

Autoregressive HMMs for speech synthesis

Matt Shannon, William Byrne

Cambridge University Engineering Department, U.K.

sms46@eng.cam.ac.uk, bill.byrne@eng.cam.ac.uk

Abstract

We propose the *autoregressive HMM* for speech synthesis. We show that the autoregressive HMM supports efficient EM parameter estimation and that we can use established effective synthesis techniques such as *synthesis considering global variance* with minimal modification. The autoregressive HMM uses the same model for parameter estimation and synthesis in a consistent way, in contrast to the standard HMM synthesis framework, and supports easy and efficient parameter estimation, in contrast to the *trajectory HMM*. We find that the autoregressive HMM gives performance comparable to the standard HMM synthesis framework on a Blizzard Challenge-style naturalness evaluation.

Index terms: HMM-based speech synthesis, acoustic modeling

1. Introduction

It has been shown that it is possible to synthesize natural sounding speech with HMMs and the quality of the best HMM-based synthesis systems now rivals the best unit selection synthesis systems [1]. A breakthrough that helped make this possible was realizing how to use dynamic feature information during synthesis, by respecting the constraints between static and dynamic features [2].

However the established approach to HMM-based synthesis is inconsistent in the enforcement of these constraints [3]. During synthesis we take the constraints between static and dynamic features into account, whereas during parameter estimation we assume the static and dynamic feature sequences are independent.

This is a recognized problem and has been addressed previously. Zen showed how a *trajectory HMM* [3] could be employed so that the same model is used for both parameter estimation and synthesis in a consistent way. Synthesis quality improved as a result [3]. However parameter estimation for the trajectory HMM is more complicated than for the standard HMM, requiring alignment with a delayed-decision Viterbi algorithm and gradient-based parameter re-estimation procedures [3]. The challenge remains to find a model which can easily and consistently be used for both parameter estimation and synthesis.

In this paper we propose using the *autoregressive HMM* [4, 5, 6, 7] for speech synthesis. The autoregressive HMM relaxes the traditional HMM conditional independence assumption, allowing state output distributions which depend on past output as well as the current state. In this way the autoregressive HMM explicitly models some of the dynamics of speech and introduces the continuity and context dependence needed for good quality synthesis.

Autoregressive HMMs have been used before for speech recognition [4, 5, 6, 8], but to our knowledge they have not

been previously investigated for speech synthesis. Note that for the autoregressive HMM considered here, the observations are acoustic feature vectors. This is distinct from the *hidden filter HMM* (also sometimes called the autoregressive HMM) [9, 10] for which the observations are waveform samples.

We show that the autoregressive HMM supports efficient parameter estimation and synthesis techniques, with the same model used in both parameter estimation and synthesis in a consistent way. From a theoretical viewpoint we highlight some of the similarities and differences between the autoregressive HMM and current models. We compare the autoregressive HMM to the standard HMM synthesis framework on a Blizzard Challenge-style naturalness evaluation and find that the autoregressive HMM gives performance that is comparable to the standard HMM synthesis framework.

In section §2, we present the autoregressive HMM. We specify the model itself and show how to do efficient parameter estimation using expectation maximization. We show that the autoregressive HMM is similar enough to the current use of HMMs in synthesis that we can use established effective synthesis techniques, such as synthesis using dynamic features and synthesis considering global variance, with little modification. In section §3, we give experimental results showing that the naturalness of speech synthesized by the autoregressive HMM is comparable to the standard framework. Finally, in section §4 we give conclusions.

2. Autoregressive HMM

We start with a general generative model for sequences of acoustic feature vectors. Conceptually we first generate a *hidden* state sequence $\theta = \theta_{1:T}$ and then generate an *observed* or *output* feature vector sequence $c = c_{1:T}$ given this state sequence. We consider models with a joint probability distribution of the form:

$$\begin{aligned} P(c, \theta) &= P(\theta)P(c|\theta) \\ &= \prod_t P(\theta_t|\theta_{t-1})P(c_t|c_{1:t-1}, \theta_t) \end{aligned} \quad (1)$$

The *state transition probabilities* $P(\theta_t|\theta_{t-1})$ are conditioned only on the previous state. The *state output distributions* $P(c_t|c_{1:t-1}, \theta_t)$ are conditioned on both the current state and all past output. This is in contrast to the standard HMM assumption that the state output distribution $P(c_t|\theta_t)$ is conditionally independent of past output.

The *autoregressive HMM with summarizers* specializes the above to a particular form of output distribution $P(c_t|c_{1:t-1}, \theta_t)$. We assume c_t is conditionally Gaussian, with covariance depending only on the state θ_t . Rather than allowing the mean for each state to be an arbitrary function of past output $c_{1:t-1}$, we restrict it to be a linear function of a fixed set of *summarizers* of past output. Each summarizer f^d is a function that

window	offset			
	-3	-2	-1	0
w^1			1.0	
w^2		-1.0	1.0	
w^3	1.0	-2.0	1.0	

Table 1: Example autoregressive window coefficients

takes the entire past output $c_{1:t-1}$ and produces a vector-valued *summary* $f^d(c_{1:t-1})$. We consider state output distributions of the form:

$$P(c_t | c_{1:t-1}, \theta_t) = \mathcal{N}(c_t | \mu_{\theta_t}(c_{1:t-1}), \Sigma_{\theta_t}) \quad (2)$$

$$\mu_q(c_{1:t-1}) \triangleq \sum_{d=1}^D A_q^d (f^d(c_{1:t-1}) - \mu_q^d) + \mu_q^0 \quad (3)$$

where Σ_q is a state-dependent covariance matrix, A_q^d is a matrix for each summary d and state q , μ_q^0 is a state-dependent bias vector, and following Woodland [6], we have introduced redundant bias vectors μ_q^d for each summary d and state q as a trick to make re-estimation easier. The set of parameters specifying the autoregressive HMM is therefore $(A_{qij}^d, \mu_{qi}^d, \mu_{qi}^0, \Sigma_{qij})$, where q ranges over states, i and j range over feature vector components, d ranges over summarizers, and A_{qij}^d is the (i, j) -component of the matrix A_q^d in (3).

We are free to choose the summarizers (f^d) to be anything which might distill useful information about past output. However, for simplicity we will take each f^d to be a fixed linear combination of the past l_d feature vectors:

$$f^d(c_{1:t-1}) = \sum_{k=-l_d}^{-1} w_k^d c_{t+k} \quad (4)$$

We call the linear summarizers *windows*, with *window coefficients* w_k^d . These window coefficients are only non-zero in the past ($k < 0$). An example of autoregressive window coefficients is shown in Table 1.

By setting the windows to be $w_k^d = \delta_k^{-d}$ we recover a canonical autoregressive HMM [4, 5, 7]. By setting the windows to be delta functions at fixed offsets from the current time we obtain the form of model used by Woodland [6] and Chin [8]. We refer to all these models, including the mild generalization with summarizers presented above, as the *autoregressive HMM*.

Note that we only explicitly deal with the *static* feature vector sequence c for the autoregressive HMM. However, the role played by linear summarizers here is somewhat similar to that of *dynamic features* in the standard HMM framework.

As is common in modelling speech, we will assume the feature vector sequence components are independent given the state sequence. This corresponds to using diagonal matrices in (3), so $A_{qij}^d = a_{qi}^d \delta_{ij}$ and $\Sigma_{qij} = \sigma_{qi}^2 \delta_{ij}$ for some a_{qi}^d and σ_{qi}^2 .

2.1. Parameter estimation

To set the parameters $(a_{qi}^d, \mu_{qi}^d, \mu_{qi}^0, \sigma_{qi}^2)$ of the autoregressive HMM we use *expectation maximization (EM)* [11]. We first compute the *state occupancies* $\gamma_q(t) \triangleq P(\theta_t = q | c)$, then use these to re-estimate the model parameters.

2.1.1. Forward-Backward algorithm

Define:

$$\alpha_q(t) \triangleq P(c_{1:t}, \theta_t = q)$$

$$\beta_q(t) \triangleq P(c_{t+1:T} | c_{1:t}, \theta_t = q)$$

Then we have the following recursions

$$\alpha_q(t) = \sum_p \alpha_p(t-1) u_{pq} P(c_t | c_{1:t-1}, \theta_t = q) \quad (5)$$

$$\beta_q(t) = \sum_r u_{qr} P(c_{t+1} | c_{1:t}, \theta_{t+1} = r) \beta_r(t+1) \quad (6)$$

where $u_{pq} \triangleq P(\theta_t = q | \theta_{t-1} = p)$ is the state transition probability. This allows us to efficiently compute α and β , and thus the state occupancies:

$$\gamma_q(t) = \frac{\alpha_q(t) \beta_q(t)}{\sum_q \alpha_q(t) \beta_q(t)}$$

2.1.2. Parameter re-estimation

We write $f^d(t)$ for the value of summarizer d at time t , that is $f^d(c_{1:t-1})$. We use the notation

$$\langle g \rangle_q \triangleq \frac{\sum_t \gamma_q(t) g(t)}{\sum_t \gamma_q(t)}$$

to denote the conditional expectation of an arbitrary function $g(t)$ with respect to the occupancies $\gamma_q(t)$ of state q .

The EM re-estimation formulae giving the updated parameter values $(\hat{a}_{qi}^d, \hat{\mu}_{qi}^d, \hat{\mu}_{qi}^0, \hat{\sigma}_{qi}^2)$ for the autoregressive HMM are [12]:

$$\hat{\mu}_{qi}^0 = \langle c_i \rangle_q \quad (7)$$

$$\hat{\mu}_{qi}^d = \langle f_i^d \rangle_q \quad (8)$$

$$\sum_{e=1}^D R_{qi}^{de} \hat{a}_{qi}^e = r_{qi}^d \quad (9)$$

$$\hat{\sigma}_{qi}^2 = r_{qi}^0 - \sum_{d=1}^D \hat{a}_{qi}^d r_{qi}^d \quad (10)$$

where q ranges over states, i ranges over feature vector components, d and e range over summarizers, and:

$$R_{qi}^{de} \triangleq \langle f_i^d f_i^e \rangle_q - \langle f_i^d \rangle_q \langle f_i^e \rangle_q$$

$$r_{qi}^d \triangleq \langle c_i f_i^d \rangle_q - \langle c_i \rangle_q \langle f_i^d \rangle_q$$

$$r_{qi}^0 \triangleq \langle c_i c_i \rangle_q - \langle c_i \rangle_q \langle c_i \rangle_q$$

Note that computing the (\hat{a}_{qi}^d) using (9) involves inverting a $D \times D$ matrix for each q and i . In our experiments, we use $D = 3$ summarizers, so this is not computationally intensive.

2.2. Synthesis

During synthesis, we produce an output feature sequence c for a given word sequence. We show how to do synthesis with the autoregressive HMM by adapting two standard algorithms.

In fact, from the point of view of synthesis there is a strong similarity between the autoregressive HMM and standard HMM synthesis framework. In both cases, $P(c|\theta)$ is a multidimensional Gaussian over vector sequences with band diagonal precision matrix [12, 13]. This common structure is what makes it possible to use current HMM synthesis methods with the autoregressive HMM.

2.2.1. Synthesis using dynamic features

For synthesis using dynamic features ([2] and case 1 in [14]), we first choose a state sequence θ and then choose the feature sequence c which maximizes $P(c|\theta)$. For standard HMM synthesis models, $P(c|\theta)$ is a multidimensional Gaussian over vector sequences [3], and so the maximum value is at its mean, which can be computed efficiently [14].

For the autoregressive HMM with linear summarizers, $P(c|\theta)$ is also a multidimensional Gaussian over vector sequences. The mean functions $\mu_q(c_{1:t-1})$ in (3) are now affine-linear, and expectation is a linear operator. Therefore the mean vector sequence $\bar{c} \triangleq \mathbb{E}[c|\theta]$ can be computed efficiently by a one-pass forward recursion over time:

$$\bar{c}_t = \mu_{\theta_t}(\bar{c}_{1:t-1}) \quad (11)$$

2.2.2. Synthesis considering global variance

Standard techniques in HMM synthesis, such as synthesis using dynamic features, are found to produce utterances that sound “flat” or “dull” [15]. In particular, it is found that synthesized utterances tend to have far less *global variance (GV)* than natural ones, where the global variance $v(c_i)$ of the i^{th} component of the feature vector sequence is defined as [15]:

$$v(c_i) \triangleq \frac{1}{T} \sum_t c_{ti}^2 - \left(\frac{1}{T} \sum_t c_{ti} \right)^2$$

Toda [15] introduced parameter generation *considering global variance* as a way to alleviate this lack of global variance, while using existing models. The distribution of global variances observed in training utterances is modelled by a Gaussian, typically treating each component of the feature vector as independent. The HMM and GV parameters are trained independently of each other. During synthesis, we use some form of gradient descent to optimize a cost function that is a weighted sum of the HMM log probability of the output sequence and the GV log probability of the output sequence (keeping the state sequence fixed). This procedure is found to dramatically improve the naturalness of synthetic speech [15].

It is trivial to adapt this for use with the autoregressive HMM. Since we keep the state sequence fixed during gradient descent, the HMM log probability is in both cases just a multidimensional Gaussian. Therefore we can do parameter generation considering global variance for the autoregressive HMM simply by passing the appropriate multidimensional Gaussian to the GV generation algorithm.

2.3. Comparison to current models

The parameterization of state output distribution for the autoregressive HMM in (2) and (3) is very different to the parameterization of output distribution shared by the standard HMM framework and the trajectory HMM (equation 9 in [2], equation 5 in [3]). Nevertheless, in all three cases the distribution over trajectories $P(c|\theta)$ given the state sequence is a multidimensional Gaussian with band diagonal precision matrix [3, 12]. This suggests the difference in parameterization of state output distribution might result in only a small difference in the kind of trajectories generated by the different models. Indeed, anecdotally we have found that the trajectories generated by synthesis using dynamic features are extremely similar for the autoregressive HMM and standard HMM synthesis framework.

Additional similarities and differences were discussed in §1. The autoregressive HMM and trajectory HMM share the

system	description
A	natural speech
B	autoregressive HMM with global variance
C	standard HMM with global variance
D	autoregressive HMM without global variance

Table 2: Systems in the listening test

property of consistency, in contrast to the standard HMM synthesis framework. The autoregressive HMM and standard HMM synthesis framework both have easy and efficient parameter estimation procedures, in contrast to the more complicated parameter estimation methods required for the trajectory HMM.

3. Experiments

To evaluate the autoregressive HMM for synthesis, we built a baseline standard HMM system and an autoregressive HMM system, and compared them in a Blizzard Challenge-style [16] mean opinion score (MOS) listening test. We chose an MOS evaluation over a preference test since we were interested in whether the two methods were broadly comparable – that is, in the magnitude of the difference between the systems – rather than a consistent preference of unknown magnitude.

Both systems were built using the *HMM-based speech synthesis system (HTS)* [17]. The similarity in parameter estimation and synthesis methods between the autoregressive HMM and standard HMM framework allowed us to implement the autoregressive HMM relatively easily in HTS. The systems were trained on the CMU ARCTIC corpus [18] for a single speaker (approximately 1 hour), with 50 held-out utterances. The static features were mel-generalized cepstra (MGC), log F0, and band aperiodicity, and we used STRAIGHT vocoding [19]. For the autoregressive system only spectral features were modelled using the autoregressive HMM. The windows in Table 1 were used for the autoregressive HMM, and standard HTS windows for the autoregressive HMM. The training regime was adapted from the HTS speaker dependent training demo [17].

For simplicity and ease of implementation, the autoregressive HMM used acoustic clustering determined by decision trees taken from the baseline HTS system. There may be scope for improvement by direct state clustering of the autoregressive observation densities, which can easily be done.

The complexity of the two systems is similar, with 5 free parameters per state per feature vector component for the autoregressive HMM and 6 for the standard HMM.

The listening test was conducted using the four systems shown in Table 2. The systems to be investigated are B and C. Systems A and D are included so that the systems of interest are less likely to be at the extreme ends of the MOS spectrum, and to help detect lack of listener seriousness, though we did not find any occurrences of this. The test consisted of 2 sections, of 25 utterances each. For both sections, listeners were asked to rate the *naturalness* of each utterance on a scale of 1 to 5. Prompts were the 50 held-out utterances in a fixed order. Listeners were allotted to one of 4 groups, and the ordering of the systems for each group was determined with a balanced Latin square design for the first 4 prompts in each section, and randomly after that. The listening test was conducted as an interactive website for one week.

system	native		non-native	
	mean	median	mean	median
A	4.7	5	4.8	5
B	2.3	2	2.2	2
C	2.3	2	2.4	2
D	2.0	2	1.6	1

Table 3: MOS listening test results

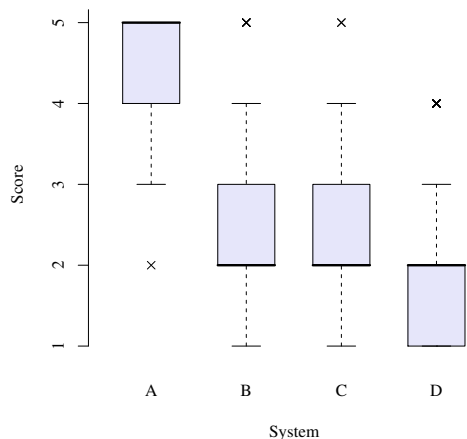


Figure 1: MOS listening test results (native)

3.1. Results

In total, 39 volunteers (24 native English speakers, 15 non-native) completed the test. A summary of the results is shown in Table 3 and an MOS box plot [20] for the native listeners is shown in Figure 1. The non-native box plot is not shown here due to lack of space and showed a slight preference for system C consistent with Table 3.

We can see that the naturalness of speech from the autoregressive HMM (system B) is broadly comparable to that of the standard HMM synthesis framework (system C), which is what we wished to establish.

4. Conclusion

We have investigated the possibility of using the autoregressive HMM for speech synthesis. We have shown that it is possible to use the same autoregressive model to do efficient parameter estimation and synthesis. The autoregressive HMM is therefore *efficient* and *consistent* whereas current models have only one of these two desirable properties. There is also considerable similarity to current models and minor modifications to existing procedures are needed for implementation. We have demonstrated that the autoregressive HMM can be used to synthesize natural speech of comparable quality to the standard HMM synthesis framework.

5. Acknowledgements

We are very grateful to the organizers of the Blizzard Challenge for providing scripts to conduct our experimental evaluation. This research was funded by the European Community’s Seventh Framework Programme (FP7/2007-2013), grant agreement 213845 (EMIME).

6. References

- [1] A. Black, H. Zen, and K. Tokuda, “Statistical parametric speech synthesis,” in *Proc. ICASSP 2007*, 2007, pp. 1229–1232.
- [2] K. Tokuda, T. Kobayashi, and S. Imai, “Speech parameter generation from HMM using dynamic features,” in *Proc. ICASSP 1995*, vol. 1, 1995.
- [3] H. Zen, K. Tokuda, and T. Kitamura, “An Introduction of Trajectory Model into HMM-Based Speech Synthesis,” in *Proc. Fifth ISCA Workshop on Speech Synthesis*, 2004.
- [4] C. Wellekens, “Explicit time correlation in hidden Markov models for speech recognition,” in *Proc. ICASSP 1987*, vol. 12, 1987.
- [5] P. Kenny, M. Lennig, and P. Mermelstein, “A linear predictive HMM for vector-valued observations with applications to speech recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 2, pp. 220–225, 1990.
- [6] P. Woodland, “Hidden Markov models using vector linear prediction and discriminative output distributions,” in *Proc. ICASSP 1992*, vol. 1, 1992, pp. 509–512.
- [7] J. Bilmes, “Graphical models and automatic speech recognition,” in *Mathematical foundations of speech and language processing*, M. Johnson, S. Khudanpur, M. Ostendorf, and R. Rosenfeld, Eds. Springer-Verlag, 2004.
- [8] K. Chin and P. Woodland, “Maximum mutual information training of hidden Markov models with vector linear predictors,” in *Proc. Interspeech 2002*, 2002.
- [9] A. Poritz, “Linear predictive hidden Markov models and the speech signal,” in *Proc. ICASSP 1982*, vol. 7, 1982.
- [10] B. Juang and L. Rabiner, “Mixture autoregressive hidden Markov models for speech signals,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 6, pp. 1404–1413, 1985.
- [11] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, Series B (Methodological)*, pp. 1–38, 1977.
- [12] M. Shannon, “A formulation of the autoregressive HMM for speech synthesis,” University of Cambridge, Technical Report CUED/F-INFENG/TR.629, 2009, <http://mi.eng.cam.ac.uk/~sms46/papers/shannon2009fah.pdf>.
- [13] H. Zen, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features,” Ph.D. dissertation, Nagoya Institute of Technology, 2006.
- [14] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *Proc. ICASSP 2000*, vol. 3, 2000.
- [15] T. Toda and K. Tokuda, “Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis,” in *Proc. Interspeech 2005*, 2005.
- [16] A. Black and K. Tokuda, “The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets,” in *Proc. Interspeech 2005*, 2005.
- [17] HTS working group, “HMM-based speech synthesis system (HTS),” <http://hts.sp.nitech.ac.jp/>, accessed 17 April 2009.
- [18] J. Kominek and A. Black, “The CMU ARCTIC databases for speech synthesis,” Carnegie Mellon University, Technical Report CMU-LTI-03-177, 2003, <http://festvox.org/cmu-arctic/>.
- [19] H. Kawahara, “Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited,” in *Proc. ICASSP 1997*, vol. 2, 1997.
- [20] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, “Statistical analysis of the Blizzard Challenge 2007 listening test results,” in *Proc. Blizzard Challenge Workshop (Sixth ISCA Workshop on Speech Synthesis)*, 2007.