# Intermediate-State HMMs to Capture Continuously-Changing Signal Features

**Gustav Eje Henter**[1] **and W. Bastiaan Kleijn**[1,2]

[1] *Sound and Image Processing Lab, School of Electrical Engineering, KTH – Royal Institute of Technology, Stockholm, Sweden*

[2] *School of Engineering and Computer Science, Victoria University of Wellington, New Zealand*

## Paper Abstract

**Traditional discrete-state HMMs are not well suited for describing steadily evolving, path-following natural processes like motion capture data or speech. HMMs cannot represent incremental progress between behaviours, and sequences sampled from the models have unnatural segment durations, unsmooth transitions, and excessive rapid variation. We propose to address these problems by permitting the state variable to occupy positions between the discrete states, and present a concrete left-right model incorporating this idea. We call this intermediate-state HMMs. The state evolution remains Markovian. We describe training using the generalized EM-algorithm and present associated update formulas. An experiment shows that the intermediate-state model is capable of gradual transitions, with more natural durations and less noise in sampled sequences compared to a conventional HMM.**

## The Problem

Hidden Markov models (HMMs) represent the dominant paradigm in model-based speech synthesis and recognition. They allow for non-linearity and long memory while still scaling linearly with database size. However, standard HMMs make a number of assumptions that are inappropriate for steadily-evolving processes like speech:

1. Durations: HMMs change states memorylessly, and durations are thus geometrically distributed. This gives rise to much greater variations in duration than real speech sounds have.
2. Transitions: Speech features tend to evolve gradually, not instantaneously and in discrete steps as HMMs assume.
3. Frame independence: Deviations from the mean behaviour tend to be correlated between frames (e.g., overshoots and undershoots), but HMMs model deviations as independent Gaussian noise.

Because of these shortcomings, random samples from speech-trained HMMs do not sound much like speech. The output is noisy, steppy, and has unnatural durations.

## Established Techniques

A number of techniques have been proposed to address the above issues. Particularly prominent are:

1. Hidden semi-Markov models (HSMMs): These let the transition probability depend on the time spent in the current state. This allows arbitrary state durations, but is difficult to make efficient.
2. Dynamics features: One can incorporate delta and delta-delta features into the HMM. This is challenging to do in a mathematically meaningful manner, and requires approximations to scale well.
3. Maximum-likelihood parameter generation (MLPG): Since random samples are noisy, don't sample—only generate the most probable outcome. This does not address the root problem, but merely hides its symptoms.

## Intermediate States

We propose a single, efficient, and mathematically consistent solution that simultaneously addresses problems 1 and 2, and somewhat mitigates problem 3.

The idea is to let the HMM state variable attain non-integer positions. This tracks incremental progress from one state to the next, so that reasonable durations naturally emerge. Positions in-between regular states also allow for intermediate sounds and gradual transitions between behaviours.

Parameters remain tied to integer positions, so the degrees of freedom do not increase notably. Integer-position parameters now act as templates, where state evolution and output properties at non-integer locations interpolate between adjacent templates'.

## Remarks

Like for HSMMs, we introduce additional memory to the underlying Markovian process, but this extra memory is a fractional state-part rather than a duration. This brings two advantages:

- Durations create a 2-D state space that increases with $T$; we maintain a 1-D state space.
- HSMM time-spent-in-state does not lend itself well to interpolation.

The underlying Markov chain $S_t$ of a left-right HMM can be seen as a kind of random walk on $\mathcal{S} \subset \mathbb{Z}$ with location-dependent evolution. Our intermediate-state variable $I_t$ performs a similar non-decreasing walk on the expanded (non-integer) space $\mathcal{I} \subset \mathbb{R}$, which can be discrete or continuous. With a discrete state space we effectively have a kind of tied HMM.

## Graphical Comparison

| | HMMs | HSMMs | Dynamics features | MLPG | MLPG + dynamics | Intermediate states |
|---|---|---|---|---|---|---|
| Natural durations | ✗ | ✓ | ✗ | ✗! | ∼ | ✓ |
| Continuous paths | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ |
| Reduced noise | ✗ | ✗ | ✗ | ✓ | ✓ | ∼ |
| Linear complexity | ✓ | ∼ | ∼ | ✓ | ∼ | ✓ |

## Training and Usage

To be practically relevant, new models must be possible to train and use.

- ⋆ Sampling from intermediate-state models is sequential and straightforward, making synthesis easy.
- ⋆ Data sequence probabilities and Viterbi paths can be computed on discretised intermediate-state HMMs with standard algorithms.
- ⋆ The same goes for the E-step of EM-training.

The M-step during training is more involved. Typically, the parameters can not all be updated simultaneously, and each update involves solving an $N \times N$ linear system. Specifically:

- Means can be optimized analytically under fixed standard deviations.
- Standard deviations can be updated using Newton's method when means are fixed, falling back on gradient updates when Newton's method does not increase the objective function.
- State evolution parameter updates also require Newton's method.

Generalized EM parameter update formulas are presented in the paper.

## An Example Application

We trained a 27-state conventional HMM and a 26-template intermediate-state HMM on eight examples of the utterance "titta bilen" (Swedish for "look, the car"). The features were the log pitch track and MFCCs of filter and aperiodicity spectra produced by the STRAIGHT system at 200 fps.

The intermediate-state model used 12 equidistantly spaced fractional states per integer, along with random walk step lengths $\Delta I_t \in \{1/12, 2/12, 4/12\}$. Each template was associated with a distribution over the different step lengths. The state-conditional output distribution $X_t \mid I_t$ was Gaussian, with a mean vector and (diagonal) standard deviation matrix linearly interpolated between those of adjacent templates.

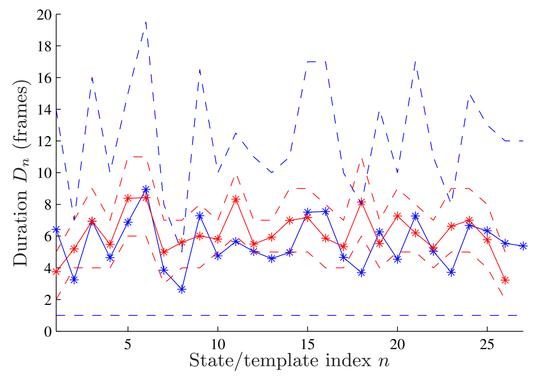


**Figure 1:** *State/template durations. The mean is solid, with 0.1 and 0.9 quantiles dashed.*

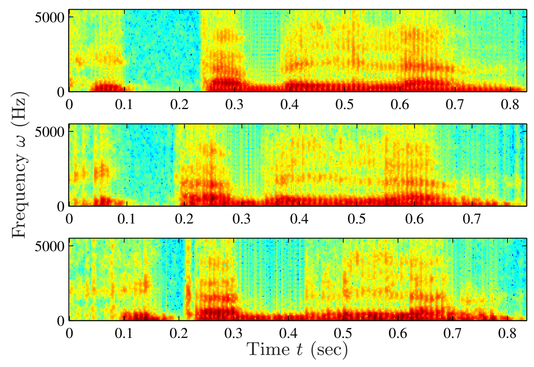


**Figure 2:** *Sample spectrograms of reference (top), intermediate-state model (middle), and discrete HMM (bottom).*

A seen in figure 1, random samples from the trained models had a similar mean duration profile, but the traditional HMM showed unnaturally large duration variation, unlike the intermediate-states model.

The better duration modelling and interpolating properties of intermediate-state models are clearly visible in figure 2, which compares a random sample from each model against a training utterance. Intermediate-states also yielded marginally smaller residual noise variance.

## Future Work

Intermediate-state HMMs do not model temporal correlations between deviations. We are working on autoregressive HMMs which account for correlations and still scale linearly in $T$. This would provide an efficient response to all three HMM shortcomings simultaneously.

We are additionally interested in different hidden-state evolution choices, such as fully continuous state spaces. This would require approximate state estimation, e.g., with the unscented or the extended Kalman filter. We are also considering improved parameter update formulas for training.