

Structure-Based Voiced/Usable Speech Detection Using State Space Embedding

Uchechukwu O. Ofoegbu, Brett Y. Smolenski and Robert E. Yantorno,

Speech Processing Laboratory, Temple University

12th & Norris Streets, Philadelphia, PA 19122-6077, USA Tel: 215-204-6984

E-mail: uche1@temple.edu, bsmolens@temple.edu, robert.yantorno@temple.edu

http://www.temple.edu/speech_lab

Abstract: The process of speech production in the human system is very complex, possesses nonlinearities, and can only be precisely modeled in terms of nonlinear dynamics. A non-linear speech classification approach is proposed, which classifies speech based on features extracted from Takens' Method of Delays, a technique used to reconstruct signals into a trajectory in multi-dimensional state space. In this research, two types of speech detection are presented, namely, voiced and usable speech (for speaker identification purposes). The proposed approach has been able to yield a probability of error of 12% in noisy environments for voiced speech detection, and 78% correct usable speech detection by comparing the structures of embedded voiced speech frames with embedded unvoiced speech frames, and embedded usable speech frames with embedded unusable speech. Some applications of this speech detection technique include the enhancement of speaker identification and speech recognition systems.

1. Introduction

Various nonlinearities exist in the production of human speech, including the variation in the shape of the glottal waveform as the amplitude of the speech changes, the non-laminar flow in the vocal tract and the prominent changes observed in the formant characteristics when the glottis is open. During speech classification, these factors need to be taken into account; i.e., a nonlinear representation of the speech signal is required.

Conventionally, speech analysis/classification has been performed based on linear properties of the given signal such as energy and autocorrelation; however, these properties cannot sufficiently represent the generation of speech in its true dynamics. Consequently, the problem arises as to how to obtain and depict the underlying dynamics from the one-dimensional speech signal. In other

words, how can the apparent one-dimension signal of speech be used in such a way as to illustrate the actual dynamics of the speech production system? One of the most popular representations of the chaotic nature of signals can be attained via Takens' embedding theorem [1], which states that a state space representation, topologically equivalent to the original state space of a system, can be obtained from a single dimension. This theorem has been applied in several signal processing applications including speech processing [2].

In this paper, speech classification based on features extracted from the state-space embedding technique is performed and results are compared with already existing speech classification measures.

2. Takens' Method Of Delays

A reasonable illustration of a nonlinear system could be to observe the system as a point moving along a trajectory in a conceptual state space where the independent degrees of freedom of the system make up the coordinates of the vector. Now, given a scalar system, a state space could be reconstructed using an embedding technique such as Takens' method of delays. In this technique, time-delayed values of the signal, x , are obtained and used to form an m -dimensional vector, v , which is mathematically represented as:

$$v(i) = [x(i), x(i-d), x(i-2d), \dots, x(i-(m-1)d)], \quad (1)$$

where d is the delay value in samples. The embedding dimension, m , depends mostly on the application of the technique, while the delay value is chosen based on the signal properties and sampling rate. ' d ' has to be large enough such that the nonlinearities are taken into account by the reconstructed trajectory but not too large to lose time resolution. Based on the non-stationary characteristics of speech signals, when applied to speech processing, the embedding technique must be

carried out on short (100 milliseconds or less) consecutive frames of speech.

3. Speech Detection

In this section, the application of state-space embedding for voiced speech detection as well as usable speech extraction systems is discussed.

3.1 Voiced Speech Detection

It has recently been shown that the use of only voiced segments of speech improves the speaker identification system, and also, that unvoiced speech contributes insignificant information about the speaker(s) for speaker identification [3]. Previously, Voiced/Unvoiced Classification had been performed by observing the energy, zero-crossings and residual energies of the speech signal [4], however, there still exists room for improvements especially when the focus is on voiced speech detection, that is, when the goal is not just to separate voiced speech from unvoiced, but to detect when voicing is present in speech as opposed to all other classes, which include transitions, unvoiced speech and silence. The generation of voiced speech constitutes a low-dimensional system as compared to non-voiced speech. Embedding voiced speech results in a well-structured signal, whereas embedded unvoiced speech is chaotic in structure. **Figure 1** below shows a frame of embedded voiced speech and a frame of embedded unvoiced speech, each with 128 points. Note the vast difference in structure between the voiced frame and the unvoiced frame. The vital question is then how to measure these structural differences between embedded voiced and unvoiced speech frames.

In this paper, two measures are introduced for voiced speech detection based on the structure of the embedded signal, and are discussed in detail in the following sections.

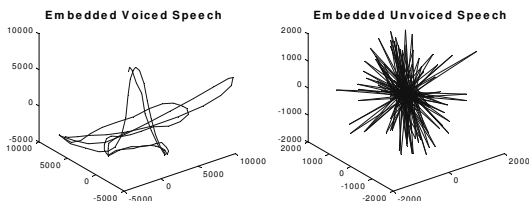


Figure 1: Embedded voiced speech frame (left panel) and embedded unvoiced speech frame (right panel)

3.2 Usable Speech Classification

Speech signals can be corrupted by two types of interference, background noise, or another speaker's

speech. The performance of speaker identification systems is known to be adversely affected by the presence of such interferences. Various techniques exist for the reduction or elimination of noise distortions in signals (including speech), however, due to the non-stationary properties of speech, complete removal of speech interferences has been a challenge to the speech processing industry. Speech interference occurs when two or more speakers are speaking simultaneously over the same channel without a significant difference in their overall energy. This research focuses on two speakers speaking on the same channel at the same time. The resulting speech is commonly termed “co-channel speech”.

When the energies of the target and interferer speeches are approximately equal, certain portions of speech still exist in co-channel speech in which the energy of one speaker is greater than the energy of the other speaker. These portions are termed “usable” while the other portions are termed “unusable”. The use of only ‘usable’ portions of speech has been shown improve the performance of speaker identification systems [3] [5] [6]. A Target (energy) to Interferer (energy) ratio (TIR) of 20dB is considered a suitable threshold for usable/unusable speech classification [6].

Usable speech classification techniques have also been introduced, which use linear-based approaches such as Spectral Auto-Correlation Peak-to-Value Ratio (SAPVR) [7], and Adjacent Peak Period Comparison (APPC) [8] along with others [9] [10] [11]. However, these methods do not take into account all the non-linear features of the signal, thereby ignoring valuable characteristics arising which could lead to more precise (or at least undiscovered) distinctions between heavily and slightly distorted speech signals, hence the reason to consider the use of the state-space embedding technique. Unvoiced speech and unusable speech are similar in structure, as the former is noise-like in nature, while the latter constitutes the presence of a significant amount of interference. Likewise, the structure of voiced speech is comparable to that of usable speech. Embedded voiced and usable speech frames are significantly more structured than embedded unvoiced and unusable speech frames.

Figure 2 below shows an embedded co-channel speech frame of 10dB TIR and 30dB TIR on the left and right panels respectively. It can be observed that the 10dB TIR (unusable) co-channel frame is more chaotic than the 30dB TIR frame. Once again, the

main question that arises is how to measure this difference in structure.

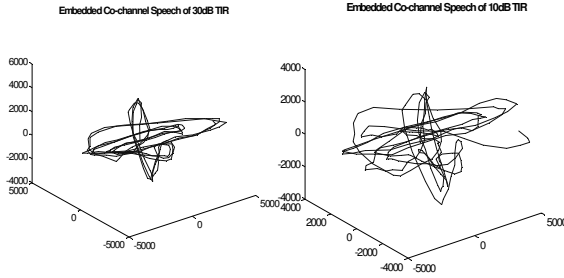


Figure 2: Embedded co-channel speech of 30dB TIR (usable speech) (left panel) and embedded co-channel speech of 10dB TIR (unusable speech) (right panel).

4. Measures

In this section, two measures of the differences between structured (voiced and usable) and unstructured (non-voiced and unusable) speech are discussed.

4.1 Difference-Mean Comparison (DMC)

Based on the difference in structure between embedded voiced/usable and unvoiced/unusable speech, it is expected that the rate of change of the embedded signal will provide a reasonable amount of separation between structured (voiced and usable) and unstructured (unvoiced and unusable) speech. **Figure 1** shows that there is a slow change in the positions of the points of the embedded voiced speech frame, whereas, the points on the embedded unvoiced speech frames change very rapidly. This rate of change can mathematically be evaluated by observing the M -th order difference (derivative) along the first non-singleton dimension of the matrix obtained by embedding the signal. For this research, since the signals are embedded in 3 dimensions, the 3rd order differences are computed. The derivatives of the unstructured signal are expected to be considerably higher than those of the structured signals. This difference between voiced and unvoiced speech can be enhanced by comparing the derivatives with the mean (magnitude) of each speech frame. Voiced speech has a considerably high magnitude as compared to unvoiced and the 3rd order derivative of its embedded signal is lower than that of unvoiced speech; therefore, comparing the derivative with the mean of the embedded speech frames provides a reasonable separation between voiced and unvoiced speech. For (structured) voiced speech, the derivative is less than the mean for most of the frames. The

reverse is the case for unvoiced speech. This measure, which is obtained by counting the number of times the third order derivatives of the embedded frames are greater than the mean of the magnitudes of any of the dimension (all three dimensions have about the same magnitudes as they are just time delayed versions of one another) is referred to as ‘**Difference-Mean Comparison (DMC)**’.

4.2 Nodal Density

Another distinguishable feature between structured (voiced and usable) and unstructured (unvoiced and unusable) speech, observable from **Figures 1** and **2** is the density of the signals. The structured signals are clearly less dense than the unstructured signals. One approach to measuring this density is by observing the volumes of the smallest enclosure of the reconstructed orbits, which are expected to be smaller for unstructured frames since the number of points (which could be considered as the mass) is constant. Another technique is to divide the space spanned by the embedded signal into n number of cubes, and define the edges of these cubes as nodes. Observing the number of nodes which are connected to points should result in a significant amount of separability between structured and unstructured frames, since unstructured frames should consist of more nodes due to its cluttered nature. This approach is referred to as the nodal density approach and has been applied to voiced as usable speech detection.

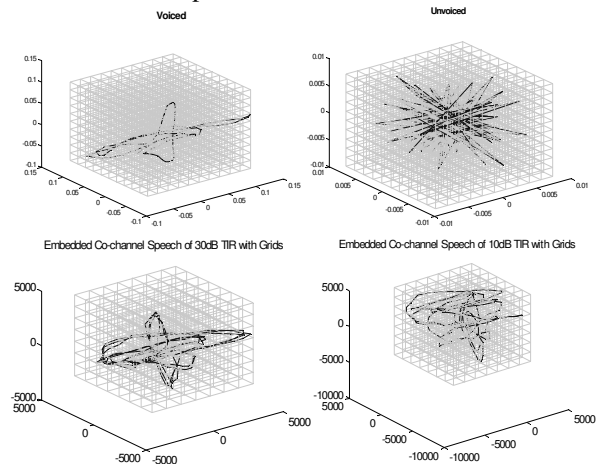


Figure 3: Embedded voiced (top-left panel) and unvoiced (top-right panel) speech frames, and Embedded co-channel speech of 30dB TIR (usable speech) (top-left panel) and 10dB TIR (unusable speech) (top-right panel), with grids to show nodes spanned.

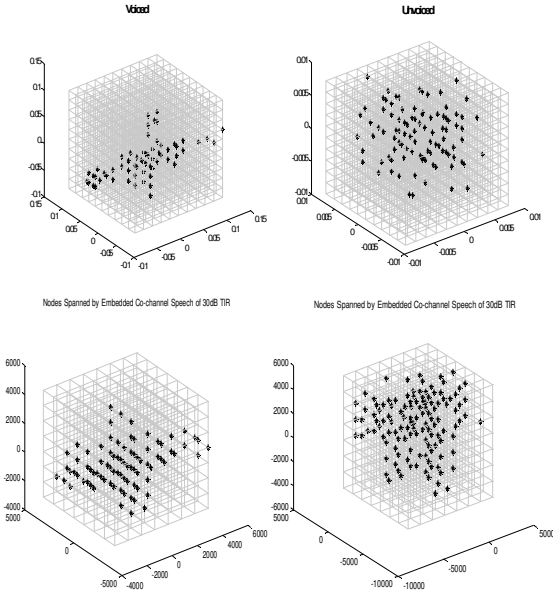


Figure 4: Embedded voiced (top-left panel) and unvoiced (top-right panel) speech frames, and Embedded co-channel speech of 30dB TIR (usable speech) (top-left panel) and 10dB TIR (unusable speech) (top-right panel), with points showing nodes spanned.

Figure 3 above shows embedded voiced (left) and unvoiced (right) speech frames (top panel) and usable (left) and unusable (right) speech frames (bottom panel) with grids to show the nodes spanned by the signal. **Figure 4** above shows the same frames as in **Figure 3**, excluding the lines but showing grid intersection points between signal and the nodes. From these figures, it is observed that the nodal densities of the unstructured speech frames are significantly higher than those of the structured speech frames.

5. Experiments And Results

All speech data used in the following experiments were obtained from the modified TIMIT database (13 female files and 12 male files). Voiced detection based on the DMC measure was carried out using the following procedure: First, the speech signal was passed through a 10th order moving average (lowpass) filter with a cut-off frequency of 1.3kHz to reduce the effect of noise on the signal (lowpass filtering significantly improved the results for noisy speech without very little effect on noiseless speech). Second, the filter output was then segmented into frames of 128 samples. Third, Takens' embedding technique was then applied on each frame. Fourth, the 3rd order

difference was computed for each embedded frame. Fifth, the mean of one dimension of the 3-dimensional matrix of each frame was then computed. Sixth, a count of the number of times the third order difference is greater than the mean was taken for each frame, which is the DMC value for the given frame. **Figure 5** below shows the block diagram of the process explained above.

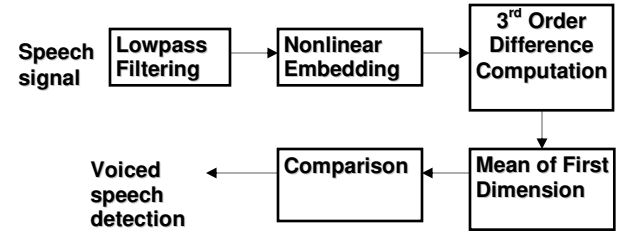


Figure 5: Illustration of the DMC-based voiced speech detection procedure

Based on the distribution of the DMC values, a threshold of 90 was found to provide decent separation between voiced and non-voiced speech even in noisy conditions; therefore, frames with DMC values of 90 and above were labeled non-voiced, while frames with values below 90 were labeled voiced.

Voiced detection based on the Nodal Density measure was performed using the following procedure: The first three steps were the same as those of the DMC measure. In the fourth step, the largest cube occupied by the each embedded frame was observed. Fifth, this cube was subdivided into 373 smaller cubes. Sixth, the number of nodes spanned by the signal was computed and divided by the total number of nodes, resulting in the nodal density. Distributions of the nodal densities were observed in various noise-conditions, and a threshold of 0.2 was chosen; therefore, frames with nodal densities of 0.2 and above were labeled non-voiced, while frames with values below 0.2 were labeled voiced.

The Nodal Density based usable speech detection was performed on all possible combinations of 41 different male and female speech files obtained from the TIMIT database. The procedure for the experiment was as follows: First, the target and interferer speech files were added together. Second, the voiced portions were then extracted from the resulting co-channel speech using the DMC measure as a voiced speech extraction system. Third, the extracted segments were then separated into frames

of 256 sample points each. The third to sixth steps of the Nodal Density voiced speed detection procedure were then carried out. The block diagram of the procedure described above is given in **Figure 6** below.

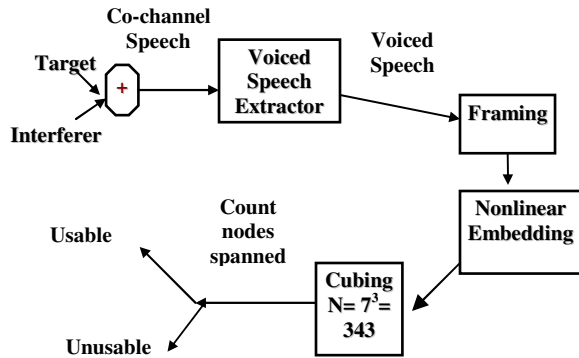


Figure 6: Illustration of the Nodal Density-based usable speech detection procedure

Figure 7 below shows the comparison of probabilities of error (PE) of the proposed measures and two traditional voiced/unvoiced classification measures [3] (used for detection voiced speech versus all other classes in this case as opposed to conventional voiced/unvoiced classification). The probabilities of error were computed by calculating the probability of misses (voiced samples detected as any other class by the measures) and the probability of false alarms (unvoiced (including transitions and silence) samples detected as voiced by the measures) and adding both probabilities. The first measure, a combination of two measures, the first order reflection co-efficient measure and the residual energy measure is shown as FR/RE, while the second, a combination of the energy measure and the zero-crossings measure is shown as E/ZC.

Note, from **Figure 7**, that the proposed measures perform better in the presence of high amounts of noise than the traditional measures, however, as the SNR increases, the performances of the proposed measure is not as good as that of the traditional methods.

Figure 8 shows the comparison of hits (usable samples – based on TIR – detected as usable by the measures) and false alarms computed for the proposed ND measure and the two recently developed usable speech detection measures that yielded the best results, SAPVR [7] and APPC [8]. The results in each case were obtained by choosing as a threshold the measure value which resulted in equal misses and false alarms. From **Figure 8**, it is

observed that the ND measure is comparable to the other two measures.

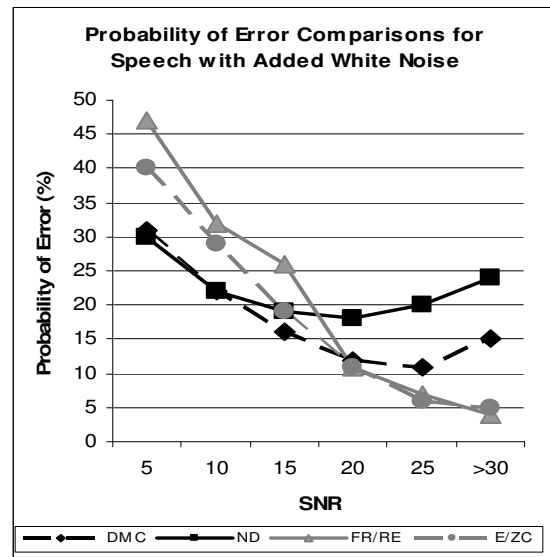


Figure 7: Probability of error curves for proposed (Difference Mean Comparison (DMC)) and Nodal Density (ND)) and traditional (First Order Reflection Coefficient and Residual Energy (FR/RE), and Energy and Zero Crossings (E/ZC)) voiced speech detection measures as a function of SNR.

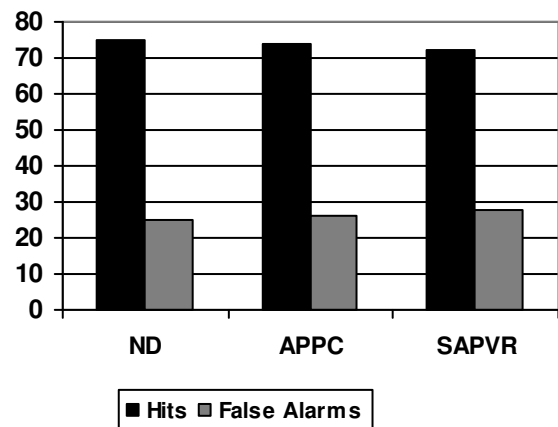


Figure 8: Hits and false alarms for Nodal density, APPC and SAPVR measures.

6. Discussion

The aim of this research was to detect if any given frame was voiced or usable (for speaker identification

and recognition purposes) based on the structure of their reconstructed trajectories. It was observed that the differences in the structure of embedded speech for voiced versus unvoiced or usable versus unusable yield a reasonable separation. The proposed technique has proven to yield better results than traditional methods in voiced speech detection classification under noisy conditions using the Difference-Mean-Comparison measure as well as the Nodal Density measure. The approach has also produced comparable results to recently developed usable speech detection classification methods using the Nodal Density Measure. Other features could be extracted from the state space embedding technique and used in voiced speech detection, one of which is the span of each embedded.

Research has shown that fusing usable speech detection systems is capable of improving the performance of the system [12] [13], therefore, fusing the ND measure with existing usable speech detection measures could yield better performance.

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. We also wish to express our appreciation to Ananth Iyer for his contribution to the Nodal Density measure.

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.

7. References

- [1] Takens, F., "Detecting strange attractors in turbulence", in *Lecture Notes in Mathematics*, Vol. 898, eds. D.A.Rand and L.S.Young, Springer, Berlin, 1981.
- [2] Terez, D. E., "Robust Pitch Determination Using Nonlinear State-Space Embedding", ICASSP, 2002.
- [3] Atal, B. S. and Rabiner, L. R., "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Transactions in Acoustic Speech and Signal Processing, vol. ASSP-24, No. 3, pp: 201-212, 1976.
- [4] Lovekin, J. M., Yantorno, R. E., Krishnamachari, K. R., Benincasa, D. B, Wenndt, S. J., "Developing Usable Speech Criteria for Speaker Identification", IEEE ICASSP 2001
- [5] Iyer, A. N., Smolenski, B. Y., Yantorno, R. E., Cupples, J., Wenndth, S., "Speaker Identification Improvement Using Usable Speech Concept", EUSIPCO 2004
- [6] Yantorno, R.E., "Co-channel speech study", Final report for Summer Research Faculty Program, Research Laboratory AFRL/IF, Speech Processing Lab, Rome Labs, New York, 1999
- [7] Krishnamachari, K. R., Yantorno, R. E, Benincasa, D. S. and Wenndt, S. J., "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions", IEEE ISPACS, 2000
- [8] Lovekin, J. M., Yantorno, R. E., Krishnamachari, K. R., Benincasa, D. B, Wenndt, S. J., "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions", IEEE ISPACS 2001
- [9] Sundaram, N., Yantorno, R. E., Smolenski, B. Y., Iyer, A. N., "Usable Speech Detection Using Linear Prediction Analysis – A Model Based Approach", ISPACS 2003.
- [10] Kizhanatham, A. and Yantorno, R.E., "Co-Channel Speech Detection Approaches Using Cyclostationarity or Wavelet Transform", IASTED SIP
- [11] Krishnamachari, K. R., Yantorno, R. E., Lovekin J. M., Benincasa, D. S., and Wenndt, S. J., "Use of Local Kurtosis Measure for Spotting Usable Speech Segments in Co-channel Speech." ICASSP 2001
- [12] Yantorno, R. E., Smolenski, B. Y., Chandra, C., "Usable speech Measures and their Fusion". ISPASS, 2003
- [13] Shah, J. K., Smolenski, B. Y., Yantorno, R. E., "Decision Level Fusion of Usable Speech Measures Using Consensus Theory", ISPACS, 2003