PHASE-SPACE REPRESENTATION OF SPEECH
— REVISITING THE DELTA AND DOUBLE DELTA FEATURES

Hua Yu

Interactive Systems Labs, Carnegie Mellon University, Pittsburgh, PA 15213
hyu@cs.cmu.edu

ABSTRACT

Speech production is essentially a nonlinear dynamic process. Motivated by ideas in dynamic system research, this paper seeks to recast the speech representation problem (front-end) as an attempt to reconstruct the phase space of the production process, or articulatory configurations. In particular, we point out that the use of the delta and double delta features, common in current ASR (Automatic Speech Recognition) systems, corresponds to time-delayed embedding, a technique in nonlinear time series analysis. We will show that due to various assumptions in the modeling framework, time-delayed embedding is naturally required for a first order HMM (Hidden Markov Model) to handle higher order dependencies. Furthermore, we examine the effect of linear transformations in the phase space and present experimental results.

1. INTRODUCTION

The configuration of the human vocal tract, which "shapes" speech acoustics, depends on the position of various speech articulators, such as tongue, lips, jaw, velum, and larynx. It is the behavior of the articulators over time that produces continually varying acoustics. A recurrent belief among speech researchers is that what the listener extracts from the speech signal might be information about the speech production process itself [1].

For automatic speech recognition purposes, it would then be natural to represent speech in a way that captures the dynamics of the production process. In dynamical system research, the dynamics of a physical system can be described mathematically in *phase space* or *state space*. Each dimension of the space represents an independent state variable of the system, such as position or velocity. Each point in the phase space corresponds to a unique state of the system. The evolution of a system over time produces a *phase portrait* in the phase space. Much can be learned about the dynamics of a system from its phase portrait. An example is shown in Figure 1, where articulatory movements are measured while the subject is producing the syllable /ba/ repeatedly [2]. The left panel shows the traditional time domain measurements of jaw and lower lip movements; the right panel shows the corresponding phase portraits for the two articulators, plotted in a plane of position vs. instantaneous velocity.

Certain aspects become readily apparent in the phase portraits. The most visible is the repetitive syllable pattern. Each circle represents an instance of /ba/, where the half denoted as CLOSED corresponds to /b/, OPEN for /a/. Inter-syllable events, such as stress, can be seen as alternating patterns of larger and smaller cycles. It is also clear that the motion of the articulators is less variable during the production of the consonant (denoted as CLOSED) than of the vowel (denoted as OPEN). In addition, inter-articulator timing (articulatory synchrony/asynchrony) can be studied if we plot a phase space that covers multiple articulators (not shown here).

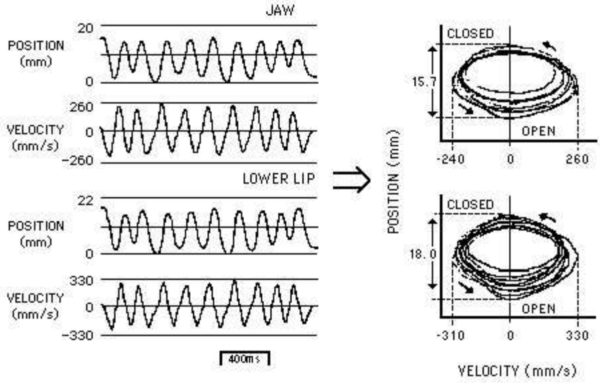


Fig. 1. Phase portraits of two articulators during production of reiterant /ba/. (From Kelso et al. 1985, ©1985 Acoustical Society of America)

If measurements of various articulators can be made easily and accurately, it will be an inherently superior representation than one based on acoustics. It gives a more direct access to the information source, and besides, there is less contamination by noise or channel distortion. Generally, however, only the speech signal is available to an ASR system. Therefore, it would be desirable to reconstruct the phase space from acoustics. Indeed, it has been shown that a surprisingly simple technique called *time-delayed embedding*, can produce a one-to-one image of the dynamics of the original system. In recent years, there has been a growing interest to represent speech using this technique [3, 4].

Rather than the straightforward approach of applying time-delayed embedding at the time domain, we argue that for ASR purposes, it is more appropriate to reconstruct phase space at a higher level, such as the cepstral level. As a matter of fact, we will show that the delta and double delta features, commonly used in ASR, is indeed a form of time-delayed embedding. Hence, the incorporation of dynamic features *is* an attempt to reconstruct the phase space of the speech production system.

The use of dynamic features would not be necessary, if the underlying model can capture higher order dependencies. For a first order Markov model, we will show that time-delayed embedding effectively transforms it into a higher order Markov model. For HMMs in speech recognition, embedding extends the feature vector from a single frame to a segment, increasing the mutual information between the feature vector and its class label.

Since embedding typically results in a high dimensional space, linear projection is commonly used to reduce dimensionality. Having a proper transformation is crucial for accurate modeling.

This paper is organized as follows. First, we give a short introduction to phase space reconstruction and time-delayed embedding. Section 3 explains why we should apply reconstruction at the cepstral domain. In Section 4, we show how the underlying HMM modeling framework naturally requires time-delayed embedding. The effect of linear transformation in the reconstructed space is discussed and experimental results presented in Section 5.

2. PHASE SPACE RECONSTRUCTION USING TIME-DELAYED EMBEDDING

2.1. The Embedding Theorem

Phase space or *state space* is an important concept, widely used in physics and dynamic system research, for studying the dynamics of a system. It is a vector space, where each dimension represents an independent variable of the system under study. A simple mechanical system can be described in a phase space of two dimensions: position versus velocity, commonly seen in physics text. A complex system with many degrees of freedom needs a high dimensional phase space. Each point in the space specifies a unique state of the system, and vice versa. When the system evolves over time, the point traces out a trajectory in the phase space: $\{\vec{\boldsymbol{x}}_n\}$.

In many cases, the system is not fully observable. We may only get a scalar measurement one at a time, denoted by $\{s_n\}$. Vectors in a new space, the embedding space, are formed from time-delayed values of the scalar measurements:

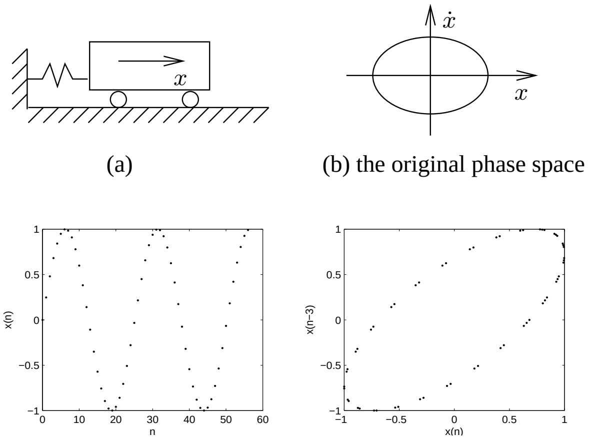
$$\vec{\boldsymbol{s}}_n = (s_{n-(m-1)\tau}, s_{n-(m-2)\tau}, \cdots, s_n)$$

The number of samples m is called the *embedding dimension*, the time τ is called *delay* or *lag*. The celebrated reconstruction theorem by Takens states that under certain general assumptions, time-delayed embedding $\{\vec{\boldsymbol{s}}_n\}$ provides a one-to-one image of the original set $\{\vec{\boldsymbol{x}}_n\}$, provided m is large enough [5].

Time-delayed embedding is a fundamental tool to investigate chaotic behavior of nonlinear systems. For a detailed discussion, as well as how to choose the right value for m and τ, readers are referred to [6].

2.2. A Linear Oscillator Example

For simplicity, we use a linear system here to illustrate the idea of phase space and phase space reconstruction.



(a) (b) the original phase space

(c) time series observation (d) reconstructed phase space

Fig. 2. A linear oscillator, its phase portrait and reconstructed phase space from time series observation

Consider a linear oscillator consisting of a mass attached to a linear elastic spring (Figure 2(a)). According to Newton's law of motion, the acceleration of the object is the total force acting on the object divided by the mass.

$$\ddot{x} = \frac{f}{m}$$

Assuming no friction, the spring force f is proportional to the amount that the spring has been compressed, which is equal to the amount that the object has been displaced.

$$f = -kx$$

Combining the two, the system dynamics can be uniquely described by

$$\ddot{x} = -\frac{k}{m}x \tag{1}$$

Solving this differential equation, we have

$$x = a\sin(wt + b)$$

where $w^2 = \frac{k}{m}$, the values of a and b depend on the initial condition.

The phase space for such a system is typically $(x, \dot{x})$. The system moves along a closed ellipse periodically (Figure 2(b)). When friction is taken into account, the phase portrait is an inward spiral, since the system will gradually lose velocity.

Now, suppose we only observe a time series $\{x_n\}$, under a certain sampling rate (Figure 2(c)). The reconstructed phase space is shown in Figure 2(d), where the embedding dimension $m = 2, \tau = 3$. Clearly the reconstructed phase portrait has the same structure as the original system, although a strong correlation exists between the delayed coordinates.

2.3. Chaotic Systems

Time-delayed embedding comes from and is used extensively in the study of chaotic systems, which have been found to be quite common in daily life. Speech, among other things, has been shown to be chaotic. The phase portrait of a chaotic system is very complex, with the existence of strange attractors as a hallmark. Examples of chaotic systems, their phase portraits, and reconstruction of their dynamcis can be found in many books and websites.

3. EMBEDDING IN THE CEPSTRAL DOMAIN

3.1. Why Embedding in the Cepstral Domain

In recent years, there has been a growing interest in applying phase space reconstruction to speech recognition [3, 4]. In the classic source-filter model, speech signal is the combined outcome of a sound source (excitation) modulated by a transfer (filter) function determined by the shape of the supralaryngeal vocal tract. This model is based on the linear system theory. So are most traditional speech parameterization, such as the linear predictive coding. It has been argued that phase space reconstruction, as a nonlinear time series analysis technique, fits better with the nonlinear nature of speech. Using delayed embedding directly on the time domain signal, various chaotic features (such as correlation dimension and Lyapunov exponents) are extracted as the basis for recognition. It is reported that although the new chaotic feature does not outperform the traditional MFCC (Mel-Frequency Cepstral Coefficients) feature, a combination of the two tends to improve recognition accuracy.

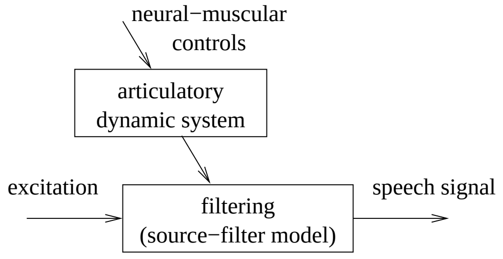


Fig. 3. Two Sub-systems in Speech production

Upon closer examination, there are really two systems involved in speech production (Figure 3): the filtering system (as in source-filter model) and the articulatory system. The coordinated motion of various articulators determines the shape of the vocal tract, which then filters the sound source, producing speech signal. Since the ultimate goal of ASR is to infer the phase space of the articulatory system, it is more appropriate to start from a representation of the instantaneous vocal tract shape, rather than directly from the speech signal.

According to the traditional theory, cepstral coefficients are designed to capture the spectral envelope, which is largely determined by the shape of the vocal tract [7]. In other words, cepstrum is a fairly good representation of the vocal tract characteristics. It gives a reasonable source/vocal tract separation. Working in the cepstral domain allows us to focus on the (nonlinear) dynamics of the articulatory system, whereas the dynamics reconstructed from the time domain contains the compounding effects of two systems.

3.2. Delta and Double-delta Features

Delta and double delta features are originally introduced in [8] to incorporate dynamic information into an ASR system:

$$\vec{\Delta}_i = -3\vec{x}_{i-3} - 2\vec{x}_{i-2} - \vec{x}_{i-1} + \vec{x}_{i+1} + 2\vec{x}_{i+2} + 3\vec{x}_{i+3} \tag{2}$$

$$\Delta\vec{\Delta}_i = -3\vec{\Delta}_{i-3} - 2\vec{\Delta}_{i-2} - \vec{\Delta}_{i-1} + \vec{\Delta}_{i+1} + 2\vec{\Delta}_{i+2} + 3\vec{\Delta}_{i+3}$$

where $\vec{x}_i$ is a 13-dimensional cepstral vector. These features lead to significant improvements and have since been widely used in ASR. The original formular (Eqn. 2) is a special case of a more general scheme, where several adjacent frames of cepstral vectors are stacked together to form a super vector $(\vec{x}_{-6}, \vec{x}_{-5}, \cdots, \vec{x}_6)$, then projected down to a lower dimension space by a linear transform.

It should be clear now that modulo the linear transform, dynamic features are exactly time-delayed embedding in the cepstral domain. This leads to a revelation that the incorporation of dynamic features has a fundamental meaning, which is to recover the phase space of the speech production system, i.e. the time-varying articulatory configuration.

There are several caveats, though. First, we are embedding a vector series, not a scalar time series. This is equivalent to taking simultaneous measurements of multiple variables of a system and not a problem at all. Second, the speech production process is not deterministic. The existence of measurement noise (environmental noise and channel distortion) further complicates the picture of the reconstructed dynamics. Some of the issues are discussed in [6, 9].

One may argue that after all, delayed embedding is only a different representation of the data, without introducing any new information. In the case of speech recognition, we need to justify any changes at the feature level with respect to the underlying modeling framework. The next section will show why time-delayed embedding is essential for HMMs.

4. TIME-DELAYED EMBEDDING AND HMMS

As many researchers have pointed out, HMMs fail to capture speech dynamics accurately, due to the conditional independence assumption: each frame is conditionally independent of each other, given the state sequence. Several alternative approaches have been proposed to compensate for this weakness, including segmental models, Gaussian transition models, etc. [10, 11, 12]. Unfortunately, these sophisticated models have yet to show improvements over the seemingly simple HMMs.

Part of the reason is due to the use of time-delayed embedding, i.e. delta and double-delta features. By changing the feature representation, each feature vector now covers a window of consecutive frames, rather than a single frame. Hence, the entity being modeled with HMMs is an entire segment, typically around 100 milliseconds in duration, rather than a single frame of ∼20 milliseconds. In a sense, this is segmental modeling in disguise.

The effect of time-delayed embedding on the underlying model can be more formally established in the following scenarios.

4.1. Deterministic Dynamic Systems

It is well known that for dynamic systems that can be described by differential equations, a set of first order differential equations is sufficiently general to represent second or higher order systems.

In the above example of a second order linear oscillator, it is easy to convert the system equation to a set of first order differential equations. By introducing a new variable $y = \dot{x}$, equation 1 can be rewritten as

$$\left\{ \begin{array}{lcl} \dot{x} & = & y \\ \dot{y} & = & -\frac{k}{m}x \end{array} \right.$$

In the phase space of (x, y), this is a first order system. In the same spirit, first order Markov models can be elevated to a higher-order model by phase space reconstruction.

4.2. Markov Models

A Markov model of order m is a model where the probability at time t depends only on the states of the last m steps. These last m steps define the state of the system. Hence, using time-delayed embedding of the past m samples, the state of the system can be accurately determined.

If the data indeed comes from an m-th order Markov source, we need m-dimensional embedding to model it properly with a first order HMM, since now the probability of the next state (or observation) depends only on the current state (or observation).

4.3. Hidden Markov Models

Markov models can be thought of as a special case of HMMs where there is a one-to-one correspondence between states and observations, i.e. states are not hidden. For HMMs, we can no longer strictly prove that a first order HMM can model an mth order source using mth dimensional delayed embedding. It may be a little difficult here to think of HMMs as a generative model in this context. Nonetheless, from a discriminative point of view, each delay vector contains more information about the identity of the HMM state than a single frame.

This is also related to the *false nearest neighbor* method, commonly used in nonlinear time series analysis to determine the minimal sufficient embedding dimension [6]. If the embedding dimension m is less than the dimensionality of the original system, the reconstructed dynamics won't be a one-to-one image of the original attractor. Instead, "folding" will occur: points are projected into neighborhoods of other points to which they don't belong to. False nearest neighbor can therefore be used as a test for insufficient embedding dimension.

Similarly, with no embedding or insufficient embedding dimension, the feature vector in ASR doesn't carry enough information to accurately determine the state of the articulatory system. Hence, embedding empowers a first order HMM by increasing the mutual information between feature vectors and their class labels.

5. LINEAR TRANSFORMATION OF THE PHASE SPACE

A linear transformation of the phase space does not change the validity of the embedding theorem. It can actually lead to a better representation of the data. As shown in Figure 2(d), a strong correlation exists between the delayed measurements, which is irrelevant to the structure of the system dynamics. Derivative coordinates (similar to delta and double delta) and principal component analysis have been proposed as alternatives to delayed coordinates [6]. Both are linear transforms of the original phase space.

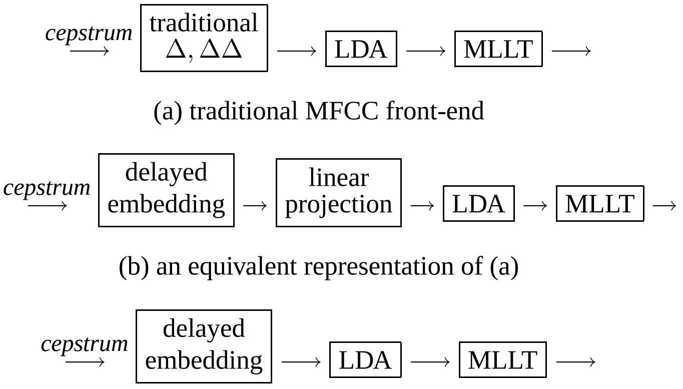
5.1. Front-End for Speech Recognition

For speech recognition, however, the situation becomes a little more complicated. The front-end in a modern ASR system has many components, some linear and some nonlinear, each serving a different purpose. A discussion on streamlining various linear transformations can be found in [13]. In short, there are two key issues in the design of a front-end. First, dimensionality reduction is inevitable. It happens at many places in a typical MFCC (Mel-Frequency Cepstral Coefficients) front-end. Some are easy to recognize, such as LDA (Linear Discriminant Analysis). Some are less obvious, such as spectral smoothing using the Mel-scale filterbank, truncating of the cepstrum, as well as the extraction of the delta and double delta features. While it makes training more feasible, dimensionality reduction loses information, and therefore, should be performed with great care. Second, the front-end needs to accommodate various assumptions made in the acoustic modeling framework, such as the use of diagonal covariance matrices (instead of full covariance matrices). It turns out that semi-tied covariance [14] with a single class can be easily implemented as a feature space transform. This is also known as Maximum Likelihood Linear Transform (MLLT). Recently, there has been an effort to achieve more sophisticated covariance tying, by constraining the inverse covariance matrices to be a linear combination of many rank one matrices (EMLLT, [15]), or more generally, symmetric matrices [16]. In a sense, the front-end has become an inseparable part of the acoustic model.

Below, we focus on the derivation of dynamic features and subsequent linear transforms.

5.2. Experiments

As discussed before, delta and double delta features are simply a linear transform of the reconstructed phase space. The traditional approach (Eqn. 2) is shown in Figure 4(a). Figure 4(b) shows the equivalent representation: a linear projection of 13 adjacent cepstral frames [1] into a 39-dimensional subspace (assuming each cepstral vector has 13 coefficients). However, as parameters of this linear projection are fixed in an ad hoc fashion, there is no guarantee that this particular subspace is optimal. As shown in Table 1, substantial improvements can be achieved by choosing the linear projection in a data-driven fashion. In this case, LDA is used to choose a subspace that better differentiates between multiple phonetic classes.



(a) traditional MFCC front-end

(b) an equivalent representation of (a)

(c) data-driven $\Delta, \Delta\Delta$

Fig. 4. Different Front-End Schemes

Δ,$\Delta\Delta$ Style	Embedding Dimension	WER (%)	
		w/o MLLT	w/ MLLT
traditional	13	21.6	21.2
data-driven	7	20.8	19.2
data-driven	13	20.1	19.0
data-driven	15	-	18.5

Table 1. Word error rates on 1998 Hub4e (Broadcast News) eval set 1. Embedding dimension is the number of frames used to form the super-vector, before LDA is applied.

For a fair comparison, LDA is used in the traditional front-end as well, in which case it does not reduce the dimensionality of the final feature space. Quinphone models with a comparable number of parameters are used for all different setups.

Clearly, the data-driven approach outperforms the ad hoc style (Eqn. 2) in choosing a subspace of the phase space. Using only 7 adjacent frames, the data-driven approach is already better than the traditional delta and double-delta front-end. It is also worth noting that a larger embedding dimension helps, but the gain quickly saturates. As a side note, there are established ways to choose a good embedding dimension as well as delay in nonlinear time series analysis. Due to the caveats discussed before, plus that we are mainly concerned with word error rates, it is easier to simply try out different dimensionalities. MLLT offers a further gain in all cases. Overall, a 14% WER reduction is obtained.

[1] Six to the left and six to the right, since $\vec{\Delta\Delta}_i$ makes use of $\vec{\Delta}_{i+3}$, which in turns uses $\vec{x}_{i+6}$. Ditto for $\vec{x}_{i-6}$.

6. CONCLUSIONS

This paper tries to establish a link between speech recognition and dynamic system research. Viewing speech production as a dynamic process, ASR can benefit from the concept of phase space and phase space reconstruction. The connection between the delta/double delta features and time-delayed embedding at the cesptral level is revealed. We also point out the importance of time-delayed embedding to the underlying HMM framework. While HMMs may seem to be unsophisticated, delayed embedding and other changes to the feature representation makes it much more powerful. We should bear this in mind when searching for advanced models beyond HMM. The effect of linear transform in the phase space is discussed. A properly chosen linear tranform gives a 14% gain on the Broadcast News task.

The analysis developed in this paper should be applicable to other tasks as well, such as handwriting recognition. It remains to be seen how far ASR can benefit from the theory and techniques in dynamic system research.

7. ACKNOWLEDGEMENTS

Many thanks to Xiaojin Zhu and other colleagues at CMU for helpful discussions, to J.A. Scott Kelso, Eric Vatikiotis-Bateson, and the American Institute of Physics for the permission to use their material.

8. REFERENCES

[1] Philip Rubin and Eric Vatikiotis-Bateson, “Measuring and modeling speech production,” in *Animal Acoustic Communication*, S.L. Hopp, M.J. Owren, and C.S. Evans, Eds. Springer-Verlag, 1998.

[2] J.A.S. Kelso, E. Vatikiotis-Bateson, EL Saltzman, and B. Kay, “A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modeling,” *Journal Acoustical Society of America*, , no. 77, pp. 266–280, 1985.

[3] V. Pitsikalis and P. Maragos, “Speech analysis and feature extraction using chaotic models,” in *Proc. ICASSP*, 2002.

[4] Andrew Lindgren, Michael Johnson, and Richard Povinelli, “Speech recognition using reconstructed phase space features,” in *Proc. ICASSP*, 2003.

[5] F. Takens, *Detecting Strange Attractors in Turbulence*, vol. 898 of *Lecture Notes in Math*, Springer, New York, 1981.

[6] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, vol. 7 of *Cambridge Nonlinear Science Series*, Cambridge University Press, 1997.

[7] L. Rabiner and B-H Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[8] S. Furui, “Speaker-independent isolated word recognition using dynamic features of speech spectrum,” *IEEE Trans ASSP*, vol. 34, no. 1, pp. 52–59, 1986.

[9] C. Diks, *Nonlinear Time Series Analysis: Methods and Applications*, vol. 4 of *Nonlinear Time Series and Chaos*, World Scientific Publishing, 1999.

[10] M. Ostendorf, V. Digilakis, and O. Kimball, “From hmms to segment models: A unified view of stochastic modeling for speech recognition,” *IEEE Transactions on Speech and Audio Processing*, 1996.

[11] R. Iyer, H. Gish, M. Siu, G. Zavaliagkos, and S. Matsoukas, “Hidden markov models for trajectory modeling,” in *Proc. ICSLP*, 1998.

[12] H. Yu and T. Schultz, “Implicit trajectory modeling through gaussian transition models for speech recognition,” in *Proceedings of the Human Language Technology Conference (HLT)*, 2003.

[13] H. Yu and A. Waibel, “Streamlining the front end of a speech recognizer,” in *Proc. ICSLP*, 2000.

[14] M. J. F. Gales, “Semi-tied covariance matrices,” in *Proc. ICASSP*, 1998.

[15] P. Olsen and R. Gopinath, “Modeling inverse covariance matrices by basis expansion,” in *Proc. ICASSP*, 2002.

[16] S. Axelrod, R. Gopinath, and P. Olsen, “Modeling with a subspace constraint on inverse covariances,” in *Proc. ICSLP*, 2002.