AUTOMATIC SPEECH SEGMENTATION COMBINING AN HMM-BASED APPROACH AND RECURRENCE TREND ANALYSIS

Runqiang Yan, Yiqing Zu, Yisheng Zhu*

Department of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200030, China
*Motorola Research Center, Shanghai 200040, China
yan_runqiang@hotmail.com, yiqing.zu@motorola.com, yszhu@sjtu.edu.cn

ABSTRACT

Aiming at improving the speech segmentation accuracy acquired from standard HMM-based approach, this paper presents a nonlinear dynamical method for phoneme boundary adjustment by discerning and measuring the nonstationarity of speech dynamics. Dynamical systems of different phones present diversified invariant attractor structures in phase space. Therefore, when analyzing adjacent phones, there may exist a point, at which the underlying dynamics changes. In this study, time-dependent recurrence trend (TDRT) is proposed to describe the local changing degree of the nonstationarity of speech dynamics as time progress and identify the largest paling slop in the windowed recurrence plots (RPs) as the phoneme boundary. The experimental result shows that 9.41% increase in agreement within 20 ms with TDRT correction is obtained on TIMIT database.

1. INTRODUCTION

The standard HMM-based approach has been broadly used for automatic speech segmentation using a process called forced alignment [1]. Although the results are quite impressive, there are also many shortcomings that prevent it from achieving prefect performance [2] in application to the corpus-based TTS synthesis. Recently, many literatures have put forward improved methods, such as iterative correction rules [3] and hybrid HMM/ANN system [4]. In this paper, a nonlinear dynamical method is investigated to adjust initial phoneme boundaries acquired from HMM-based approach by detecting the underlying dynamical changes between adjacent phones.

Recent work demonstrates that rich nonlinearity may exist during speech production [5]. Generally, there are three mechanisms of the physical modeling: vocal fold oscillation for voiced speech, the turbulent sound source for unvoiced speech and interaction phenomena [6]. Dynamical systems of different phones present diversified invariant attractor structures in phase space [7]. Therefore, when analyzing adjacent phones, there may exist a point, at which the underlying dynamics changes. A relatively simple and straightforward two-dimensional visualization technique, recurrence plot (RP) introduced by Eckmann et al. [8], can graphically display such abrupt changes or slow time-varying nature of the underlying driving force. As a quantification parameter to describe the paling of recurrent points in RP, recurrence trend (RT) can be used to discern and measure the nonstationarity of speech dynamics.

The rest of the paper is organized as follows. In section 2, a review of the technique of RP is firstly described and the concept of RT is given. In section 3, the parameters in RP of speech signal are statistically determined and some RP representatives of diphone segments are presented. Following this, the explanation of phoneme boundary determination framework (time-dependent recurrence trend, TDRT) is discussed. The experimental result in section 4 shows that 9.41% increase in agreement within 20 ms with TDRT correction is obtained on TIMIT database. Finally, conclusions are provided in section 5.

2. RECURRENCE PLOT AND RECURRENCE TREND ANALYSIS

The basic idea of RP is local recurrence or neighborliness of vector points in reconstructed phase space. According to Takens' embedding theorem, a scalar time series $\{x(i), i = 1, 2, ...\}$ is firstly applied to construct vectors in the form: $X_i = \left(x(i), x(i+\tau), ..., x(i+(m-1)\tau)\right)$ where m is embedding dimension and τ is time delay. Such temporal point sets $\{X_i, i = 1, 2, ..., N\}$ represent certain trajectories in an m-dimensional phase space. Then the following important step is the calculation of the $N * N$ matrix:

$$R_{i,j} = \Theta\left(\varepsilon - \left\|X_i - X_j\right\|\right), i, j = 1, 2, ..., N \quad (1)$$

Where ε is a predefined cutoff distance, $\|*\|$ is the norm (e.g. the Euclidean norm) and $\Theta(x)$ is the Heaviside function. The matrix only consists of the value 0 and 1, and then its graphical representation is a $N * N$ grid of points, which are visualized as white for 0 and black for 1. Such

kinds of graphics, RPs, give us visual inspection of higher dimensional phase space trajectories in a two-dimensional space.

RP demonstrates characteristic large-scale and small-scale patterns. The former pattern is denoted as topology and the latter as texture. The topological structure offers a global impression, which can be characterized as homogeneous, periodic, drift and disrupted. The appearing textural figures in RPs exhibit different forms and represent different dynamical behaviors [9], e.g., diagonal (similar local evolution of pair of trajectories) and vertical or horizontal line (state does not change for some time).

Zbilut and Webber [10] have developed the recurrence quantification analysis (RQA) to quantify the recurrent behaviors of the underlying dynamics by using the recurrence point density and the diagonal structures exposed in RPs. Five types of RQA variables are always examined: recurrence rate, determinism, maximal length of diagonal structures, entropy and trend. Among these quantification parameters, recurrence trend is a measure of the paling of recurrent points away from the main diagonal to the plot's corners in so far as a flat slop indicates strong stationarity, whereas large slope indicates poor stationarity due to the nonstationarity of the underlying dynamics. RT is calculated using linear regression method as follows:

$$RT = \frac{\sum_{i=1}^{K}(i-(K+1)/2)(RR_i - \overline{RR})}{\sum_{i=1}^{K}(i-(K+1)/2)^2} \qquad (2)$$

Because of symmetry along the main diagonal, RP can be described by its half part (i.e., the upper left or lower right triangle). $K-1$ lines equally spaced paralleling to the main diagonal can divide the triangle into K parts. RR_i represents the percentage of recurrence in the i th part and $\overline{RR}$ is the mean of $\{RR_i, i=1,2,...,K\}$.

The computation of RT in short time frame overlap-moving along the time sequence can yield time-dependent recurrent behaviors of the underlying dynamics. For the stationary dynamics, TDRT curve should be regular and smooth, therefore, the acute variation, if possible appearing, can make the identification of the dynamical changes underlying the time series.

3. DETECTING CHANGE POINTS BETWEEN PHONES

In order to introduce the TDRT approach for speech segmentation, a set of English phone and diphone segments are extracted from TIMIT database. Before RP analysis, each wave is firstly rescaled to the interval [-0.5, 0.5].

3.1. Parameters in RPs

When presented with experimental dynamical data in the absence of expert or priori knowledge, heuristics are developed for selecting proper embedding parameters of speech dynamics. In this paper, false nearest neighbor technique and average mutual information method [11] are used to calculate embedding dimension m and delay time τ respectively. Figure 1-A1 shows the fraction of false nearest neighbors of the six selected phonemes: monophthong 'aa', diphthong 'aw', voiced stop 'd', voiced fricative 'dh', lateral 'l' and nasal 'm'. Figure 1-B1 describes average mutual information of each corresponding phoneme. Statistical results for 39 phones, which 61 possible phonetic and phonemic labels in TIMIT are folded according to the conventions discussed in the section 4, are given in Figure 1-A2 and 1-B2. These graphs show that an appropriate choice of embedding dimension is five and time lag is eight.

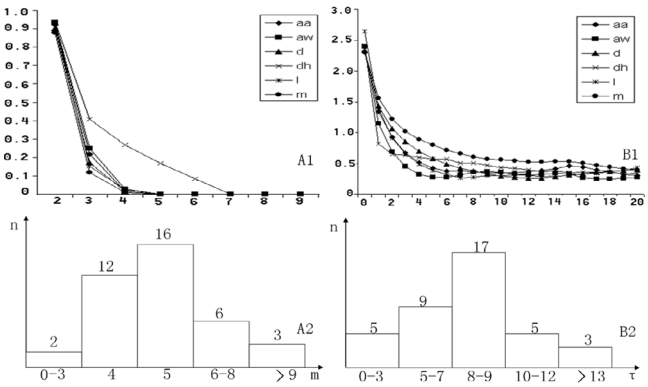

Figure 1. Embedding parameters selection on a sample of speech phonemes. (A1, A2) for embedding dimension m and (B1, B2) for time delay τ.

To avoid the false recurrent results from choosing too small or too large predefined cutoff distance ε and also in the interest of RPs comparison, $\varepsilon = \sigma$ is chosen, where σ is the standard deviation of wave time series. In the latter experiments, the phase space reconstruction and matrix grid representation to establish RPs will use the above results.

3.2. RPs of diphones

In order to present the dynamical changes between adjacent phones, six 'half' RP representatives of diphones and corresponding waveforms are exhibited in Figure 2, Figure 2-A represents the diphone 'iyaa', which is extracted from the word 'periodical'; in 2-B, 'tuw' from 'too'; in 2-C, 'shiy' from 'she'; in 2-D, 'vow' from 'votes'; in 2-E, 'mey' from 'may'; in 2-F, 'ruw' from 'room'. These RP graphics display different phonetic connections between vowel and vowel or consonant and vowel.

With different production mechanisms alternating in continuous speech, such as oscillations of vocal folds for voiced sounds and turbulent sources for unvoiced sounds,

the state of vocal tract and articulators inside the mouth changes timely. RPs can describe such time-varying phenomena by reconstructing their underlying dynamics explicitly exhibited in Figure 2. For example, in the half RP of 'iyaa', the part of 'iy' exhibits periodic structures and the same as the part of 'aa', but the upper left rectangular area is much more different from these two triangular parts, where appears many blank white strips caused by different underlying dynamics between 'iy' and 'aa'. Therefore, the phoneme boundary can be mapped to the right borderline of the rectangle in the RP.

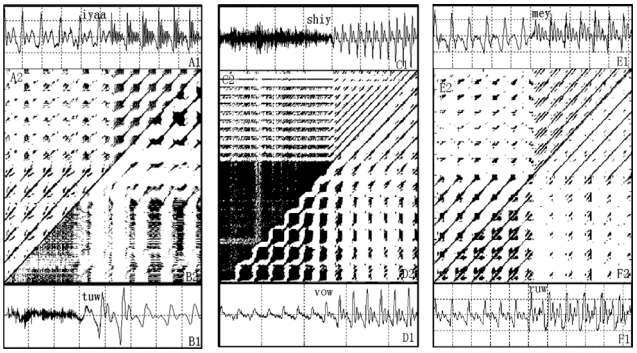

Figure 2. The waveforms and 'half' RPs of some diphone representatives. (A1, A2) 'iyaa'; (B1, B2) 'tuw'; (C1, C2) 'shiy'; (D1, D2) 'vow'; (E1, E2) 'mey' and (F1, F2) 'ruw'.

3.3 Time-dependent recurrence trend analysis

Because the dynamical change point between phones cannot be determined by a single, global RT, time-dependent recurrence trend is thus proposed to describe the local changing degree of the nonstationarity of speech dynamics as time progress and identify the largest paling slop in the windowed RPs as phoneme boundary.

When applying TDRT analysis, the sampled diphone segments is firstly blocked into frames of 20 ms with frame shift of 5 ms. Then calculate the RT of each frame using equation 2, so the TDRT curve can be plotted to expose the recurrent behaviors of the underlying dynamical changing as time progress.

In Figure 3, six TDRT curves of corresponding diphone representatives in Figure 2 are exhibited. The undermost value point of each curve, "a", "b", "c", "d", "e" and "f", will be chosen as the dynamical change point between phones, because in this frame the unstable character is prominent. Therefore, the period between the starting and the extreme point belongs to the left phone and the residual belongs to the right. During these two periods, TDRT also exhibits nonstationary phenomena of the phones, such as relative drastic fluctuation in Figure 3-D for the voiced fricative 'v' and the monophthong 'ow'.

The time point on the diphone sequence, which reflects the underlying dynamics changing, can be calculated by:

$$trans = (k-1)*fs + \left(\frac{RT_{k-1} - RT_k}{RT_{k-1} + RT_{k+1} - 2*RT_k}\right)*ws \quad (3)$$

where k is the index of the extreme point, RT_{k-1}, RT_k and RT_{k+1} refer to the RT values of the left, the extreme and the right point respectively, such as 'aL', 'a' and 'aR' in Figure 3-A. fs is the frame shift, 5 ms, and ws refers to the frame size, 20 ms.

The *trans* of each diphone in Figure 3 calculated by equation 3 is much more coincident with its corresponding speech from visual and acoustical perceptual meanings.

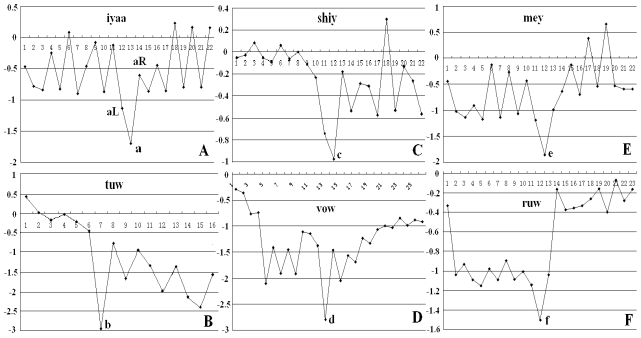

Figure 3. The TDRT curves of corresponding diphone representatives in Figure 2. The index of time frame along the x-axis and the RT value of each frame along the y-axis.

4. EXPERIMENTS

The TIMIT acoustic/phonetic database consists of 2360 independent sentences spoken by 630 speakers from eight different dialect regions. These speech sentences also have phonemic labels checked manually. In order to evaluate the performance of TDRT, experiments are carried out using TDRT approach for short diphone speech segmentation and HMM approach with TDRT correction for continuous speech segmentation. By comparison, the results of standard HMM-based approach are also given. In these two experiments, the sentences in 'train' 'dr1' are used, which consist of 380 sentences spoken by 38 speakers. Diphone segments are consecutively extracted. By folding the allophone {dx, nx, q, hv, ux, ax-h} to its corresponding phoneme, amalgamating closure {bcl, dcl, gcl, pcl, kcl, tcl} to its owning and neglecting the difference {ah, ax}, {axr, er}, {ix, ih}, {en, n}, {el, l}, {em, m}, {eng, ng}, {h#, epi, pau}, 61 existing phonetic and phonemic labels are folded to 40 labels including 'silence'. The number of phoneme occurrences is 12421 and the number of phoneme boundaries to be determined is 12041.

In the experiments, a simple left-to-right 3 states multi-Gaussian phoneme-based HMM system is used. The Parameter vectors comprise 12th order MFCC parameters and energy, plus delta and acceleration coefficients (total 39

coefficients). The number of mixtures of each state is set to 16 and the frame size is 20 ms with 5 ms frame shifting.

In the diphone speech segmentation experiment, 11991 diphone segments, with more than 45 ms duration, are used for training and testing by HTK [12] and TDRT.

Occurrence	351	4093	2114	4219	1214	11991
	VV	VC	CC	CV	#	Total
	(%)	(%)	(%)	(%)	(%)	(%)
HMM (ms)						
<10	79.41	80.08	80.09	80.32	79.93	80.13
<20	89.86	91.13	90.52	92.07	90.78	91.28
<30	94.94	95.22	94.78	95.09	95.38	95.11
TDRT (ms)						
<10	85.59	91.41	86.14	91.08	91.68	90.22
<20	90.64	95.11	91.12	95.18	94.73	94.26
<30	95.34	98.76	94.84	98.38	99.14	97.87

Table 1. Segmentation accuracy of diphone speech using HMM and TDRT approaches.

Table 1 shows the segmentation accuracy using HMM and TDRT respectively for four broad phonetic classes (VV, VC, CC and CV) and diphones including 'silence' (V refers to vowel, C to consonant and # to diphones including 'silence'). By comparison, the TDRT approach explicitly outperforms the HMM for each category and provides more accurate segmentation. Considering the accuracy within 20 ms, a relative 3.26% increase in agreement is acquired.

Although HMM approach cannot precisely anchor the phoneme boundaries between phones, it provides appropriate range for further adjustment of the inner phoneme boundaries when applied to continuous speech segmentation. In the following experiment, when the standard HMM-based segmentation results are acquired, TDRT correction process is consecutively applied around the initial phone boundaries.

Occurrence	351	4098	2123	4253	1216	12041
	VV	VC	CC	CV	#	Total
	(%)	(%)	(%)	(%)	(%)	(%)
HMM (ms)						
<10	64.22	67.32	66.76	68.08	65.92	67.26
<20	80.45	82.69	81.48	83.32	81.01	82.46
<30	88.76	90.27	90.23	90.82	89.27	90.31
HMM & TDRT (ms)						
<10	71.42	75.13	71.81	74.96	74.74	74.34
<20	86.94	90.94	87.67	90.82	91.07	90.22
<30	91.87	96.49	92.74	96.62	96.41	95.73

Table 2. Segmentation accuracy of continuous speech using HMM approach with and without TDRT correction.

Table 2 shows the segmentation accuracy using HMM approach with and without TDRT correction (organization is the same as for Table 1). It demonstrates that the use of TDRT correction results in a 10.53% increase in agreement within 10 ms, 9.41% within 20 ms and 6% within 30 ms. Therefore, TDRT correction process effectively improves the segmentation accuracy.

5. CONCLUSIONS

HMM-based approach may not provide the best solution towards high-accuracy segmentation because its models are intently built to identify phonetic segments. Therefore, for the corpus-based TTS synthesis, where the performance is closely related to the accuracy of the phonetic labeling, further boundary adjustment is necessary. The TDRT correction process gives an effective way toward such proposal by discerning and measuring the nonstationarity of speech dynamics. The experimental result shows that the segmentation accuracy on the continuous speech database using HMM approach can be effectively improved with TDRT correction. However, there are still a lot of issues for further investigation, e.g., the influence of the phonemic annotation errors and the nonstationary dynamics during phone production.

6. REFERENCES

[1] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models," Speech Communication, vol. 12, pp. 357-370, 1993.

[2] J. P. H. van Santen and R. W. Sproat, "High-Accuracy Automatic Segmentation," presented at EUROSPEECH, 1999.

[3] F. Chou, C. Tseng, and L. Lee, "A Set of Corpus-Based Text-to-Speech Synthesis Technologies for Mandarin Chinese," IEEE Transactions on Speech and Audio Processing, vol. 10, pp. 481-494, 2002.

[4] F. Malfrere, O. Deroo, T. Dutoit, and C. Ris, "Phonetic Alignment: Speech Synthesis-Based vs. Viterbi-Based," Speech Communication, vol. 40, pp. 503-515, 2003.

[5] M. Faundez-Zanuy, G. Kubin, W. B. Kleijin, P. Maragos, S. McLaughlin, A. Esposito, A. Hussain, and J. Schoentgen, "Nonlinear Speech Processing: Overview and Applications," Control and Intelligent Systems, vol. 30, pp. 1-10, 2002.

[6] W. B. Kleijn and K. K. Paliwal, Speech Coding and Synthesis: Elsevier Science B. V., 1995.

[7] X. Liu, R. J. Povinelli, and M. T. Johnson, "Detecting Determinism in Speech Phonemes," presented at IEEE Digital Signal Processing Workshop 2002, 2002.

[8] J. P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence Plots of Dynamical Systems," Europhysics Letter, pp. 973-977, 1987.

[9] J. B. Gao and H. Q. Cai, "On the Structures and Quantification of Recurrence Plots," Physics Letters A, vol. 270, pp. 75-87, 2000.

[10] C. L. Webber, Jr. and J. P. Zbilut, "Dynamical Assessment of Physiological Systems and States Using Recurrence Plot Strategies," Journal of Applied Physics, pp. 965-973, 1994.

[11] R. Hegger, H. Kantz, and T. Schreiber, "Practical Implementation of Nonlinear Time Series Methods: the TISEAN Package," Chaos, vol. 9, pp. 413-435, 1999.

[12] HMM Tool Kits (HTK), [Available]: http://htk.eng.cam.ac.uk.