# VOICING-STATE CLASSIFICATION OF CO-CHANNEL SPEECH USING NONLINEAR STATE-SPACE RECONSTRUCTION

*Y. A. Mahgoub and R. M. Dansereau*

Carleton University, Department of Systems & Computer Engineering
1125 Colonel By Drive, Ottawa, ON, K1S 5B6, Canada

## ABSTRACT

This paper presents a new classification method to determine the voicing-state of co-channel speech based on nonlinear state-space reconstruction. Nonlinear approaches are known to give a better access to the full dynamics of speech system than linear techniques. Three voicing-states of co-channel speech are considered; Unvoiced/Unvoiced (U/U), Voiced/Unvoiced (V/U), and Voiced/Voiced (V/V). The proposed method requires neither *a priori* information nor speech training data. Nonetheless, simulation results show enhanced performance in identifying the three voicing-states using the proposed method compared to other existing techniques.

## 1. INTRODUCTION

Co-channel speech is defined as the composite speech signal of two or more talkers [1]. This phenomenon commonly occurs due to the combination of speech signals from simultaneous and independent sources into one signal at the receiving microphone, or when two speech signals are transmitted simultaneously over a single channel. The use of a voicing-state classifier is essential in many applications where co-channel speech might occur. This includes Automatic Speech Recognition (ASR), Speaker Identification (SID), and speech enhancement techniques.

Most of the conventional techniques used to classify voicing-state of single and co-channel speech rely on the pattern recognition approach and treat the speech system as a linear system [2], [3], [7]. However, the application of nonlinear dynamical methods to speech characterization and analysis has produced numerous new and promising approaches over the last two decades. For example, the results in [4] and [5] have shown enhanced performance in solving the general problems of pitch determination and speech enhancement by using nonlinear methods compared to the linear techniques. Nonlinear approaches are known to give a better access to the full dynamics of speech system than linear techniques.

Previous work on voicing-state classification of co-channel speech has shown some success using either *a priori* information about the individual speakers [6] or training data sets [7]. However, *a priori* information is not always available in many practical situations. Also, methods that use training data sets are speaker- and environment-dependent. Every time the recording conditions or the background noise level change, a new set of training data is required. In [8], it has been attempted to locate only the “usable speech” segments (single-talker voiced frames) using the Spectral Autocorrelation Peak Valley Ratio (SAPVR) technique. No great attention has been given to the other voicing-state classes of co-channel speech.

In this paper, a new voicing-state classification algorithm for co-channel speech, based on nonlinear state-space reconstruction, is proposed. Three voicing-states are considered in this study:

1. Unvoiced/Unvoiced (U/U): where both speakers are either in the unvoiced state or the silence state.
2. Voiced/Unvoiced (V/U): where only one speaker is in the voiced state.
3. Voiced/Voiced (V/V): where both speakers are in the voiced state.

The silence state is assumed to be a subset of the unvoiced class. Also, no need to differentiate between speakers in the V/U class is assumed. In Sec. 2, the principle of state-space reconstruction is explained. The new proposed method is described in Sec. 3. Comparisons of the proposed algorithm to the other existing techniques with the aid of computer simulations are presented in Sec. 4. Finally, Sec. 5 includes the conclusions reached.

## 2. STATE-SPACE RECONSTRUCTION

State-space (also called phase-space) reconstruction is the first step in nonlinear time series analysis. It basically views a single-dimensional data series, $s(n)$; $n = 1, 2, \ldots, N$, in an $m$-dimensional Euclidean space, $\mathbb{R}^m$. Using this method, the trajectories that connect data points (vectors) in the state-space are expected to form an attractor that preserves the topological properties of the original unknown attractor. A common way to reconstruct the state-space is the method of delays introduced by

0-7803-8874-7/05/$20.00 ©2005 IEEE

Takens [9]. In this method, $m$-dimensional vectors, $\mathbf{x}_n$, in the state-space are formed from the time-delayed samples of the original signal, $s(n)$, as follows

$$\mathbf{x}_n = \left[s(n), s(n-d), s(n-2d), \cdots, s(n-(m-1)d)\right] \quad (1)$$

where $d$ is the embedding delay, and $m$ is the embedding dimension (number of coordinates). Delay reconstruction requires a proper choice of the parameters $d$ and $m$. The value of $d$ can be calculated as the time (in samples) of the first zero crossing of the autocorrelation function. This allows opening up the attractor in the state-space reconstruction. The embedding dimension, $m$, has to be chosen large enough (e.g., $m = 5$ in the present application) in order to avoid false neighbor trajectories.

## 3. METHOD DESCRIPTION

Voiced-speech is well known with its quasi-periodic nature. Thus, it can be fully represented in a low dimensional state-space. This reconstruction is pitch synchronous such that one revolution of the attractor is equivalent to one pitch period. Furthermore, the trajectories of voiced-speech with temporal distance of multiple of pitch periods tend to be close and parallel to each other. Co-channel and unvoiced speech, on the other hand, do not have this quasi-periodic nature and therefore require a higher embedding dimension. Figure 1 shows three speech frames with different voicing-states; V/U, V/V and U/U, and their corresponding state-space reconstructions, respectively, for $m = 3$. It is obvious from the state-space plots that the degree of similarity in the neighbor trajectories can be used as a good indicator for the presence of voiced-speech. This can be achieved by calculating the Trajectory Parallel Measure (TPM) of the attractor. The TPM idea was first proposed in [10] and modified in [11] to detect determinism in speech phonemes. If $\mathbf{a}$ is the query trajectory at time index $n$ and $\mathbf{b}_i$ is the $i^{th}$ nearest trajectory to $\mathbf{a}$, then

$$\begin{aligned} \mathbf{a} &= \mathbf{x}_{n+1} - \mathbf{x}_n \\ \mathbf{b}_i &= \mathbf{x}_{m+1} - \mathbf{x}_m \end{aligned} \quad (2)$$

where $\mathbf{x}_n$ is the query point and $\mathbf{x}_m$ is the nearest neighbor point on the $i^{th}$ trajectory. The angle between $\mathbf{a}$ and $\mathbf{b}_i$ is given by

$$\theta_i(n) = \cos^{-1}\left(\frac{\mathbf{a} \cdot \mathbf{b}_i}{|\mathbf{a}||\mathbf{b}_i|}\right) \quad (3)$$

The TPM of the whole attractor is calculated as

$$TPM\ \% = \frac{100}{N}\sum_{n=1}^{N} H\left(\alpha - \frac{1}{\pi L}\sum_{i=1}^{L}\theta_i(n)\right) \quad (4)$$

where $N$ is the total number of samples in the frame, $\alpha$ is a constant threshold, $L$ is the total number of nearest trajectories, and $H(\cdot)$ is the Heaviside function.

The proposed algorithm of classifying a co-channel speech frame into one of the three voicing states; U/U, V/U, or V/V works as follows:

1. Determine whether the present frame is unvoiced or not by using the following two measures [2], [3].

- STE, the frame short-time energy in dB.

$$STE = E = 10 * \log\left(\varepsilon + \frac{1}{N}\sum_{n=1}^{N} s^2(n)\right) \quad (5)$$

where $\varepsilon$ is a small constant to avoid calculating $\log(0)$.

- HILO, the ratio of the energy in the signal above $f_H$ Hz to that below $f_L$ Hz.

An STE threshold, $E_{th}$, is used and is defined as

$$E_{th} = 0.1 * \frac{1}{M'}\sum_{k=1}^{M} E_k \quad (6)$$

where $E_k$ is the STE of the $k^{th}$ frame as given by (5), $M$ is the total number of frames and $M'$ is the total number of non-silent frames. The reason for averaging frame energies over $M'$ instead of $M$ in (6) is to reduce the effect of silent frames on the energy threshold calculation. All speech frames are classified as unvoiced (U/U) except when the STE is higher than $E_{th}$ and the HILO ratio is lower than a chosen threshold, $HILO_{th}$.

2. For the frames not classified as unvoiced in step 1, reconstruct the state-space by using the method of delay embedding (1) with a suitable embedding dimension $m$ and time delay $d$.

3. For a vector $\mathbf{x}_n$ in the state-space, locate the nearest neighbor points on the nearest neighbor trajectories. A nearest neighbor, $\mathbf{x}_m$, is found by searching for the point that minimizes the distance to the query vector, $\mathbf{x}_n$

$$\min_{\mathbf{x}_m} \|\mathbf{x}_n - \mathbf{x}_m\| \quad (7)$$

A further constraint is imposed such that nearest neighbors must have a temporal separation greater than the minimum pitch period (about 2.5 ms). The search for the nearest neighbors is also limited to a maximum spatial distance of $r\sigma_x$ in the state-space; where $r$ is a constant and $\sigma_x$ is the standard deviation of the frame samples.

4. Calculate the overall TPM of the given frame using (4). If the TPM is greater than a threshold $TPM_{th}$, then this frame is classified as V/U. Otherwise it is considered as a V/V frame.

## 5. SIMULATION RESULTS

In order to evaluate the performance of the proposed algorithm, a computer simulation is carried out using the TIMIT database. A total of 60 co-channel mixtures consisting of the speech of male/male, female/female and female/male is created. All the mixtures are tested at a signal to interference ratio (SIR) of 0 dB.

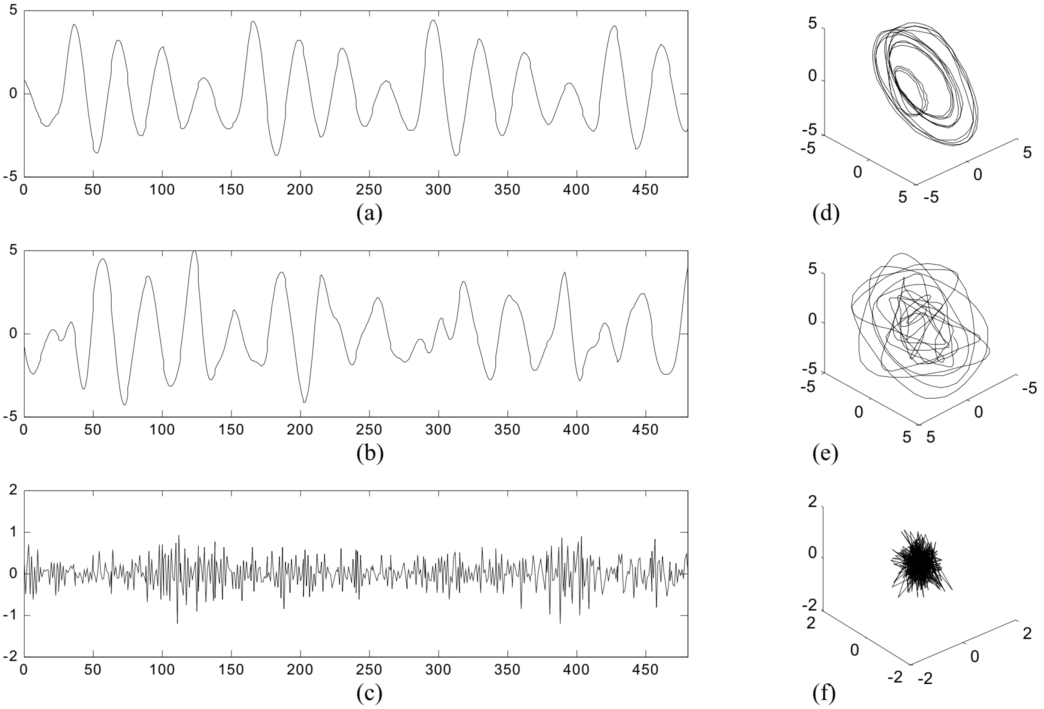

Figure 1. Three frames of (a) single-voiced, (b) double-voiced and (c) unvoiced speech signals and their corresponding state-space reconstructions (d), (e) and (f), respectively, for $m = 3$.

The speech data are sampled at 16 kHz and segmented into 30 ms frames with 50% overlap. The following values of different parameters are used:

$$f_L = 2 \text{ kHz} \quad f_H = 4 \text{ kHz} \quad HILO_{th} = -3 \text{ dB}$$

$$m = 5 \quad \alpha = 0.25 \quad r = 0.6 \quad TPM_{th} = 70\%$$

First, the voicing-states of the speech of every speaker alone are determined to have a baseline for comparison. This produces two sets of 2-level labels for speaker #1 and speaker #2 in which a value of 1 is given to every voiced frame and a value of 0 is given to every unvoiced frame. The two sets of labels are then added together to create a third set of 3-level labels. This set is called the reference set in which a value of 0 corresponds to U/U class, a value of 1 corresponds to V/U class and finally, a value of 2 refers to V/V class. Reference sets are further examined by visual inspection to correct any errors. Voiced-speech frames with SIR less than 15 dB are considered V/V. Finally, a fourth set of 3-level labels is obtained from the mixed speech by applying the proposed approach. The last two sets are compared to determine if any error has been occurred.

Tables 1, 2, and 3 summarize the final results obtained for the three different mixtures. It can be noticed from the tables that the algorithm's capability in determining the U/U state is better than V/U and V/V. However, the overall performance still exceeds 85%. The accuracy of determining V/U frames is higher than the V/V frames for female/female mixtures while the opposite is true for male/male mixtures. The performance of determining the two classes is approximately the same for female/male mixtures. This is due to female voiced-speech having a higher pitch frequency (and consequently greater number of pitch periods per frame) than male voiced-speech. This results in more neighbor trajectories in the state-space.

Three major sources of error are observed from the simulations:

1. Transition frames (onsets and offsets of voiced-speech).
2. Frames with mixed excitation such as voiced fricatives and plosives.
3. Frames when the pitch frequency of one speaker is approximately an integer multiple of the other speaker's pitch frequency.

While it is sometimes easy to locate transition frames during single talking, it is quite difficult to do so with co-channel speech. To reduce the effect of the transition frames on the final performance, a three-tap median filtering is used on the 3-level label sets.

Table 4 shows a comparison of the presented results to the results given in [7] using the Bayesian approach and the results given in [8] using the SAPVR approach. A total increase of at least 7% on the overall percentage of correctly identified segments is achieved. Taking into consideration that the proposed algorithm does not use training data, this gives it a great advantage over the other two algorithms.

Table 1: Performance of the proposed classifier on male/female mixtures (overall performance: 85.1%).

| Estimated voicing-state | Reference voicing-state | | |
|---|---|---|---|
| | U/U | V/U | V/V |
| U/U | 99 % | 2.8 % | 2.4 % |
| V/U | 0.2 % | 75.5 % | 16.8 % |
| V/V | 0.8 % | 21.7 % | 80.8 % |

Table 2: Performance of the proposed classifier on male/male mixtures (overall performance: 84%).

| Estimated voicing-state | Reference voicing-state | | |
|---|---|---|---|
| | U/U | V/U | V/V |
| U/U | 98.4 % | 5.4 % | 4.6 % |
| V/U | 0.2 % | 68.8 % | 10.5 % |
| V/V | 1.4 % | 25.8 % | 84.9 % |

Table 3: Performance of the proposed classifier on female/female mixtures (overall performance: 86%).

| Estimated voicing-state | Reference voicing-state | | |
|---|---|---|---|
| | U/U | V/U | V/V |
| U/U | 96.2 % | 3.6 % | 2.9 % |
| V/U | 1.5 % | 87.7 % | 22.9 % |
| V/V | 2.3 % | 8.7 % | 74.2 % |

Table 4: Performance comparison of the proposed classifier with other algorithms.

| Voicing-state | Percentage of correct identifications | | |
|---|---|---|---|
| | Bayesian | SAPVR | Proposed |
| U/U | 85.9 % | N/A | 97.9 % |
| V/U | 59.8 % | 71 % | 77.3 % |
| V/V | 90.8 % | N/A | 80 % |

## 4. CONCLUSIONS

A new classification method to determine the voicing-state of co-channel speech is introduced. The method is based on nonlinear state-space reconstruction of speech data. Simulation results show that a total percentage of 85% correctly identified segments is achieved. The proposed method therefore outperforms other techniques with the advantage that no *a priori* information or training is required.

## 5. REFERENCES

[1] R. E. Yantorno, "Co-Channel Speech Study," *Final Report for Summer Research Faculty Program*, Air Force Office of Scientific Research, Speech Processing Lab, Rome Labs, September 1999.

[2] B. S. Atal and L. R. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-24, no. 3, pp. 201-212, June 1976.

[3] L. J. Siegel and A. C. Bessey, "Voiced/Unvoiced/Mixed Excitation Classification of Speech," *IEEE Trans. on Acoust., Speech, Signal Processing*, vol. ASSP-30, no. 3, pp. 451-460, June 1982.

[4] D. Terez, "Robust Pitch Determination Using Nonlinear State-Space Embedding," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing (ICASSP'02)*, vol. 1, pp. I-345 - I-348, May 2002.

[5] M. T. Johnson, A. C. Lindgren, R. J. Povinelli, and X. Yuan, "Performance of nonlinear speech enhancement using phase space reconstruction," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing (ICASSP'03)*, vol. 1, pp. I-920 - I-923, April 2003.

[6] M. Weintraub, "A Computational Model for Separating Two Simultaneous Talkers," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing (ICASSP'86)*, vol. 11, pp. 81-84, April 1986.

[7] D. S. Benincasa and M. I. Savic, "Voicing State Determination of Co-channel Speech," *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing (ICASSP'98)*, vol. 2, pp. 1021-1024, May 1998.

[8] N. Chandra, and R. E. Yantorno, "Usable Speech Detection Using the Modified Spectral Autocorrelation Peak to Valley Ratio Using the LPC Residual," *Proc. of 4th IASTED International Conference, Signal and Image Processing*, pp. 146-149, 2002.

[9] F. Takens, "Detecting strange attractors in turbulence," In D. A. Rand and L. S. Young, editors, *Dynamical Systems and Turbulence*, Warwick 1980, Lecture Notes in Mathematics 898, Springer, Berlin, 1981.

[10] J. McNames, "A Nearest Trajectory Strategy for Time Series Prediction," *Proc. Int. Workshop on Advanced Black-Box Techniques for Nonlinear Modeling*, pp. 112-128, July 1998.

[11] X. Liu, R. J. Povinelli, M. T. Johnson, "Detecting Determinism in Speech Phonemes," *Proc. IEEE 10th Digital Signal Processing Workshop and the 2nd Signal Processing Education Workshop*, pp. 41–46, October 2002.