

NONLINEAR STATE SPACE EMBEDDING FEATURES AND THEIR APPLICATION TO ROBUST SPEECH SEGMENTATION

Brett Y. Smolenski, Uchechukwu O. Ofoegbu, Jashmin K. Shah, Robert E. Yantorno

Temple University/ECE Dept. 12th & Norris Streets, Philadelphia, PA 19122-6077, USA
bsmolens@temple.edu, uche1@temple.edu, shah@temple.edu, robert.yantorno@temple.edu
http://www.temple.edu/speech_lab

ABSTRACT

The aim of this research is to develop features from a state-space embedded signal useful for high resolution and robust real-time speech segmentation. Since the set of state-space trajectories of a system can completely describe the system, a state-space embedding of a signal is typically used to qualitatively study any nonlinearities of the system generating a signal. However, while it may be easy for one to observe patterns in the state-space trajectories of a system, it is often difficult to quantify what is observed. In this paper two novel features are extracted from a state-space embedded signal using concepts from differential geometry. These features are computed iteratively on the 1-dimensional speech signal and they completely characterize the state space trajectories formed by the signal. The results obtained show that these features are particularly useful for classifying voicing states and can detect these phoneme boundaries with a resolution of four samples.

1. INTRODUCTION

Most speech processing applications partition the speech signal into short fixed length frames with some degree of overlap [1]. Framing is always necessary, since the speech signal is non-stationary [1]. However, a more intelligent approach to segmenting the speech signal would be to identify the stationary regions in the speech signal and then process those entire segments.

The Hidden Markov Model (HMM) is commonly used to model speech and other non-stationary signals [2]. The idea behind this model stems from the observation that the speech signal can be decomposed to a sequence of states called phonemes [2]. While in a particular state, the speech signal is quasi-stationary and it is often modeled as an all pole system being excited by either a periodic impulse train (voiced speech) or white Gaussian noise (unvoiced speech) [1]. Given this information, it may be more prudent to segment the speech signal into stationary segments as opposed to arbitrarily breaking the signal into frames.

Traditional approaches to speech segmentation used linear models of the speech signal [3]. However, it is well known that the speech production mechanism, in general, is a nonlinear

system, which can only be approximated using linear models [4]. The Navier-Stokes equation is a nonlinear partial differential equation that represents one of the most general models of the human speech production mechanism [1]. For example, unvoiced speech corresponds to turbulent flow that is described by chaotic solutions of the Navier-Stokes equation [5]. A chaotic signal is a signal that appears random, but is actually the result of a deterministic nonlinear system. One also observes chaotic signal properties when the speech is co-channel [6].

The evolution of a nonlinear dynamical system can be described by a point moving along a trajectory in its state space, where the coordinates of the point are independent degrees of freedom of the system (memory elements) [5]. The signal state-space embedding method was developed for analyzing chaotic signals generated by nonlinear systems [4]. When a signal is embedded in state-space it is transformed into a trajectory in an m -dimensional space. The number of necessary dimensions corresponds to the number of state variables necessary to describe the system [5]. Unfortunately these state variables are not directly observable. However, according to Takens' embedding theorem, it is possible to reconstruct a state-space representation topologically equivalent to the original state-space of a system using the 1-dimensional observable signal [7]. Topologically equivalent means that there exists a one-to-one transformation between the embedded signal and the actual state-space trajectory.

Using Takens' method of delays, points $\mathbf{x}(i)$ in an m -dimensional space are formed from time-delayed values of a signal $s(i)$:

$$\mathbf{x}(i) = [s(i), s(i-d), s(i-2d), \dots, s(i-(m-1)d)], \quad (1)$$

where m is the embedding dimension, and d is the chosen delay value in samples. For this research, a constant embedding dimension of 3 was used, since it has been shown that voiced speech can be adequately embedded in 3 dimensions [4]. The choice of an optimal delay parameter d depends on the sampling rate and mutual information between samples in the signal. The delay should be large enough so that adjacent points $\mathbf{x}(i)$ have a minimum of mutual information between them [7]. However, one can not make the delay arbitrarily large, since one would sacrifice time resolution. A constant d value of 12 samples has been found to produce good embedding results [4] and this was the value used for this research.

2. BACKGROUND

Figure 1 (below) shows a segment of a hypothetical state space trajectory in three dimensions, which could describe the dynamics of a nonlinear system having three memory elements. As the state space trajectory of the system unfolds, one can consider the rectangular **TNB** frame of reference moving along the curve characterized by the three perpendicular *osculating*, *normal*, and *rectifying* planes [8].

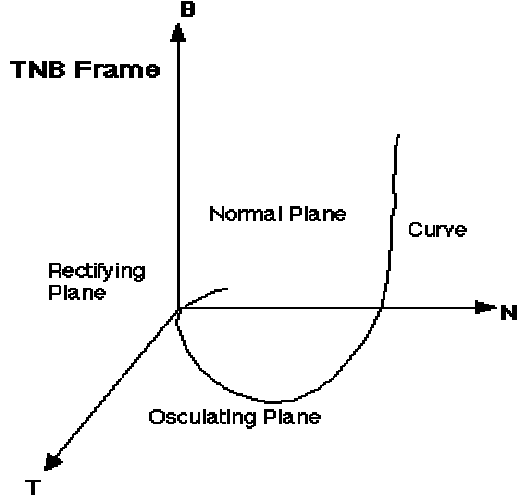


Figure 1: Illustration of the rectangular **TNB** frame of reference along the segment of a state space trajectory.

The Serret-Frenet theorem from the subject of Differential Geometry [9] states that any 3-dimensional space curve can be completely characterized by the following matrix equation relating the **T**, **N**, and **B** vectors:

$$\begin{bmatrix} \dot{T} \\ \dot{N} \\ \dot{B} \end{bmatrix} = \begin{bmatrix} 0 & \tau & 0 \\ -\tau & 0 & \kappa \\ 0 & -\kappa & 0 \end{bmatrix} \begin{bmatrix} T \\ N \\ B \end{bmatrix} \quad (2)$$

where κ is the *curvature* and τ is the *torsion* (defined below). The derivatives of the vectors on the left side of Equation 1 are with respect to s , the arc length of the curve [9].

The curvature and torsion are defined as:

$$\kappa = \lim_{\Delta s \rightarrow 0} \frac{\Delta \theta}{\Delta s} \quad (3)$$

$$\tau = \lim_{\Delta s \rightarrow 0} \frac{\Delta \Phi}{\Delta s} \quad (4)$$

where $\Delta \theta$ is the angle between the tangents **T** to the curve and $\Delta \phi$ is the angle between the binormals **B** to the curve [9]. Thus κ is the rate at which the tangent at a point P rotates as it moves along the curve. Hence, the reciprocal of κ is the radius of curvature. The torsion τ is the rate at which the unit binormal **B** at the point P rotates as it moves along the curve. Since the torsion and curvature parameters completely describe the curve, it is likely that they would serve as useful features.

It should be noted that the space curve formed by the state space embedding procedure is really an estimate and sampled version of the actual state-space trajectory. Hence, the curvature, torsion, and necessary derivatives in Equation 1 must be estimated from the discrete embedding curve. This was accomplished by using the following formulas:

$$S_n = |A_n| \quad (5)$$

where $|\cdot|$ is the Euclidean norm and the vector A_n is defined as:

$$A_n = \langle x_n - x_{n-1}, y_n - y_{n-1}, z_n - z_{n-1} \rangle \quad (6)$$

which is the elemental arc length of the discrete embedding curve. To approximate the curvature K_n and torsion T_n , the formulas below were derived:

$$K_n = \cos^{-1} \left(\frac{-A_n \cdot A_{n+1}}{|A_n| |A_{n+1}|} \right) \quad (7)$$

and:

$$T_n = \cos^{-1} \left(\frac{\langle -A_n \times A_{n+1} \rangle \cdot \langle -A_{n+1} \times A_{n+2} \rangle}{|A_n \times A_{n+1}| |A_{n+1} \times A_{n+2}|} \right) \quad (8)$$

where \times and \cdot represent the vector cross and dot products, respectively.

3. PROCEDURE

The speech signals used were obtained from the TIMIT database after being down sampled to 8 kHz. 13 Female and 12 male utterances were used. Pink noise was added and scaled to produce an overall SNR of 15dB and 30dB. Pink noise was chosen due to its prevalence in operational environments [6]. These noisy signals were embedded using Takens' method of delays and the curvature and torsion features were calculated using Equations 8 and 9 respectively. The data was labeled, using manual inspection, as voiced, unvoiced, silence, mixed, and transition; however, the curvature feature currently classifies speech only as voiced or unvoiced (no phonation). Hence, unvoiced detections by the curvature features were evaluated using the unvoiced and silence data, while voiced detections were compared to the voiced portions only. Transitions were not considered in the curvature's assessment.

4. RESULTS AND DISCUSSION

Figures 2 and 3 (below) show the state-space embedding for unvoiced and voiced speech, respectively. The unvoiced embedding was obtained using 500 samples of the phoneme /s/ and the voiced embedding was obtained using 500 samples from the word "we'll". One can observe that the embedded signal for unvoiced speech is highly chaotic and random while voiced speech generates a very structured embedded signal. This is because voiced speech can be described by a system having a lower number of state variables; where as the turbulence in unvoiced speech requires a large number of state variables [4].

For voiced speech it has been shown that the embedded trajectories should be in the general form of an ellipse with loops, where the ellipse corresponds to the pitch period and the loops correspond to the resonances in the vocal tract [4]. The torsion for any space curve that can be represented in a plane, such as an ellipse, is always zero [9]. Hence, for a particular vowel, one would expect the curvature signal to be consistently

large (near the value of π), and the torsion signal to be consistently small (near the value of zero). Since the 500 samples used for Figure 3 were taken across the three voiced phonemes comprising the word “we’ll”, several elliptical structures can be observed.

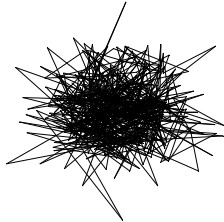


Figure 2: Embedded 500 samples of the unvoiced phoneme /s/.

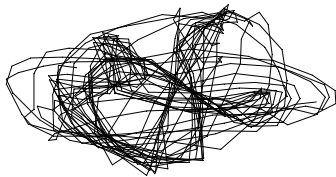


Figure 3: Embedded 500 samples from the word “we’ll”.

In order to obtain the threshold for the voiced/unvoiced classification (V/UV), the histogram of the curvature signal for the entire data set was plotted. This resulted in a bimodal histogram – one mode for unvoiced speech and the other for voiced speech. The threshold was then chosen as the midpoint between these two modes.

Figure 4 (below) shows the raw curvature as well as the median filtered curvature histogram for the entire data set prior to adding noise. Note the two distinct modes in these histograms. Since curvature values are higher for voiced segments than unvoiced, the mode representing voiced speech is on the right, while the mode represent unvoiced speech is on the left. The curvature signal was 79-point median filtered in order to smooth it and reduce the overlap between the two classes. Since it is well known that median filters preserve jump discontinuities [10], a median filter was necessary in order to preserve the resolution of the V/UV classification. In addition, the individual modes are approximately exponentially

distributed. In this situation it can be shown that the best predetection filter is the median filter [10].

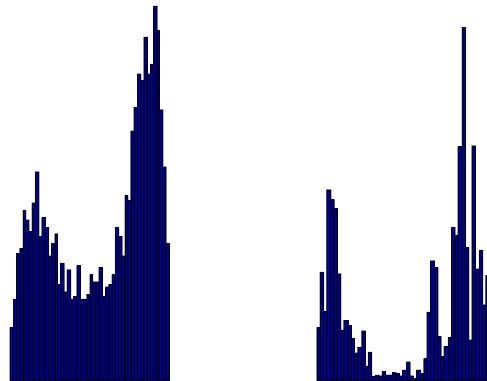


Figure 4: Histogram of raw curvature values for entire data set with no added noise.

It is interesting to note the level of bimodality seen in the raw curvature signal. The energy and zero-crossing rates have traditionally been used to classify voicing state; however, even their averaged values do not have such a high level of bimodality [3].

Figure 5 (below) shows the 79-point median filtered curvature signal plotted with the words “we’ll serve” where the voiced fricative /v/ has been truncated. For ease of viewing, both signals were normalized between (-1, 1) and the curvature signal had a constant of 1 added to it. One can clearly see how the curvature signal decreases significantly during the unvoiced /s/ phoneme.

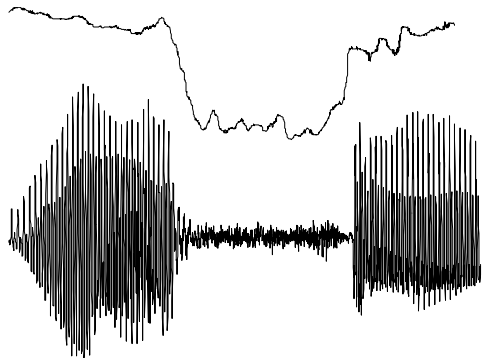


Figure 5: Median filtered curvature signal plotted with the speech “we’ll ser...”.

Figure 6 (below) shows the 79-point mean filtered torsion signal plotted with the same speech as in Figure 5. As expected, the torsion signal increases during unvoiced speech; however the transition is not as pronounced as with the curvature signal and there exists more variability. Since median filters preserve impulses as well [10], mean filtering was necessary for the

torsion signal due to its impulsive structure. More sophisticated predetection filters are currently being studied.

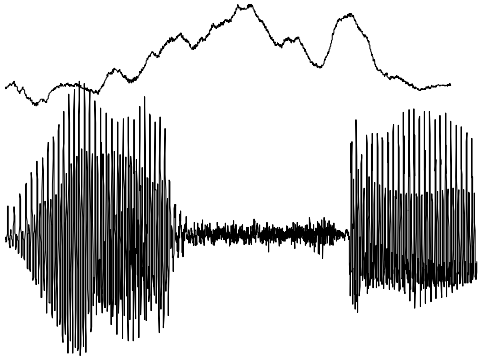


Figure 6: Mean filtered torsion signal plotted with the speech “we’ll ser...”.

In addition, the curvature signal was used to detect voiced and unvoiced speech on a sample-by-sample basis for the entire clean speech data set, as well as, with 15dB and 30dB added pink noise. The percentage of hits and false alarms for these classifications are given in Table 1 (below).

Table1: Percentage of hits and false alarms.

	V Hits	V False Alarms	UV Hits	UV False Alarms
Clean	86.8	8.3	91.7	13.2
30dB	88.7	9.9	90.1	11.3
15dB	93.2	17.4	82.6	6.8

From Table 1, it is observed that pink noise has little effect on the performance of the curvature measure.

5. FURTHER RESEARCH

An alternative approach to embedding a signal can be accomplished using Singular Value Decomposition (SVD) of the Hadarmond matrix [4]. This approach has been shown to be more robust in the presents of additive noise [4]. Further improvement may be obtainable by working with the LPC-residual or, equivalently, using Generalized SVD (GSVD), which pre-whitens the signal [11]. Since the effects of the vocal track resonances would be removed, pre-whitening should ‘straighten out’ the embedded signal for voiced speech so that only quasi elliptical orbits would remain for the state-space trajectories. This should yield a more consistent curvature and torsion values during voiced speech.

In addition, the average parameters of the ellipses could be linearly transformed into circles having an average unit radius, hence, producing a lower variance curvature and torsion features for detecting pitch period, voicing state, and phoneme boundaries. This is possible, since it can be shown that, for a circle, the curvature is the radius of the circle and the torsion is

zero [9]. Further, these transformations and the whitening process would have the effect of making the curvature and torsion features for unvoiced speech more uniformly distributed on the interval $(0,\pi)$.

Further, it is hoped that these features could be applied to detecting *usable speech* (speech that has been corrupted but is still usable for tasks such as speaker identification), since usable speech is known to have more structure than unusable [6].

ACKNOWLEDGEMENT

This effort was sponsored by the Air Force Research Laboratory, Air Force Material Command, and USAF, under agreement number F30602-03-1-0216. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

DISCLAIMER

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory, or the U.S. Government.

6. REFERENCES

- [1] T. F. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*, Upper Saddle River, NJ: Prentice-Hall, 2002.
- [2] X. D. Huang, Y. Ariki, and J. A. Mervyn, *Hidden Markov Models for Speech Recognition*, Edinburgh: Edinburgh University Press, 1990.
- [3] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, New York, NY: John Wiley, 2000.
- [4] G. Kubin, “Nonlinear Processing of Speech”, in *Speech Coding and Synthesis*, Elsevier, 1995.
- [5] R.C. Hilborn, *Chaos and Nonlinear Dynamics: An Introduction for Scientists and Engineers*, New York, NY: Oxford University Press, 2000.
- [6] R.E. Yantorno, “Co-Channel Speech Study,” Final Report for Summer Research Faculty, Sponsored by AFRL/IF Laboratory, Rome, NY: 1999.
- [7] F. Takens, “Detecting Strange Attractors in Turbulence,” *Lecture Notes in Mathematics*, Vol. 898, eds. D.A. Rand and L.S. Young, Springer, Berlin, 1981.
- [8] M. Rahman and I. Mulolani, *Applied Vector Analysis*, Boca Raton: CRC Press, 2001.
- [9] Y. Talpaert, *Differential Geometry: With Applications to Mechanics and Physics*, New York: Marcel Dekker, 2001.
- [10] J. Astola and P. Kuosmanen, *Fundamentals of Nonlinear Digital Filtering*, Boca Raton, FL: CRC Press, 1997.
- [11] R.J. Vaccaro ed., *SVD and Signal Processing II: Algorithms, Analysis, and Applications*, Elsevier, New York, NY: 1991.