The Effect of Using Normalized Models in Statistical Speech Synthesis

Matt Shannon[1]*, Heiga Zen*[2]*, William Byrne*[1]

[1]Cambridge University Engineering Department, Cambridge, U.K.
[2]Toshiba Research Europe Ltd., Cambridge Research Lab., Cambridge, U.K.
sms46@eng.cam.ac.uk, heiga.zen@crl.toshiba.co.uk, bill.byrne@eng.cam.ac.uk

Abstract

The standard approach to HMM-based speech synthesis is inconsistent in the enforcement of the deterministic constraints between static and dynamic features. The *trajectory HMM* and *autoregressive HMM* have been proposed as normalized models which rectify this inconsistency. This paper investigates the practical effects of using these normalized models, and examines the strengths and weaknesses of the different models as probabilistic models of speech. The most striking difference observed is that the standard approach greatly underestimates predictive variance. We argue that the normalized models have better predictive distributions than the standard approach, but that all the models we consider are still far from satisfactory probabilistic models of speech. We also present evidence that better intra-frame correlation modelling goes some way towards improving existing normalized models.

Index terms: HMM-based speech synthesis, acoustic modelling, autoregressive HMM, trajectory HMM, normalization

1. Introduction

The standard approach to HMM-based speech synthesis [1] is inconsistent in the enforcement of the deterministic constraints between static and dynamic features [2]. During synthesis we explicitly impose these constraints [3] whereas the standard model used during parameter estimation ignores them. Alternatively, the standard model used during parameter estimation can be viewed as a model defined over static features only, in which case it correctly enforces the constraints between static and dynamic features but is *unnormalized* as a probability distribution, i.e. the probability of the set of all sequences of static features is not one [2].

Previously models such as the *trajectory HMM* [2] and the *autoregressive HMM* [4] have been proposed to address this problem by using the same valid, normalized model for parameter estimation and synthesis.

The lack of normalization in the standard model used during parameter estimation means that the probabilistic justification for conventional training procedures in terms of maximizing the likelihood strictly speaking does not apply. However the question remains what the practical consequences of this lack of normalization are.

To investigate the effect of normalization we look at the predictive distribution for the standard approach, the trajectory HMM and the autoregressive HMM. We focus on the predictive distribution since this is the quantity of interest when viewing the three approaches as probabilistic models of speech, and because it is the basis for almost all current synthesis algorithms.

We also investigate a possible improvement to current normalized models, by using full covariance matrices to model the correlations between different components of the feature vector.

The rest of this paper is organized as follows. Section 2 reviews the acoustic models used in the three approaches to statistical speech synthesis we will be comparing. Section 3 investigates the effect of normalization on the predictive distribution. Section 4 considers improving current normalized models by using full covariance modelling for the trajectory HMM.

2. Acoustic modelling in speech synthesis

2.1. Framework

In statistical parametric speech synthesis [5] we typically build a probabilistic model $P(C \,|\, \boldsymbol{l}, \lambda)$, where $C = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_T]$ is a representation of speech as a sequence of acoustic feature vectors, T is the number of frames in C, $\boldsymbol{l} = [l_1, \ldots, l_J]$ is a representation of text as a sequence of labels, J is the number of labels in $\boldsymbol{l}$, and λ is a set of model parameters [5]. As an aid to modelling we introduce a hidden *state sequence* $\boldsymbol{q} = [q_1, \ldots, q_T]$, where each q_t is a label together with an integer state index. We then decompose the model $P(C \,|\, \boldsymbol{l}, \lambda)$ into a *duration model* $P(\boldsymbol{q} \,|\, \boldsymbol{l}, \lambda)$ and an *acoustic model* $P(C \,|\, \boldsymbol{q}, \lambda)$. We use the same semi-Markov form of duration model throughout [6].

The sequence over time of a single component of the feature vector (e.g. the 6$^{\text{th}}$ mel-cepstral coefficient) forms a *trajectory*. We will mostly follow the common assumption that the trajectories $\boldsymbol{c}_{(1:T)i}$ for different feature vector components i are independent given the state sequence. From now on we will focus on one component of the feature vector sequence, and for clarity of notation drop the index i. Thus $c_t \in \mathbb{R}$ is a scalar, $\boldsymbol{c} \in \mathbb{R}^T$ is a trajectory, and $P(\boldsymbol{c} \,|\, \boldsymbol{q}, \lambda)$ is a distribution over trajectories.

2.2. Model used in standard approach during training

The standard model used during parameter estimation is as follows. We augment the static feature vector $o_{t0} \triangleq c_t$ with dynamic features $o_{t1} \triangleq \frac{1}{2}c_{t+1} - \frac{1}{2}c_{t-1}$ and $o_{t2} \triangleq c_{t+1} - 2c_t + c_{t-1}$ (for the standard HTS windows) to obtain an *observation* $\boldsymbol{o}_t = [o_{t0}, o_{t1}, o_{t2}]$. Note that the map $w : \mathbb{R}^T \to \mathbb{R}^{T \times 3}$ taking a static feature vector sequence $\boldsymbol{c}$ to its corresponding *observation sequence* $O = [\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T]$ is linear. We then build a model over O instead of over $\boldsymbol{c}$, setting $P_{\text{obs}}(O \,|\, \boldsymbol{q}, \lambda) = \prod_t P_{\text{obs}}(\boldsymbol{o}_t \,|\, q_t, \lambda)$ where

$$P_{\text{obs}}(\boldsymbol{o}_t \,|\, q_t = q, \lambda) = \prod_{d=0}^{2} \mathcal{N}\left(o_{td}; \mu_{qd}, \sigma_{qd}^2\right) \qquad (1)$$

This work was partly supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (EMIME).

and $\lambda = \left[\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2 : q\right]$. Note that this ignores the deterministic relationship between $\boldsymbol{o}_{(1:T)0}$, $\boldsymbol{o}_{(1:T)1}$ and $\boldsymbol{o}_{(1:T)2}$. The model places almost all of its probability mass on observations sequences that can never occur, since the set of *realizable* observation sequences $\{O : O = w(\boldsymbol{c}) \text{ for some } \boldsymbol{c}\}$ forms a thin T-dimensional subspace of the overall $3T$-dimensional space.

For a realizable sequence $O = w(\boldsymbol{c})$ we have

$$P_{\text{obs}}(\boldsymbol{o}_t \mid q_t = q, \lambda) = \psi(\boldsymbol{c}_{t-L:t+L}, q_t, \lambda) \tag{2}$$

where $L = 1$ for the standard settings of these windows. Here $\log \psi$ is a quadratic form in $\boldsymbol{c}_{t-L:t+L}$, i.e. $\psi(\cdot, q_t, \lambda)$ can be thought of as an unnormalized Gaussian [2, 7]. This follows from the fact that w is linear and that $[w(\boldsymbol{c})]_t$ depends only on $\boldsymbol{c}_{t-L:t+L}$.

Equivalently, rather than viewing the standard model above as a model over O, we can view it as an *unnormalized* model over $\boldsymbol{c}$

$$\text{“}P_{\text{std}}\text{”}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) \triangleq P_{\text{obs}}(w(\boldsymbol{c}) \mid \boldsymbol{q}, \lambda) \tag{3}$$

$$= \prod_t \psi(\boldsymbol{c}_{t-L:t+L}, q_t, \lambda) \tag{4}$$

This is unnormalized since integrating over all possible $\boldsymbol{c}$ does not necessarily give 1.

Thus the standard model used during training can be viewed either as a model over O that places most of its probability mass on unrealizable sequences, or as an unnormalized model over $\boldsymbol{c}$.

2.3. Trajectory HMM

The *trajectory HMM* [2] explicitly normalizes “P_{std}”

$$P_{\text{traj}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) \triangleq \frac{1}{Z(\boldsymbol{q}, \lambda)} P_{\text{obs}}(w(\boldsymbol{c}) \mid \boldsymbol{q}, \lambda) \tag{5}$$

$$= \frac{1}{Z(\boldsymbol{q}, \lambda)} \prod_t \psi(\boldsymbol{c}_{t-L:t+L}, q_t, \lambda) \tag{6}$$

where $Z(\boldsymbol{q}, \lambda)$ is the normalization constant required to obtain a valid probability distribution. This normalization constant $Z(\boldsymbol{q}, \lambda)$ does not factorize over time with respect to $\boldsymbol{q}$, which means that training for the trajectory HMM is more computationally demanding than for the standard approach [2].

The trajectory HMM is *globally normalized* at the level of $\boldsymbol{c} \mid \boldsymbol{q}$ – we first take the product of the unnormalized individual factors for each time t, then normalize.

2.4. Autoregressive HMM

The *autoregressive HMM* [4] achieves normalization by building up the overall distribution $P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$ from locally-normalized pieces

$$P_{\text{ar}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) \triangleq \prod_t P_{\text{ar}}(c_t \mid q_t, \boldsymbol{c}_{t-K:t-1}, \lambda) \tag{7}$$

where

$$P_{\text{ar}}(c_t \mid q_t = q, \boldsymbol{c}_{t-K:t-1}, \lambda) \triangleq \mathcal{N}\left(c_t; \sum_{k=1}^{K} a_{qk} c_{t-k} + b_q, \sigma_q^2\right) \tag{8}$$

and $\lambda = \left[\boldsymbol{a}_q, b_q, \sigma_q^2 : q\right]$. The individual factors $P_{\text{ar}}(c_t \mid q_t, \boldsymbol{c}_{t-K:t-1}, \lambda)$ are *linear-Gaussian*. Typically $K = 3$.

The autoregressive HMM is *locally normalized* – the overall distribution $P_{\text{ar}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$ is the product of the individual factors for each time t, each of which is normalized. The fact that c_t only depends on the past $\boldsymbol{c}_{1:t-1}$ for each factor ensures that the overall distribution is normalized.

approach	model during training	model during synth
std	$P_{\text{obs}}(\boldsymbol{o} \mid \boldsymbol{q}, \lambda) \Leftrightarrow$ “P_{std}”$(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$	$P_{\text{traj}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$
traj	$P_{\text{traj}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$	$P_{\text{traj}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$
AR	$P_{\text{ar}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$	$P_{\text{ar}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$

Table 1: Summary of how the various acoustic models are used.

2.5. Model used in standard approach during synthesis

In the standard approach we do take the constraints between static and dynamic features into account during synthesis [3]. Rather than computing the most likely observation sequence $\arg\max_{\boldsymbol{o}} P_{\text{obs}}(\boldsymbol{o} \mid \boldsymbol{q}, \lambda)$ we compute the most likely *realizable* observation sequence $\arg\max_{\boldsymbol{c}} P_{\text{obs}}(w(\boldsymbol{c}) \mid \boldsymbol{q}, \lambda)$. Note that this is equal to $\arg\max_{\boldsymbol{c}} P_{\text{traj}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$ since the normalization constant does not depend on $\boldsymbol{c}$. Therefore we may say that the standard approach is effectively to use P_{traj} during synthesis. Since $P_{\text{traj}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$ is Gaussian, the most likely trajectory is the mean trajectory.

2.6. Summary

To summarize we have

$$\text{“}P_{\text{std}}\text{”}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) = \prod_t \psi(\boldsymbol{c}_{t-L:t+L}, q_t, \lambda) \tag{9}$$

$$P_{\text{traj}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) = \frac{1}{Z(\boldsymbol{q}, \lambda)} \prod_t \psi(\boldsymbol{c}_{t-L:t+L}, q_t, \lambda) \tag{10}$$

$$P_{\text{ar}}(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) = \prod_t P_{\text{ar}}(c_t \mid q_t, \boldsymbol{c}_{t-K:t-1}, \lambda) \tag{11}$$

A summary of how these acoustic models are used is given in Table 1. The difference between the standard approach and the trajectory HMM is the use of a normalized model during parameter estimation.

For all three of these models the distribution of $\boldsymbol{c} \mid \boldsymbol{q}$ is Gaussian, i.e. $P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda) = \mathcal{N}(\boldsymbol{c}; \overline{\boldsymbol{\mu}}_{\boldsymbol{q}}, \overline{\Sigma}_{\boldsymbol{q}})$ for some mean trajectory $\overline{\boldsymbol{\mu}}_{\boldsymbol{q}}$ and covariance matrix $\overline{\Sigma}_{\boldsymbol{q}}$ [2, 7].

3. Effect of normalization

In this section we investigate the effect of training with a normalized model on the predictive distribution $P(\boldsymbol{c} \mid \boldsymbol{q}, \lambda)$ used during synthesis. We first compare the three models qualitatively by visualizing the predictive distribution for some unseen test utterances. We hope this will give the reader some insight into the qualitative differences between the models. We then show that our observations based on these examples generalize by looking at an objective measure.

For these experiments we randomly selected 50 *test set* utterances from the CMU ARCTIC corpus, and trained standard, autoregressive HMM and trajectory HMM systems on the remainder of the corpus.[1] For simplicity of computation and visualization we use a fixed state sequence $\boldsymbol{q}$ throughout. In all cases alignments based on P_{obs} are used for the standard and trajectory HMM systems, and alignments based on P_{ar} are used for the autoregressive HMM system.[2]

[1] The standard and autoregressive HMM systems are as used in previous work [4]. The trajectory HMM system was trained using a Viterbi alignment obtained from the standard system.

[2] Specifically the alignments used are *median alignments* obtained by at each time picking the state with the median posterior occupancy.

3.1. Visualization of the predictive distribution

3.1.1. Mean trajectory with pointwise variance

For each of the systems we plot the mean trajectory ± 1.5 standard deviations along with the natural trajectory actually generated by the speaker. The standard deviation at time t is the square root of the *marginal* or *pointwise* variance $[\overline{\Sigma}_{\boldsymbol{q}}]_{tt}$ predicted by the model. This gives some insight into the distribution over trajectories encoded by the various models by showing the range of values each model expects to see at each time.

Figure 1 shows an example of a distribution over trajectories for the 6$^{\text{th}}$ mel-cepstral coefficient. We can see that the standard approach underestimates the variance – the natural trajectory is often outside the range predicted by the model, and we see a few events that are so many standard deviations from the mean that they should only happen *extremely* rarely according to the model. The normalized models have much larger variances, and this looks more reasonable – the natural trajectory is a long way outside the predicted range much less often.

The different models also have different mean trajectories, though the effect of normalization on the mean trajectory is smaller than the effect on the variance around the mean. In this example the trajectory HMM has the mean trajectory that lies closest to the natural trajectory.

3.1.2. Sampling trajectories from the predictive distribution

Another way to investigate the characteristics of the predictive distribution is to sample from it. Our implicit assumption during maximum likelihood parameter estimation is that the speaker generated the training corpus by *sampling* from $P(\boldsymbol{c} \,|\, \boldsymbol{q}, \lambda)$ for each utterance. Therefore a good way to assess the accuracy of our probabilistic model is to draw samples from our trained model $P(\boldsymbol{c} \,|\, \boldsymbol{q}, \lambda)$ and compare these to natural trajectories.

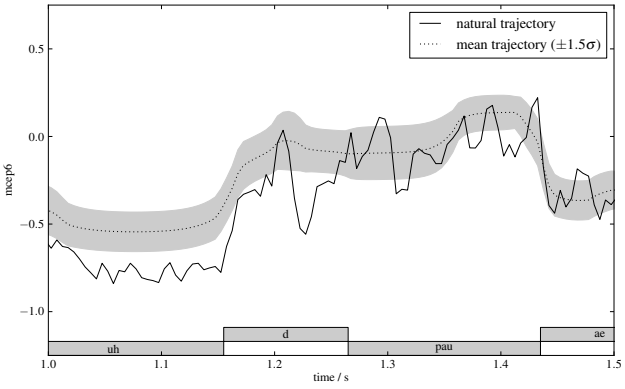
By modifying the procedure used to compute the mean trajectory in conventional *most likely (ML)* synthesis [3], it is possible to efficiently sample trajectories [8]. To allow us to visualize the trajectories for all mel-cepstral components simultaneously we plot running spectra derived from these trajectories.

Figure 2 compares a running spectrum for natural speech with running spectra for sampled trajectories for the three approaches, all for the same utterance. We can see that for the two normalized models, sampling produces a running spectrum that looks qualitatively similar to natural speech, and captures some of its characteristic roughness, whereas for the unnormalized standard approach sampling produces a running spectrum that is slightly too smooth, due to the fact the standard approach underestimates predictive variance.

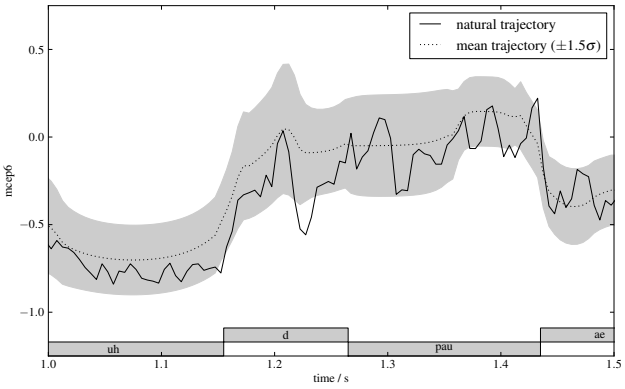
We can also see that sampling produces a much more natural-looking running spectrum than taking the mean (illustrated here for the trajectory HMM), in keeping with our observation that maximum likelihood training implicitly assumes that natural trajectories are generated by sampling.

3.2. Test set log probability

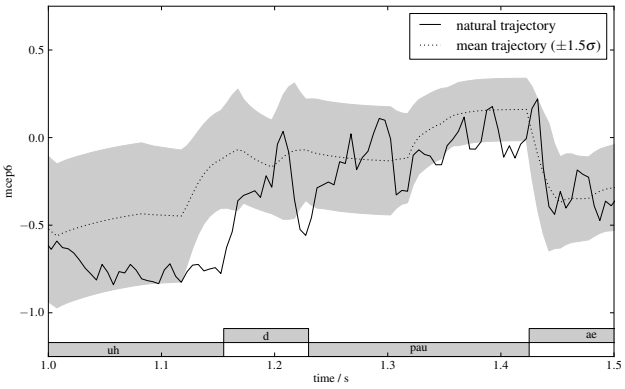
The log probability $\log P(\boldsymbol{c} \,|\, \boldsymbol{l}, \lambda)$ evaluated on a held out test set provides a natural measure of the accuracy of each system as a probabilistic model. As a score, *test set log probability* provides a natural compromise between the model's accuracy in terms of the mean trajectory, the expected pointwise variation around that mean, and the correlations over time present in the variation around the mean. Here we look at the log probability $\log P(\boldsymbol{c} \,|\, \boldsymbol{q}, \lambda)$ for a fixed state sequence $\boldsymbol{q}$.



(a) standard approach



(b) trajectory HMM



(c) autoregressive HMM

Figure 1: Visualization of the distribution over trajectories for each of the three models, together with the natural trajectory actually generated by the speaker (6$^{\text{th}}$ mel-cepstral coefficient, 0.5 seconds of speech, given fixed state sequence).

system	log prob
standard	29.3
trajectory HMM	47.6
autoregressive HMM	47.8

Table 2: Log probability on 50 unseen test set utterances for the three systems (per frame, all mel-cepstral coefficients, given fixed state sequence).

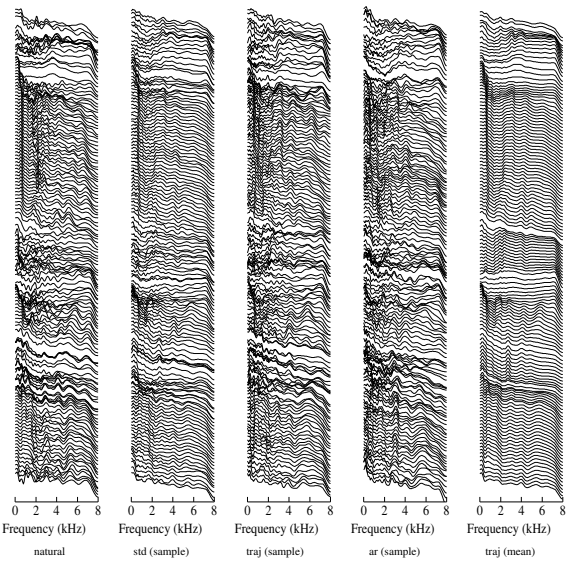


Figure 2: Running spectra for natural speech, for sampled trajectories for each of the three approaches, and for mean trajectories (0.7 seconds of speech, given fixed state sequence).

Table 2 shows the test set log probability for the three systems. We can see that the normalized models have greatly increased test set log probabilities compared to the unnormalized standard approach. This suggests that the normalized models are better as probabilistic models of speech.

The low test set log probability of the standard system is to a large extent due to its lack of predictive variance – artificially boosting the predictive variance by multiplying the covariance matrix $\overline{\Sigma}_{\boldsymbol{q}}$ by a factor of 3 while keeping the same mean trajectory $\overline{\boldsymbol{\mu}}_{\boldsymbol{q}}$ increased the test set log probability of the standard system to 46.9. This is strong evidence that the standard approach systematically underestimates predictive variance as Figure 1 and Figure 2 suggested.

4. Improving the model

We discussed in §3.1.2 that drawing samples from the predictive distribution allows us to investigate whether the probabilistic generative model we are using is reasonable or not. However preliminary experiments showed that speech synthesized from sampled trajectories sounds very artificial and unnatural for all three models, and in particular it sounds much less natural than speech synthesized using mean trajectories. This shows that while current normalized models have *better* predictive distributions than the unnormalized standard approach, they are still far from *good* – they have some major deficiency as probabilistic models of speech.

In this section we look at one possible improvement to current models, namely using full rather than diagonal covariance matrices. Full covariance matrices explicitly model the correlations between different feature vector components within one frame ($\boldsymbol{c}_{t\cdot}$), which are ignored by current normalized models.

We conducted a subjective listening test to compare full and diagonal covariance models. The speech samples to be evaluated were synthesized from the standard HMM by synthesis considering global variance, the full covariance trajectory HMM by ML generation, the full covariance trajectory HMM

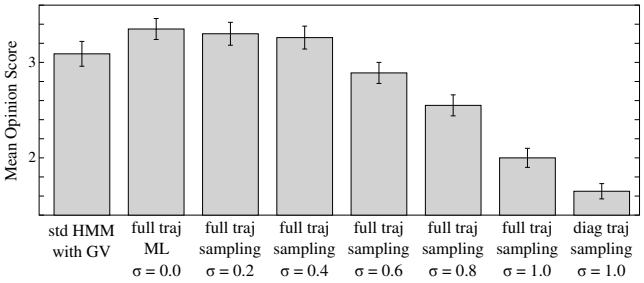


Figure 3: MOS test results.

by random sampling with different values of Gaussian source noise ($\sigma = 0.2$, 0.4, 0.6, 0.8 or 1.0, where $\sigma = 1.0$ corresponds to true random sampling), and the diagonal covariance trajectory HMM by random sampling. There were 800 (100 sentences $\times$ 8 systems) samples in the test. One subject could evaluate up to 320 test samples in the test, which were randomly chosen and presented for each subject. Each test sample was evaluated by three subjects. In the test, after the subjects had listened to a test sample, they were asked to assign it a similarity score from a five-point Likert scale where 5 is completely natural and 1 is completely unnatural. In total 16 subjects participated in the MOS test. The full covariance trajectory HMM system used feature-space MLLR to approximate a true full covariance system. Because of time constraints, only the trajectory HMM was used.

Figure 3 shows the subjective listening test results. We can see that sampled trajectories have significantly worse quality than mean ones, but that samples from the full covariance trajectory HMM do sound more natural than samples from the standard trajectory HMM. This indicates that better intra-frame correlation modelling improves the predictive distribution, but that even this more powerful model is not a satisfactory probabilistic model of speech.

5. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.

[2] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features," *Computer Speech and Language*, vol. 21, no. 1, pp. 153–173s, 2007.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, 2000, pp. 1315–1318.

[4] M. Shannon and W. Byrne, "Autoregressive HMMs for speech synthesis," in *Proc. Interspeech 2009*, 2009, pp. 400–403.

[5] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[6] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[7] M. Shannon and W. Byrne, "A formulation of the autoregressive HMM for speech synthesis," Department of Engineering, University of Cambridge, UK, Technical Report CUED/F-INFENG/TR.629, 2009.

[8] K. Tokuda, H. Zen, and T. Kitamura, "Reformulating the HMM as a trajectory model," in *Proc. of Beyond HMM – Workshop on statistical modeling approach for speech recognition*, 2004.