Autoregressive clustering for HMM speech synthesis

Matt Shannon, William Byrne

Cambridge University Engineering Department, U.K.

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

Abstract

The *autoregressive HMM* has been shown to provide efficient parameter estimation and high-quality synthesis, but in previous experiments decision trees derived from a non-autoregressive system were used.

In this paper we investigate the use of autoregressive clustering for autoregressive HMM-based speech synthesis. We describe decision tree clustering for the autoregressive HMM and highlight differences to the standard clustering procedure. Subjective listening evaluation results suggest that autoregressive clustering improves the naturalness of the resulting speech.

We find that the standard *minimum description length (MDL)* criterion for selecting model complexity is inappropriate for the autoregressive HMM. Investigating the effect of model complexity on naturalness, we find that a large degree of overfitting is tolerated without a substantial decrease in naturalness.

Index terms: HMM-based speech synthesis, decision tree clustering, autoregressive HMM

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 take the constraints between static and dynamic features into account during synthesis [2]. However the established approach to HMM-based synthesis ignores these constraints during parameter estimation [3].

The *autoregressive HMM* [4, 5, 6, 7] provides an alternative to the standard HMM synthesis framework. It supports efficient parameter estimation using expectation maximization and allows high quality synthesis of comparable naturalness to the standard framework, while treating static and dynamic features consistently [8]. However previous experiments with the autoregressive HMM have used decision trees derived from a standard, non-autoregressive system [8], which is not theoretically well-motivated.

In this paper we investigate using autoregressive clustering for the autoregressive HMM. We describe autoregressive decision tree clustering and highlight differences to the standard clustering procedure. We assess the naturalness of the resulting speech in a subjective listening evaluation. Decision tree clustering requires choosing an appropriate model complexity, and so we explore the effect of model complexity on naturalness. We discuss selecting an appropriate model complexity automatically using the *minimum description length (MDL)* criterion [9]. We use the log probability on a held-out test set to inform our investigation into model complexity and to measure the degree of overfitting.
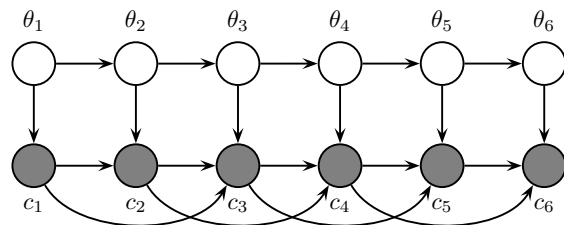

Figure 1: graphical model for autoregressive HMM of depth 2

2. Autoregressive HMM

We briefly review the autoregressive HMM model. The autoregressive HMM [4, 5, 6, 7, 8] is a generative model for sequences of pairs of an acoustic feature vector and a hidden state. The joint distribution over the *hidden* state sequence $\theta = \theta_{1:T}$ and the *observed* or *output* static feature vector sequence $c = c_{1:T}$ is

$$P(c, \theta) = \prod_t P(\theta_t | \theta_{t-1}) \prod_i P(c_t^i | c_{t-K:t-1}^i, \theta_t) \quad (1)$$

where c_t^i is the ith component of the output feature vector at time t and $K \in \mathbb{N}$ is referred to as the *depth* of the model. A graphical model for the case $K = 2$ is shown in Figure 1. The *state output distributions* $P(c_t^i | c_{t-K:t-1}^i, \theta_t)$ are normal with a mean that depends on past output:

$$P(c_t^i | c_{t-K:t-1}^i, \theta_t = q) = \mathcal{N}(c_t^i | \mu_q^i(c_{t-K:t-1}^i), (\sigma^2)_q^i) \quad (2)$$

$$\mu_q^i(v) = \sum_{d=1}^{D} a_q^{id} f^{id}(v) \quad (3)$$

where each $f^{id} : \mathbb{R}^K \to \mathbb{R}$ is a fixed *summarizer* that computes a real-valued summary of the recent past output $c_{t-K:t-1}^i$. The parameters of the autoregressive HMM are $(a_q^{id}, (\sigma^2)_q^i)$. We typically use $D = K + 1$ summarizers of the form $f^{id}(v) = v^d$ (the dth component of v) for $1 \leq d < D$ and $f^{iD}(v) = 1$, so the mean $\mu_q^i(c_{t-K:t-1}^i)$ is a state-dependent linear combination of the recent past output plus a bias.

We have taken the summarizers f^{id} to be functions of the recent past output in the same feature vector component i. This is consistent with the common assumption when modelling speech that the feature vector components (c^i) are independent given the state sequence θ. However it causes no problem if the summarizers depend on all recent past output $c_{t-K:t-1}$, or even on the present output up to the given component $c_t^{1:i-1}$.

3. Parameter estimation

The autoregressive HMM permits efficient parameter estimation using expectation maximization (EM) [6, 8]. Here we

summarize the re-estimation formulae used to compute updated parameter values given the *state occupancies* $\gamma_q(t)$ obtained using the Forward-Backward algorithm [10].

Defining a dummy summarizer $f^{i0}(t) \triangleq c_t^i$ for notational convenience, we define *accumulators*

$$S_q^{ide} \triangleq \sum_t \gamma_q(t) f^{id}(t) f^{ie}(t) \qquad (4)$$

where q ranges over states, i ranges over feature vector components, and $0 \le d, e \le D$. Here we are considering f^{id} as a function over time $f^{id}(t) = f^{id}(c_{t-K:t-1}^i)$.

The re-estimation formulae giving the updated parameter values $(\widehat{a}_q^{id}, (\widehat{\sigma}^2)_q^i)$ are

$$\sum_{e=1}^{D} S_q^{ide} \widehat{a}_q^{ie} = S_q^{id0} \qquad (5)$$

$$(\widehat{\sigma}^2)_q^i = \frac{1}{\gamma_q} \left(S_q^{i00} - \sum_{d=1}^{D} \widehat{a}_q^{id} S_q^{id0} \right) \qquad (6)$$

where q ranges over states, i ranges over feature vector components, $1 \le d \le D$, and $\gamma_q \triangleq \sum_t \gamma_q(t)$. Note that computing the $(\widehat{a}_q^{id})$ using (5) involves storing and inverting a $D \times D$ matrix for each q and i. For the experiments below $D = 4$.

The value of the EM auxiliary function at its maximum $(\widehat{a}_q^{id}, (\widehat{\sigma}^2)_q^i)$ is

$$-\frac{T}{2}(\log 2\pi + 1) - \frac{1}{2} \sum_q \gamma_q \sum_i \log(\widehat{\sigma}^2)_q^i \qquad (7)$$

4. Decision tree clustering

The standard approach to *decision tree clustering* [11] is modified for the autoregressive HMM. The objective function is based on the auxiliary function (7) in the usual way. Ignoring constants the objective function is

$$-\frac{1}{2} \sum_C \gamma_C \sum_i \log(\widehat{\sigma}^2)_C^i - \xi \cdot \#\{\text{leaves}\} \qquad (8)$$

where ξ is referred to as the *clustering threshold*.

From (7) we can see that the change in likelihood for a split of cluster C into two pieces C_1 and C_2 is

$$\frac{1}{2}\gamma_C \sum_i \log(\widehat{\sigma}^2)_C^i - \frac{1}{2}\gamma_{C_1} \sum_i \log(\widehat{\sigma}^2)_{C_1}^i$$
$$-\frac{1}{2}\gamma_{C_2} \sum_i \log(\widehat{\sigma}^2)_{C_2}^i \qquad (9)$$

The corresponding change in the number of leaves is 1. Therefore to greedily optimize the objective function we recursively split each leaf using the question that maximizes the change in likelihood, unless the maximum achievable change is less than ξ in which case we do not split that leaf. Note that (9) depends only on C, C_1 and C_2 and not on the other clusters, so the order in which we choose to split leaves makes no difference.

We can compute the accumulators for an arbitrary cluster just by summing the state-level accumulators (4) for that cluster. Thus we can use (6) to compute $(\widehat{\sigma}^2)_C^i$ for each cluster C and so compute the change in likelihood (9) for a hypothesized split.

Typically a hard minimum occupancy constraint is also imposed on each leaf, which can be incorporated above by setting the objective function to $-\infty$ for any tree that violates this constraint.

4.1. Differences to the standard HMM framework

For the autoregressive HMM the updated parameter values $(\widehat{a}_q^{id}, (\widehat{\sigma}^2)_q^i)$ together with state occupancies (γ_q) are *not* sufficient to recover the accumulators (4). This means we must pass the decision tree clustering algorithm the accumulators themselves, and not just the re-estimated parameter values together with occupancies as for the standard HMM framework.

4.2. Minimum Description Length (MDL)

The *minimum description length (MDL)* criterion [9] for the standard HMM framework allows automated setting of the clustering threshold ξ. It sets

$$\xi = \frac{1}{2}k \log N \qquad (10)$$

where k is the number of free parameters per leaf and N is the total occupancy of the root node. Our experimental results show that (10) is not directly applicable to the autoregressive HMM.

5. Experiments

To evaluate autoregressive clustering for synthesis, we built baseline autoregressive HMM systems (S1 and S2) using standard, non-autoregressive decision trees, and compared these to fully autoregressive systems (A1-7) using the clustering procedure outlined in §4. We compared the systems in a Blizzard Challenge-style [12] subjective listening evaluation.

All systems were built using the *HMM-based speech synthesis system (HTS)* [13]. The systems were trained on the CMU ARCTIC corpus [14] for a single speaker (approximately 1 hour) with 50 held-out utterances. The static features were 40-dim mel-generalized cepstra ($\gamma = 0$, $\alpha = 0.42$) (mgc), 0/1-dim log F0 (lf0), and 5-dim band aperiodicity (bap), and we used STRAIGHT vocoding [15]. The mgc and bap streams were modelled using the autoregressive HMM. For the baseline systems (S1 and S2) we first trained a standard HMM synthesis system including standard clustering, and then used the Forward-Backward occupancies from the standard system to estimate the parameters of an autoregressive system for the mgc and bap streams, a procedure we call *cross-training*. There were a similar number of free parameters per leaf per static feature vector component during standard clustering as during autoregressive clustering (8 vs 5). Windows were as in previous work [8] and were the same for all systems. The mgc and bap windows had depth $K = 3$. Other details of the systems were standard [13]. For each system the 50 test set utterances were synthesized using *synthesis considering global variance* [16].

Clustering thresholds were chosen as follows. For the standard clustering systems (S1 and S2) the mgc clustering threshold was set by MDL using (10). Since this is an initial investigation into autoregressive clustering we tested several mgc thresholds (A1-7). Systems A1-7 and S1 are directly comparable in terms of non-mgc model complexity, with the clustering thresholds for S1 set near but not at their MDL values. For comparison we also built a baseline system S2 using MDL clustering thresholds for all streams.

The listening test was conducted using the systems shown in Table 1. The test consisted of 5 sections of 10 utterances each. For all 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 10 groups, and the ordering of the systems for each group was determined with a balanced Latin square design. The listening test was conducted as an interactive website for two weeks.

type	system	thresh	mean	median
natural	N		4.8	5
AR with std trees	S1	MDL	2.5	2
	S2	MDL	2.5	2
AR	A1	180	2.6	3
	A2	190	2.6	3
	A3	220	2.7	3
	A4	300	2.7	3
	A5	500	2.3	2
	A6	1200	1.9	2
	A7	5000	1.2	1

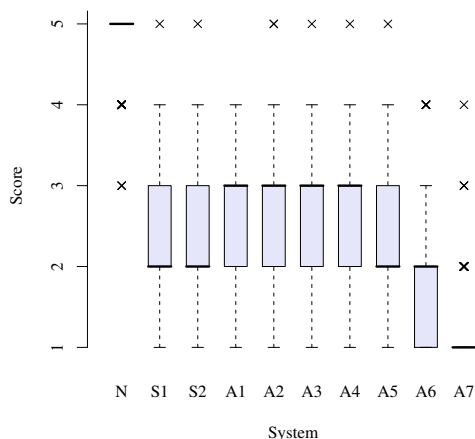Table 1: systems and naturalness opinion score results (each system rated on a total of 170 utterance-listens)

Figure 2: naturalness opinion score results

	N	S1	S2	A1	A2	A3	A4	A5	A6	A7
N		■	■	■	■	■	■	■	■	■
S1	■		□	□	□	□	□	□	■	■
S2	■	□		□	□	□	□	□	■	■
A1	■	□	□		□	□	□	□	■	■
A2	■	□	□	□		□	□	□	■	■
A3	■	□	□	□	□		□	■	■	■
A4	■	□	□	□	□	□		■	■	■
A5	■	□	□	□	□	■	■		■	■
A6	■	■	■	■	■	■	■	■		■
A7	■	■	■	■	■	■	■	■	■	

Table 2: pairwise comparisons of significant differences between naturalness using Bonferroni-corrected Mann-Whitney U tests (■ indicates a significant difference at 1%)
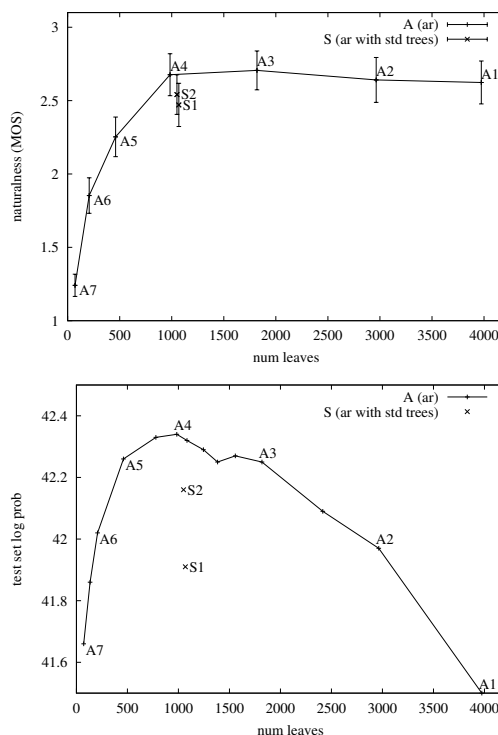

Figure 3: naturalness (mean opinion score) and test set log probability against number of mgc leaves. 1000 leaves corresponds to an average of 650 frames per leaf.

5.1. Results

In total 34 native English speakers completed the evaluation. Table 1 shows a summary of the results. Figure 2 is an opinion score box plot [17], and a matrix of statistically significant differences between the various systems is shown in Table 2.

The best autoregressive clustering systems (A2-4) appear to be slightly more natural than the two baseline systems (S1-2), though the difference between their score distributions is not statistically significant at the 1% level. The overly simple systems (A6-7) perform significantly worse. The two baseline systems (S1-2) are very similar to each other as we would expect.

Varying the clustering threshold varies the number of free parameters in the model. Too many free parameters will cause overfitting. Too few free parameters will lead to a model that is too simple to accurately model the data. To investigate this relationship we plot naturalness against the number of free parameters in Figure 3 (with approximate 95% confidence intervals[1]). We can see that using anywhere from 1000 to 3000 leaves (corresponding to an average of 650 to 220 frames per leaf) is roughly optimal.

---

[1] using the central limit theorem to assume a normal distribution on the sample mean (using estimated variance of the distribution over scores for each system). Also assumes each individual opinion scoring event is independent of all other scoring events (no sequential effects), and that the score for a given (system, utterance text, listener) is independent of the utterance text and listener.

The computation of mean opinion score and confidence intervals in Figure 3 is justified by the fact that people do appear to treat the 5-point Likert scale as an interval scale in this context, even though a priori we can only assume it is an ordinal scale [18].

5.2. Test set log probability

For each of the systems we measured the *test set log probability* – the log probability per frame that the model assigns to the 50 held-out utterances. This allows us to measure the degree of underfitting (a low test set log probability caused by too few parameters) and overfitting (a low test set log probability caused by too many parameters).

Test set log probability is plotted against number of free parameters in Figure 3. The optimal model complexity is around 1000 leaves (650 frames per leaf). This is also near-optimal in terms of naturalness. Interestingly this is roughly the same as the number of leaves selected by MDL for standard clustering. Underfitting degrades naturalness noticeably (A5-7). However a large degree of overfitting (A1-2) is tolerated without a substantial decrease in naturalness. The difference between the baseline systems (S1-2) and the best autoregressive clustering systems is more pronounced for test set log probability than for naturalness.

5.3. Negative effects of standard clustering

The test set log probability for S1 is a surprising amount lower than for S2. After further investigation we found that this can largely be attributed to an extremely low log probability for the mgc stream of one particular state within one particular utterance. The occupancy of this state drops from 32.2 to 3.0 during cross-training. In fact there is a similarly bad state for S2, but it happens not to occur in the 50 test utterances. This bad state results in a clearly audible pop in the synthesized audio.

The fact such bad states can arise is a weakness in re-using the clustering computed for one acoustic model framework for a second framework.

5.4. MDL for autoregressive clustering

The above results imply that it is not appropriate to directly use the MDL formula (10) for the autoregressive HMM. Since the autoregressive HMM has $k = 5 \times 40$ free parameters per mgc leaf, (10) selects a threshold of around 1200 for each state leading to a model similar to A6, which shows substantial underfitting and has very low naturalness.

Preliminary experiments suggest that (8) is still a good proxy for test set log probability on this data set if we set the mgc clustering threshold as $\xi = \frac{1}{2}\rho k \log N$ with $\rho = 0.3$.

6. Conclusion

We have described autoregressive decision tree clustering and highlighted differences to the standard clustering procedure. Our experimental results suggest that autoregressive clustering improves the naturalness of autoregressive HMM-based speech synthesis. We have seen that the standard *minimum description length (MDL)* criterion for setting model complexity is inappropriate for the autoregressive HMM. We have seen that the optimal model complexity (as measured by the log probability of a held-out test set) is near-optimal in terms of naturalness. Underfitting degrades naturalness noticeably, whereas a large degree of overfitting leads to only a very small decrease in naturalness.

8. 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] M. Shannon and W. Byrne, "Autoregressive HMMs for speech synthesis," in *Proc. Interspeech 2009*, 2009, http://mi.eng.cam.ac.uk/~sms46/papers/shannon2009ahs.pdf.

[9] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn. (E)*, vol. 21, no. 2, pp. 79–86, 2000.

[10] 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, http://mi.eng.cam.ac.uk/~sms46/papers/shannon2009fah.pdf.

[11] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. ARPA Human Language Technology Workshop*, 1994, pp. 307–312.

[12] A. Black and K. Tokuda, "The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets," in *Proc. Interspeech 2005*, 2005.

[13] HTS working group, "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/, accessed 17 April 2009.

[14] 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/.

[15] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. ICASSP 1997*, vol. 2, 1997.

[16] T. Toda and K. Tokuda, "Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis," in *Proc. Interspeech 2005*, 2005.

[17] 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.

[18] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop (Interspeech 2008)*, 2008.