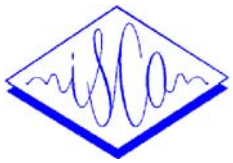
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Towards Better Understanding of the Model Implied by the use of Dynamic Features in HMMs

John S Bridle

Novauris Laboratories UK Ltd
Cheltenham, UK
john.bridle@novauris.com

Abstract

We examine a widely-used kind of Hidden Markov Model (HMM), in which "dynamic features" are included along with the direct measurements. We conclude that the generative model implied by the use of dynamic features is quite different from the conventional view and that such models are capable of providing a surprising amount of the sort of dynamics that we thought were necessary to describe the important properties of speech patterns. We suggest that one reason why several attempts to replace HMMs with models with explicit dynamics have failed is that the dynamics already implicit in standard HMMs are roughly equivalent.

10.21437/Interspeech.2004-281

1. Introduction

For many years a widely-used type of acoustic model for automatic speech recognition has been a Hidden Markov Model with an output distribution for each state defined in terms of a mixture of gaussians. The "observation vector" for the output distributions is usually made by appending to the real, direct (static) acoustic vector one or more vectors computed from the local sequence of static vectors. The use of such augmented feature vectors can significantly enhance the performance of automatic speech recognition systems [2].

In the standard view, HMMs are generative models: given a trained HMM, it is possible to draw samples from the output distribution by first deciding what state the system will be at each time, then at each time, sampling from the output distribution of the corresponding state. The distribution at each time depends only on the (hidden) state at that time, so the output distribution has piecewise-constant statistics. Although the generative model aspect of HMMs is not much used in practice, it is very important conceptually. We argue here that the standard interpretation of the generative model is quite wrong when dynamic features are included.

Several authors have attempted to improve on HMMs by including explicit dynamics and segmental properties. (*e.g.* [6, 3]) The usual argument is that conventional HMMs are inadequate for dealing with the real nature of speech patterns, so we should be able to find something even better. Generally speaking, results from these attempts have been disappointing, and the basic HMM structure has not been superseded.

It is well known that HMMs make very speech synthesizers, because of the piece-wise constant statistics. However, Tokuda *et al.* have shown that much more interesting and dynamically-rich acoustic patterns can be produced from HMMs if they are looked at in a different way, and suprisingly good speech can be synthesized from them [7]. The key insight is that the output statistics are piecewise-constant for the *augmented* vectors, and the relationship between the static and dynamic features must be taken into account when contructing the maximum-likelihood sequence of static vectors [8].

2. The Conventional HMM Output Distribution

In this section we define what we are calling the conventional model.

We are mainly concerned, in the present paper, with what happens within an interval during which the HMM state does not change, so generally we do not mark the dependence on state in the notation that follows. We also ignore end-effects, although these can be very important, especially as the number of frames controlled by a state is often less than the width of the window over which the dynamic features are computed.

Given a sequence, $[y_1, y_2, \cdots y_n]$, of observation vectors, a sequence of augmented vectors, $[z_1, z_2, \cdots z_n]$, is constructed by appending to each observation vector some more elements that are computed from adjacent observations. In general,

$$z_i = F(y_{i-w}, \ldots, y_{i+w}), \quad (1)$$

where F is linear and w is the half-width of the window. We intend that z_i includes y_i. As a specific example, the default in HTK [11] is to append a vector of *delta coefficients*

$$d_i = (-2y_{i-2} - y_{i-1} + y_{i+1} + 2y_{i+2})/10 \quad (2)$$

and a vector of *acceleration coefficients*

$$a_i = (-2d_{i-2} - d_{i-1} + d_{i+1} + 2d_{i+2})/10 \quad (3)$$

In this case the augmented vectors z_i can be expressed in terms of y_{i-4} to y_{i+4} ($w = 4$).

The distribution over the augmented vectors is written as a mixture of diagonal-covariance gaussians (and the parameters of the distribution at each frame are controlled by the corresponding HMM state).

For any sequence of augmented vectors, $Z = [z_1, ...z_n]$, we compute a score that would be the probability density if the augmented vectors could be considered independent (given the state at each time):

$$P(Z) = \prod_{i=1}^{n} p(z_i) \quad (4)$$

Our central question is: What is the generative model implied by this construction?

The conventional answer is that n augmented vectors are drawn independently from p, and experiments have been published using synthetic data generated this way (*e.g.* [5]). This will obviously not do, because the constraints of equations 1 will not be satisfied (unless we are very lucky!).

At this point it would be possible to say "Well, the model may be inconsistent, but it works, and works rather well!" A different view is "HMMs are a mess, so we should build something clean and powerful to replace them". In this paper we take a third view, that equations 1 and 4 *do* define a distribution, and a generative model, but it is not the obvious one.

We want a generative model for the real observations, $[y_1, \cdots y_n]$. The dynamic features are simply intermediates in the definition of the model. The key notion is that however we do it, the dynamic features must conform to the algebraic constraints of equation 1, as exemplified by relationships 2 and 3.

One approach (conceptually) is to generate a proposed sequence Z, and check that the dynamic features obey the constraints. If not, try again. This will take too long, but the argument reminds us that the density in 4 is not correctly normalized, since most of the Z space is invalid because of 1.

Equations 1 and 4 certainly define a function from any sequence of observations to a scalar that looks like an unnormalized probability. We have only to normalize it to define a density. In principle, it is clear: the product of mixtures of gaussians is a (possibly rather large) mixture of gaussians. (The product of M mixtures, each with C gaussian components, has C^M components in general.) The big MoGs in the augmented vectors can be re-written in terms of the original observation vectors sequence, again as a mixture of gaussians (if we neglect end effects). See Tokuda *et al.* [8] for details.

Williams shows [9] that the use of augmented vectors can be analysed in terms of an autocovariance model, or as a product of gaussians.

One way to understand a generator is to look at the mode of the output distribution. This is the sequence $[y_1, \cdots y_n]$ that maximises equation 4 while constrained by 1. This maximum likelihood pattern is used for synthesis by Tokuda *et al.* [7].

3. Sampling from the Generative Model

One of the best ways to understand a stochastic model is to construct the generator, and to draw samples from it. Indeed, such a generator is a standard tool when testing implementations of inference algorithms for stochastic models.

By "a sample" we mean a complete sequence of "static" vectors, $[y_1, y_2, \cdots y_N]$, corresponding to measurements of spectrum shape etc. (We shall assume that the sequence of HMM states has been decided beforehand)

The Gibbs Sampler technique [1, 4], also known as Markov Chain Monte Carlo (MCMC), is a convenient way to sample from the joint distribution defined by equations such as 1 and 4. (Note that the markov chain in MCMC is quite distinct from the markov chain of hidden states in an HMM.)

There are several ways to make a Gibbs Sampler for our problem. One simple version is as follows: We choose a frame, i, compute the conditional distribution of y_i given the current values at the other times, and sample from that conditional distribution. It is well known that the (markov chain) process defined by repeating such procedures converges to the joint distribution. If the distributions over augmented vectors are mixtures of gaussians, then the conditional distributions are also mixtures of gaussians, and the means and mixture weights are functions of the adjacent values.

4. A Simple Model with Dynamics: Delta Chains

We shall focus on a simple case, to illustrate some of the ideas.

We use one-dimensional observations (y is a scalar), and simple asymmetric delta: $d_i = y_i - y_{i-1}$. The equivalent of the dynamic features is a probabilistic constraint on the difference between adjacent observations:

$$P(y_1, ...y_n) = P_s(y_1)P_d(y_2 - y_1)P_s(y_2)P_d(y_3 - y_2)...P_s(y_n) \quad (5)$$

where P_s is a distribution for the static, original observation values, and P_d is the distribution of the differences or deltas. P_s and P_d are each mixtures of gaussians. Equation 5 describes a chain with no delta constraints at the two ends. We also sometimes connect the ends in a ring.

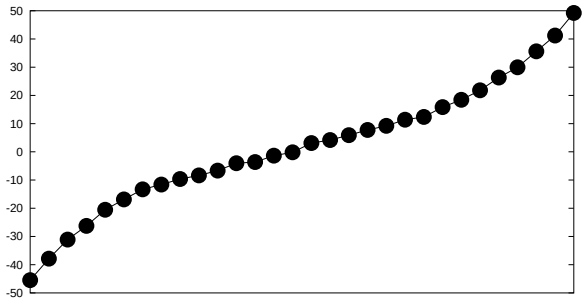


Figure 1: *Sample from a simple Gaussian DeltaChain with a positive delta mean.*

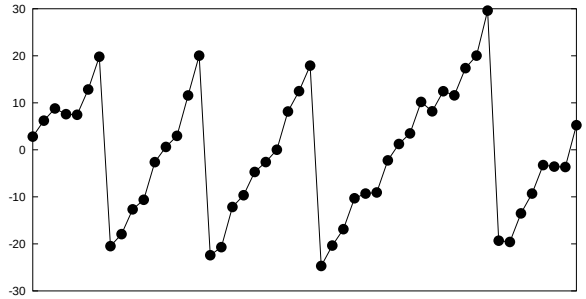


Figure 2: *Sample from a DeltaChain with a bimodal delta.*

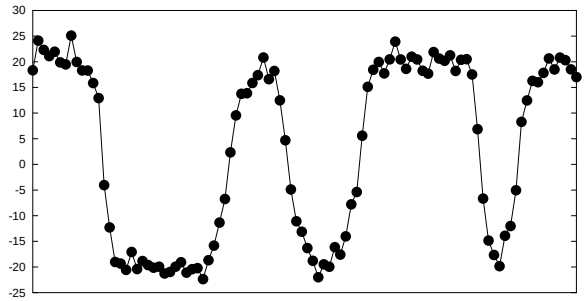


Figure 3: *Sample from a DeltaChain with bimodal static.*

Our simple *delta chain* model is different from the conventional model in the following ways:

1. The observations are one dimensional, so we can plot the sequence as a simple graph. The diagonal covariance matrix of the conventional model means that to a first approximation we can treat the different dimensions (features) separately, but in most systems they are coupled through the choice of mixture component.

2. The simple delta is offset (between frames, rather than being centered on the frames), and the distributions of the deltas are independent of the statics, whereas in the conventional model the dynamic features are centered, and the distribution is joint, so the choice of mixture component in made for the statics and dynamics together. (In turns out that our simpler, independent, model is a special case of the joint model, so anything that the simple model can do can be done with the more complex one.)

4.1. Some samples from scalar delta chains

In this section we look at single samples from a few simple delta chains, each controlled by a single output distribution defined on static and dynamic features, so it is relevant to what could be generated by an HMM for a sequence of frames during which the state does not change. (It is more enlightening to watch as the sampling proceeds, and we plan to make some dynamic examples generally available.)

Figure 1 is a (Gibbs Sampler) sample from a delta chain of 30 frames. The static distribution P_s is gaussian with zero mean ($P_s = \mathcal{N}(0, 30)$) and the delta distribution is gaussian with positive mean ($P_d = \mathcal{N}(10, 1)$). The ends of the chain are unconstrained in figure 1, but loop-connected in figures 2, 3 and 4. (It is not strictly necessary to use a Gibbs Sampler in this simple case. In principle any system with single-gaussian output distributions can be inverted, as explained by Tokuda [8].) As expected, there is a slope approaching 10 vertical units per frame at the edges. The static variance constrains the slope in the middle of the "segment".

Things get more interesting if we use mixtures of gaussians for the static or delta distributions, or for both. In figure 2 we have a zero mean static distribution with a large variance: $P_s = \mathcal{N}(0, 100)$. The delta is a mixture of a positive slope and a less likely negative jump: $P_d = 0.9\mathcal{N}(5, 10) + 0.1\mathcal{N}(-50, 10)$. The result is a sawtooth.

Figures 3 and 4 show two ways to make the system alternate on a relatively long cycle between two levels. The recipe for figure 3 is:
$P_s = 0.5\mathcal{N}(20, 20) + 0.5\mathcal{N}(-20, 20)$
$P_d = \mathcal{N}(0, 10)$.
The delta introduces a persistence in the choice of the static component, and smooths the transitions.

The recipe for figure 4 is:
$P_s = 0.5\mathcal{N}(20, 20) + 0.5\mathcal{N}(-20, 20)$
$P_d = 0.6\mathcal{N}(0, 1) + 0.2\mathcal{N}(40, 10) + 0.2\mathcal{N}(-40, 10)$.
The extra delta components provide the opportunity for clean jumps between the two modes of the static distribution.

5. Discussion

Segmental HMMs are one class of models that are an attempt to go beyond the assumed limitations of standard HMMs [3]. In the simplest form, the sequence of vectors in a segment is determined by drawing a sample from a first distribution to set the mean for this segment, then drawing samples from a distribution centered on this

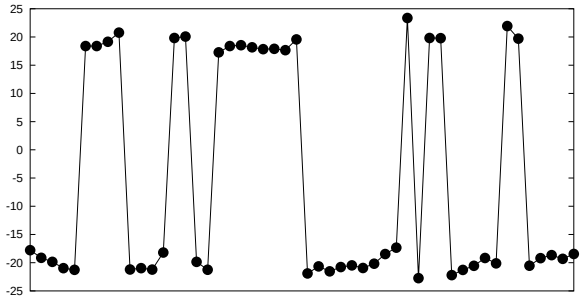


Figure 4: *Sample from a DeltaChain with bimodal static and trimodal delta.*

mean, but with a new (tighter) variance. The result is to allow much larger differences between frames in different segments of the same type than between frames within the same segment. A very similar effect can be achieved with conventional HMMs. We can illustrate using a simple delta chain, by using the static distribution for the initial distribution, and using a zero-mean delta distribution to achieve the reduced variance within the segment. There certainly are differences — for instance the delta distributions will impose their own notion of continuity at "segment" boundaries — but it is not clear that the overall effect will be worse in terms of modeling speech patterns. Models with trends within a segment can also be dealt with using a non-zero mean for the delta distribution.

It will be interesting to see the spectrum patterns produced by running the Gibbs sampler on distributions (and constraints) derived from real acoustic models. The maximum-likelihood trajectories produced by Tokuda *et al.* [7], which correspond to the mode of the full distribution, are a foretaste.

It is an open question whether this view of the generative model implied by current HMMs will lead to improvements in the way that the parameters are estimated. Tokuda *et al.* [8] show that they can improve the accuracy of a system that uses a special "trajectory model" scoring method, but do not report improvements when the "trajectory trained" models are used with conventional decoders.

6. Conclusions

We hope that the viewpoint expounded here (and in some of the references) will help to put the alternatives to the conventional HMM in context, and lead eventually to improvements in automatic speech recognition.

7. Acknowledgements

The author would like to thank Chris Williams, Joe Frankel, Mark Bedworth Hywel Richards and Melvyn Hunt for stimulating discussions.

8. References

[1] S.Geman, and D.Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images" IEEE Trans. PAMI, 6: 721–742, 1984.

[2] S.Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", IEEE TRans. ASSP-34 52–59, 1986.

[3] W.Holmes and M.Russell, "Experimental evaluation of segmental HMMs", Proc ICASSP 1995, 536–539.

[4] A.Ihler, E.Sudderth, W.Freeman and A.Willsky, "Efficient Multiscale Sampling from Products of Gaussian Mixtures", Advances in Neural Information Processing Systems 16, MIT Press, 2004.

[5] D.McAllaster, L.Gillick, F.Scattone, and M.Newman, "Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch", Proc. ICSLP98 1847–1850.

[6] J.Picone, S.Pike, R.Regan, T.Kamm, J.Bridle, L.Deng, Z.Ma, R.Richards and M.Schuster, "Initial evaluation of Hidden Dynamic Models on conversational Speech", Proc. ICASSP99.

[7] K.Tokuda, T.Masuko, T.Yamada, T.Kobayashi, and S.Imai, "An Algorithm for speech parameter generation from continuous mixture HMMs with dynamic features", Proc. EuroSpeech 1995, 757–760.

[8] K.Tokuda, H.Zen, and T.Kitamura, "Trajectory modeling based on HMMs with the explicit relationship between static and dynamic Features", Proc EuroSpeech 2003.

[9] C.Williams, "How to Pretend that correlated variables are independent by using difference observations", Neural Computation (to appear)

[11] S.Young, *et al.* , The HTK Book, Entropic Cambridge Research Laboratories.