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. 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 for phase space reconstruction. The traditional delta and double features also impose a suboptimal linear transform in the reconstructed space. We show that a significant improvement in recognition accuracy can be achieved by choosing the transform in a data-driven fashion.

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 a phase space or a 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.

Figure 1: Phase portraits of two articulators during production of reiterant /ba/. (Courtesy of J.A.S. Kelso, E. Vatikiotis-Bateson, and the American Institute of Physics. ©1985 Acoustical Society of America)

If measurements of various articulators could be made easily and accurately, it would 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. This is discussed in Section 2.

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 in Section 3 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.

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. The effect of linear transformation in the reconstructed space is discussed in Section 4.

The use of dynamic features would not be necessary, if the underlying model can capture higher order dependencies. Appendix A discusses why delayed embedding is essential for
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 $m = 2, \tau = 3$ (other values may work just as well). Clearly, the reconstructed phase portrait has the same structure as the original system, although a strong correlation exists between the delayed coordinates. A linear transform can be used to decorrelate the two dimensions, which is the subject of Section 4.

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.

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 (non-linear) dynamics of the articulatory system, whereas the dynamics reconstructed from the time domain contains the compounding effects of two systems.

Figure 3: Two Sub-systems in Speech production
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:

\[
\begin{align*}
\vec{x}_i & = -3\vec{x}_{i-3} - 2\vec{x}_{i-2} - \vec{x}_{i-1} + \vec{x}_{i+1} + 2\vec{x}_{i+2} + 3\vec{x}_{i+3} \\
\Delta\vec{x}_i & = -3\Delta\vec{x}_{i-3} - 2\Delta\vec{x}_{i-2} - \Delta\vec{x}_{i-1} + \Delta\vec{x}_{i+1} + 2\Delta\vec{x}_{i+2} + 3\Delta\vec{x}_{i+3}
\end{align*}
\]

(1)

where \(\vec{x}_i\) is a 13-dimensional cepstral vector. These features lead to significant improvements and have since been widely used in ASR. Equation 1 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:

\[
\begin{pmatrix}
\vec{x}_1 \\
\Delta\vec{x}_1 \\
\Delta\Delta\vec{x}_1
\end{pmatrix} =
\begin{pmatrix}
-3 & -2 & -1 & 1 & 0 & 1 & 2 & 3 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vec{x}_1 & \cdots & \vec{x}_1
\end{pmatrix}
\]

(2)

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.  

It is well known that dynamic features lead to improved recognition performance. The particular choice of the linear transform in Equation 2 turns out to be quite important as well, as we will discuss next.

4. Linear Transforms in 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.

4.1. Optimizing Linear Transforms in Front-Ends

For speech recognition, however, the situation becomes a little more complicated. The front-end in a modern ASR system can be regarded as a pipeline consisting of many components. Some of them are linear and some are nonlinear, each serving a different purpose. A discussion on streamlining various linear transforms can be found in [12]. Below, we will focus on a typical MFCC (Mel-Frequency Cepstral Coefficients) front-end as an example. Our discussion should extend to other popular front-ends as well. In short, there are two key issues in the design of a front-end.

---

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

2One caveat is that 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. These issues are discussed in [6, 9].

---

Figure 4: Data-driven derivation of dynamic features. Feature dimensionalities are shown for all stages.

First, dimensionality reduction happens at many places in the MFCC front-end. While it makes training more feasible, dimensionality reduction loses information, and therefore, should be performed with great care. Some reductions are easy to recognize, such as spectral smoothing using the Mel-scale filterbank, truncating of the cepstrum, and LDA. Some are less obvious, such as the extraction of the dynamic features. As discussed before, the traditional dynamic features (Eqn. 1) are simply a linear projection of the delayed vector (Eqn. 2). Hence, it can be separated into two steps, delayed embedding (stacking of cepstral frames) and a linear projection, as shown in Figure 4(b).

When the input is 14-dimensional (cepstrum+energy), embedding increases the dimensionality to 14 * 13 = 182, which is then reduced to 42. The problem is that dimensionality reduction is performed by the ad hoc linear projection (as in Eqn. 2). There is no guarantee that the resulted subspace is optimal. Although LDA is also used, it does not reduce the dimensionality of the final feature space. A better approach is to skip this ad hoc transform and let LDA choose the best projection, as shown in Figure 4(c). \(\Delta, \Delta\Delta\) is never explicitly computed, but we can think of the entire LDA matrix as a data-driven derivation of that.

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) [15]. It can be regarded as a principled way to decorrelate feature dimensions.

4.2. Experiments

We experimented with different ways to derive dynamic features and the use of MLLT. The test set is the 1998 Hub4e (Broadcast News) evaluation set 1. All models have 105k Gaussians, and are trained on the 1996 Broadcast News training corpus (80 hours). Vocal tract length normalization, cluster-based cepstral mean normalization are used. All results are based on first-pass decoding using a 40k lexicon and a trigram language model.

As shown in Table 1, the data-driven approach outperforms...
Table 1: Word error rates on 1998 Hub-4e (Broadcast News) eval set 1. Embedding dimension is the number of frames used to form the super-vector, before LDA is applied.

<table>
<thead>
<tr>
<th>Style</th>
<th>Embedding Dimension</th>
<th>WER (%) w/o MLLT</th>
<th>WER (%) w/ MLLT</th>
</tr>
</thead>
<tbody>
<tr>
<td>traditional</td>
<td>13</td>
<td>21.6</td>
<td>21.2</td>
</tr>
<tr>
<td>data-driven</td>
<td>7</td>
<td>20.8</td>
<td>19.2</td>
</tr>
<tr>
<td>data-driven</td>
<td>13</td>
<td>20.1</td>
<td>19.0</td>
</tr>
<tr>
<td>data-driven</td>
<td>15</td>
<td>-</td>
<td>18.5</td>
</tr>
</tbody>
</table>

5. 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 cepstral level is revealed. We also point out the importance of time-delayed embedding to the underlying HMM framework. The effect of linear transform in the phase space is discussed.

We have obtained a 14% gain by switching from the traditional way of choosing a subspace of the phase space. Using only 7 adjacent frames, the data-driven approach is already better than the traditional MFCC front-end, which uses 13 adjacent frames. It is also worth noting that a larger embedding dimension helps, but the gain quickly saturates. MLLT offers a further gain in all cases. Overall, a 14% WER reduction (21.6% → 18.5%) is obtained.

6. References


A. Time-Delayed Embedding and HMMs

One may argue that after all, delayed embedding causes just 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 modeling framework.

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, parallel path HMMs [10, 11]. Unfortunately, these sophisticated models have yet to show improvements over the seemingly simple HMMs.

Part of the reason is due to the use of dynamic features, i.e. time-delayed embedding. 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.

Due to space limitations, we summarize the effect of time-delayed embedding on the underlying model as follows:

- Any higher order deterministic system can be described by a set of first order differential equations, through the use of delayed embedding;
- An m-th order Markov systems can be represented by a first order Markov model, through delayed embedding of m samples;
- For HMMs, we can no longer establish that a first order HMM can model a higher order source. Nevertheless, embedding increases the discriminative ability of (first order) HMMs by increasing the mutual information between feature vectors and their class labels.

Hence, while HMMs may seem to be unsophisticated, delayed embedding and other changes to the feature representation make it much more powerful. We should bear this in mind when searching for advanced models beyond HMMs.