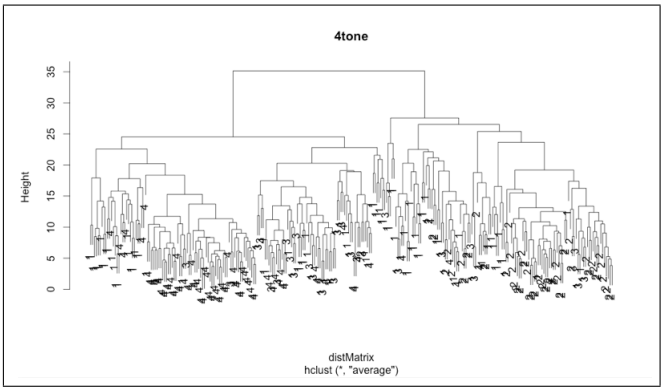

Figure 6. Hierarchical clustering on a well-behaved dataset

5.4 K-medoids Clustering with DTW Distance

5.4.1 Procedure

In this experiment, K-medoids clustering is applied to the 30-point pitch vector with Dynamic Time Warping (DTW) distance measure instead of a standard Euclidian distance measure. DTW is a well-established distance measure that uses dynamic programming to find the best alignment between two time-series data of equal or unequal length, even in the case of time-shift between the two vectors. The implementation is performed using the PAM function in the package `cluster` in *R*. The number of cluster is set to 4.

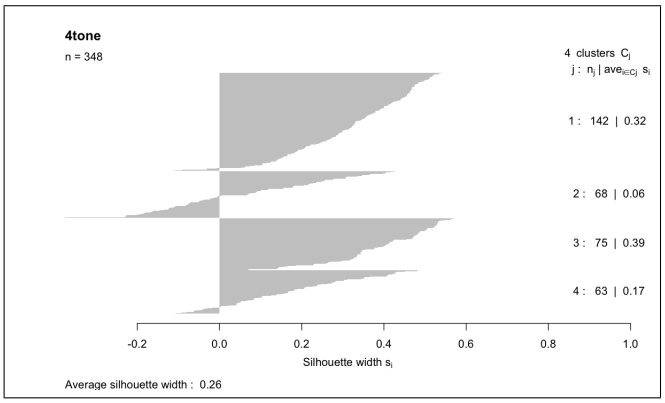

Figure 7. Silhouette plot of K-medoids clustering results

5.4.2 Results

The k-medoids results is shown in Figure 7 and Figure 8. The silhouette figure shows that there is one big cluster and three more evenly distributed cluster. Interestly, the histogram showed that, similar to the results of the SAX representation (notice that the current experiment is not performed with SAX representation), tone 3 and tone 4, while bearing similarity to each other, are in general more well separated than tone 1 and tone 2. Meanwhile, contour shapes in tone 1 and tone 2 more evenly distribute into all four clusters.

6. DISCUSSION

In this paper I have showed that while there are some interesting observations to be made about the different time-series representations and the clustering results, there are no algorithms/representations tested in the current study that showed a performance of clustering/classification of tone categories in connected speech that is comparable to the non-dimensionality reduced pitch and D1 (first derivative) feature vectors in Gauthier [11]. First, it was observed in the current study that different methods indicate that tone 3 and tone 4 bear considerable similarities in their contour shapes, which are more well separated from other shapes in the data and are mostly concentrated in one cluster. In contrast, tone 1 and tone 2 showed more variances in the shapes that are present in all clusters. Second, the experiment with SAX and the results with hierarchical clustering on the 'well-behaved' dataset showed that the SAX-

representation is successful in capturing the coarse shapes of the contours while ignoring the fine pitch movements that are insignificant in human tone classification and perception. Third, the failure of polynomial and qTA parameters in model-based clustering shows that these two kinds of representations are not suited to model the mental representation of tone perception. Lastly, comparing to the results of Gauthier [11], the current dataset is very small, which is a possible reason that the within-category consistency is more subject to the influence of the presence of noise.

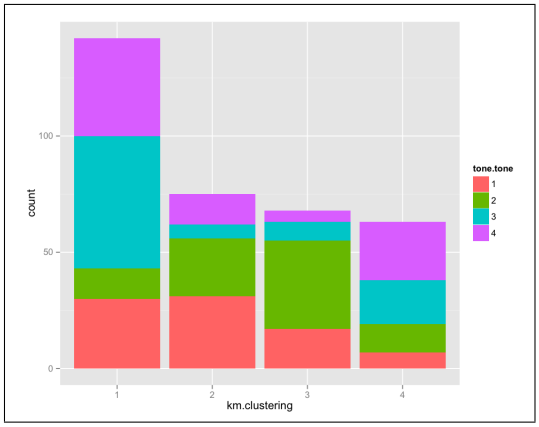

Figure 8. K-medoids Clustering results in tone categories

7. REFERENCES

[1] Blicher, D. L., Diehl, R. L., and Cohen, L. B.: Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: Evidence of auditory enhancement. Journal of Phonetics,18(1), 37-49.1990.

[2] Chang, Yung-hsiang Shawn.: Distinction Between Mandarin Tones 2 and 3 for L1 and L2 Listeners Proceedings of 23rd NACCL, 1: 8496.2011.

[3] Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C.: The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology, 54(4), 358-368.1957.

[4] Fry, D. B., Abramson, A. S., Eimas, P. D., and Liberman, A. M.: The identification and discrimination of synthetic vowels. Language and Speech, 5, 171-189.1962.

[5] Iverson, P., and Kuhl, P. K.: Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. Journal of the Acoustical Society of America, 97(1), 553-562.1995.

[6] Diesch, E., Iverson, P., Kettermann, A., and Siebert, C: Measuring the perceptual magnet effect in the perception of /i/ by German listeners. Psychological Research, 62, 1-19.1999.

[7] Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom,B.: Linguistic experience alters phonetic perception in infants by 6 months of age. Science, 255(5044), 606-608.1992.

[8] Davidoff, J., Davies, I., and Roberson, D: Colour categories in a stone-age tribe. Nature, 398, 203-204.1999.

[9] Feldman, N. H., Griffiths, T. L., and Morgan, J. L.: "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference." Psychological Review, 116(4), 752-782.2009.

[10] Feldman, N. H., Griffiths, T. L., Goldwater, S., and Morgan, J. L.: "A role for the developing lexicon in phonetic category acquisition." Psychological Review, 120(4), 751-778. 2013.

[11] Gautheir,B., Shi,R, Xu,Y.: Learning phonetic categories by tracking movements. Cognition, 103, (2007),80-106.

[12] Han, M. S., and Kim, K.-O.: Phonetic variation of Vietnamese tones in disyllabic utterances. Journal of Phonetics, 2, 223232.1974.

[13] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.2009.

[14] Lin,J., Keogh,E., Wei,L.,and Lonardi,S.: Experiencing SAX: a novel symbolic representation of time series. Data Mining and Knowledge Discovery. Oct.2007, Vol.15, Issue.2, pp107-144.2007.

[15] Prom-on, S., Xu, Y. and Thipakorn, B.: Modeling tone and intonation in Mandarin and English as a process of target approximation. Journal of the Acoustical Society of America 125: 405-424.(2009).

[16] Shen, X-N. S.: Tonal coarticulation in Mandatin. Journal of Phonetics, 18, 281- 295.1990.

[17] Valero,J: Measuring similarity of automatically extracted melodic pitch contours for audio-based query by humming of polyphonic music collections. Master's Thesis, MTG, DTIC, UPF, 2013.

[18] Xu, Y.: Contextual tonal variations in Mandarin. Journal of Phonetics, 25, 6183.1997.

[19] Xu, Y. and Sun X.: Maximum speed of pitch change and how it may relate to speech. Journal of the Acoustical Society of America 111: 1399-1413.2002.