Context Adaptive Training with Factorized Decision Trees for HMM-Based Speech Synthesis

Kai Yu[1], Heiga Zen[2], Francois Mairesse[1], and Steve Young[1]

[1] Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK
[2] Toshiba Research Europe Ltd., Cambridge Research Laboratory, Cambridge, CB4 0GZ, UK

Email: `{ky219,farm2,sjy}@eng.cam.ac.uk, heiga.zen@crl.toshiba.co.uk`

Abstract

To achieve natural high quality synthesised speech in HMM-based speech synthesis, the effective modelling of complex acoustic and linguistic contexts is critical. Traditional approaches use context-dependent HMMs with decision tree based parameter clustering to model the full combination of contexts. However, weak contexts, such as word-level emphasis in neutral speech, are difficult to capture using this approach. To effectively model weak contexts and reduce the data sparsity problem, weak and normal contexts should be treated independently. *Context adaptive training* provides a structured framework for this whereby standard HMMs represent normal contexts and linear transforms represent additional effects of weak contexts. In contrast to speaker adaptive training, separate decision trees have to be built for the weak and normal context factors. This paper describes the general framework of context adaptive training and investigates three concrete forms: MLLR, CMLLR and CAT based systems. Experiments on a word-level emphasis synthesis task show that all context adaptive training approaches can outperform the standard full-context-dependent HMM approach. However, the MLLR based system achieved the best performance.

Index Terms: HMM-based speech synthesis, context adaptive training, factorized decision tree

1. Introduction

Statistical parametric speech synthesis based on hidden Markov models (HMMs) [1] has grown in popularity in recent years. In this framework, the spectrum, excitation, and durations of speech are modelled simultaneously in a unified HMM framework. For a given text sentence to be synthesized, speech parameter trajectories that maximise their output probabilities are generated from the estimated HMMs under consistency constraints between static and dynamic features [2].

It is well known that the spectral and prosodic features of a particular phone in human speech are not only determined by the individual phonetic content, but also heavily affected by various background events associated with the phone, such as neighbouring phones, phone positions, linguistic role of words, etc. The background events which can affect the acoustic realization of a phone are referred to as its *contexts*. Compared to speech recognition, speech synthesis requires a much larger and more complex set of contexts to be represented to achieve high quality synthesized speech. Effective modelling of these complex context dependencies consequently becomes one of the most critical problems for HMM-based speech synthesis. The traditional approach for handling complex contexts is to use a distinct HMM for each particular combination of all possible contexts, referred to as *context-dependent HMM*. The amount of training data is normally not sufficient for robustly estimating all context-dependent HMMs and it is common that the training data does not cover all context combinations required for synthesising new texts.

To address these problems, top-down decision tree based context clustering is normally used [3]. In this approach, states (or streams) of context-dependent HMMs are grouped into "clusters" and the distribution parameters within each cluster are shared. The assignment of HMMs to clusters is performed by examining the context combination of each HMM through a binary decision tree[1], where one context-related yes/no question is associated with each non-leaf node. The decision tree is constructed by sequentially selecting the questions which yield the largest likelihood increase of the training data. The size of the tree is controlled using a pre-determined threshold of likelihood increase or by introducing a model complexity penalty, such as minimum description length [4]. With the use of context questions and parameter sharing, the unseen contexts and data sparsity problems are effectively addressed.

Although context-dependent HMMs with a decision tree-based state clustering technique can effectively model strong contextual effects, it is difficult to model weak contexts, such as word-level emphasis in neutral speech [5], because weak contexts have less influence on the likelihood of data. When pooled with other more influential contexts, they are rarely selected during the decision tree construction. Consequently, the set of final clustered context-dependent HMMs have a poor representation of these contexts. One approach to address this problem is to split the decision tree construction into two stages. In the first stage, a decision tree is constructed using only the weak context questions. In the second stage, the other questions are used to further extend the decision tree [5]. Although this approach can effectively exploit weak context questions, it fragments the training data and leads to a reduction in the amount of the data that can be used in clustering the other contexts. Consequently, the quality of synthesised speech is degraded [5].

In this paper, an alternative approach, *context adaptive training* with factorized decision trees is investigated for weak context modelling. Here, two separate sets of parameters are used to model weak and normal contexts (such as phonetic or position contexts) respectively. Standard HMM parameters are used for normal contexts, while transforms are estimated for weak contexts. Full context-dependent HMMs are

---

[1]In HMM-based synthesis, normally one decision tree is constructed for a particular stream at a particular state position.

then constructed using HMM parameters for normal contexts, transformed by weak-context specific transformations. Standard adaptive training techniques [6–8] can be used to perform interleaved updates of the two sets of model parameters. However, compared to adaptive training for speech recognition, in addition to the structured HMM representation using two sets of parameters, context adaptive training also needs to change the decision tree clustering process due to the nature of contexts being adapted. *Factorized decision trees* are used to fulfil this requirement. The basic idea is to construct independent decision trees for weak and normal contexts and then combine them to construct a structured common decision tree to represent the full context information [5, 9–11]. This paper describes the general framework of context adaptive training and investigates three specific implementations of systems: maximum likelihood linear regression (MLLR), constrained MLLR and cluster adaptive training. The effectiveness of these systems is demonstrated on a word-level emphasis synthesis task.

The rest of the paper is structured as follows. Section 2 describes the framework of context adaptive training with factorized decision trees. Section 3 discusses the three alternative implementations and experimental results are given in section 4, followed by conclusions in section 5.

2. Context adaptive training with factorized decision tree

As discussed in section 1, full context-dependent HMMs can not effectively capture weak contexts and a two-pass decision tree fragments the data remaining for the normal contexts. *Context adaptive training* is proposed here to address these problems.

Adaptive training has been widely used in automatic speech recognition (ASR) to build compact acoustic models on non-homogeneous data [6–8]. A set of transforms are trained to represent different acoustic conditions (or homogeneous blocks), and canonical context-dependent HMMs represent pure speech variabilities. The HMMs for a particular acoustic condition are constructed by adapting the canonical HMMs using the corresponding transforms. The two sets of parameters, transforms and HMMs, are updated in an interleaved fashion. Each update is done holding the other set of parameters fixed. When adaptive training is used in ASR, the acoustic conditions are normally defined for complete data blocks, such as speaker or noise environment. To obtain robust transform estimations, a regression tree is usually constructed to allow a group of Gaussian components to share a transform. Assuming there is only one Gaussian in each clustered state $\mathbf{r_c}$, the adapted Gaussian parameters can be represented as

$$\hat{\mathbf{\Lambda}}_{\mathbf{r_c}} = \mathcal{F}_{\mathbf{r_t}}\big(\mathbf{\Lambda}_{\mathbf{r_c}}\big) \qquad s.t. \qquad \mathbf{r_c} \in \mathbf{r_t} \qquad (1)$$

where $\mathbf{\Lambda}_{\mathbf{r_c}}$ is the Gaussian parameter set of state cluster $\mathbf{r_c}$, $\hat{\mathbf{\Lambda}}$ denotes the adapted parameters, $\mathcal{F}_{\mathbf{r_t}}(\cdot)$ is the transform associated with regression base class $\mathbf{r_t}$. When modelling a range of acoustic conditions using adaptive training, transforms and canonical HMMs always use the same state tying structure. Hence, $\mathbf{r_c}$, as the atomic cluster for adaptation, is always a subset of any transform regression base class $\mathbf{r_t}$.

In contrast to ASR, contexts in HMM-based speech synthesis are much more complex. Though all contexts affect the acoustic realization of phones, different contexts are not always correlated. For example, weak contexts, such as word emphasis, may be independent of phonetic/position contexts because

they are resulting from different underlying phenomena. Therefore, not only the model parameter representation, but also the decision tree clustering process needs to be factorized in context adaptive training.

The idea of using factorized decision trees was used for acoustic modelling in recognition and synthesis [9, 10, 12] but has not been investigated within the framework of adaptive training until recently [5, 11]. In context adaptive training, the main purpose of constructing factorized decision trees is to fully model both sets of contexts and define atomic adaptation units where both normal and weak contexts can apply their effects. To achieve this goal, rather than pooling all context questions together to form a single decision tree, two decision trees are built independently, with normal context questions (e.g. phone and position questions) and weak context questions (in this paper word-level emphasis). They are then combined to form a larger common decision tree by intersecting the leaf nodes as shown in Fig 1.
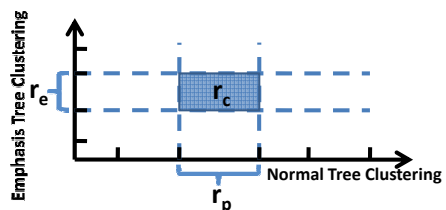

Figure 1: Combination of normal and emphasis decision trees

The leaf nodes $\mathbf{r_e}$ and $\mathbf{r_p}$ are the final state clusters for emphasis and normal decision trees respectively. The leaf nodes of the combined decision tree, $\mathbf{r_c}$ correspond to the intersections of the leaf nodes of the emphasis decision tree $\mathbf{r_e}$ and the normal decision tree $\mathbf{r_p}$. Hence, $\mathbf{r_c}$ are atomic units for adaptation, on which $\mathbf{r_e}$ and $\mathbf{r_p}$ will have effect. Assuming there are N_e and N_p clustered states from the emphasis tree and normal decision tree respectively, the combined decision tree could have $N_e \times N_p$ different context-dependent states. This structured representation is therefore more compact than direct full context-dependent modelling. With this factorized representation, both sets of contexts can make full use of all training data and be extensively explored. The data sparsity problem is effectively addressed without fragmenting the training data.

Once the combined decision tree is constructed, the state output distributions, a single Gaussian in this paper, within $\mathbf{r_c}$ are tied. The Gaussian parameters can then be represented using a structured form similar to adaptive training

$$\hat{\mathbf{\Lambda}}_{\mathbf{r_c}} = \mathcal{F}_{\mathbf{r_e}}\big(\mathbf{\Lambda}_{\mathbf{r_p}}\big) \qquad s.t. \qquad \mathbf{r_c} = \mathbf{r_p} \cap \mathbf{r_e} \qquad (2)$$

where $\mathbf{r_e}$ are the regression base classes, which are equivalent to the leaf nodes of the emphasis decision tree. Compared to equation (1), equation (2) has a different constraint on the regression base class, resulting from factorized decision tree clustering.

In adaptive training, both the transform and the canonical HMM may take various forms. Linear transforms [7, 13] and cluster adaptive training [8] are the two main categories. Equation (2) shows the general form of linear transform based context adaptive training. Cluster-based context adaptive training has the general form as shown in equation (3):

$$\hat{\mathbf{\Lambda}}_{\mathbf{r_c}} = \mathcal{F}_{\mathbf{r_t}}\big(\mathbf{\Lambda}_{\mathbf{r_p}}, \mathbf{\Lambda}_{\mathbf{r_e}}\big) \quad s.t. \quad \mathbf{r_c} = \mathbf{r_p} \cap \mathbf{r_e} \quad \mathbf{r_c} \in \mathbf{r_t} \quad (3)$$

Here, the multiple-cluster Gaussian for state $\mathbf{r_c}$ consists of bases from different context groups $\mathbf{r_p}$ and $\mathbf{r_e}$. Adaptation may then be performed on an additional full context specific regression class $\mathbf{r_t}$ constructed from the atomic state clusters $\mathbf{r_c}$.

3. Implementation

Section 2 describes the general framework of context adaptive training with factorized decision trees. This section will discuss concrete forms of context adaptive training.

The first form uses maximum likelihood linear regression (MLLR) [13] as the weak context transform and standard HMMs for modelling normal contexts [5]. The mean and covariance of the atomic state cluster r_c are represented by

$$\hat{\boldsymbol{\mu}}_{r_c} = \mathbf{A}_{r_e}\boldsymbol{\mu}_{r_p} + \mathbf{b}_{r_e}$$
$$\hat{\boldsymbol{\Sigma}}_{r_c} = \boldsymbol{\Sigma}_{r_p} \tag{4}$$

where r_p and r_e represent the leaf nodes of normal and weak context decision trees respectively. $\mathbf{A}_{r_e}$ and $\mathbf{b}_{r_e}$ are weak-context transform parameters and $\boldsymbol{\mu}_{r_p}$ and $\boldsymbol{\Sigma}_{r_p}$ are Gaussian parameters corresponding to normal contexts. The update formulae of these parameters can be found in [5].

Instead of MLLR, constrained MLLR (CMLLR) [7] can also be used to represent weak contexts. With CMLLR, both mean and covariance are adapted

$$\hat{\boldsymbol{\mu}}_{r_c} = \mathbf{A}_{r_e}\boldsymbol{\mu}_{r_p} + \mathbf{b}_{r_e}$$
$$\hat{\boldsymbol{\Sigma}}_{r_c} = \mathbf{A}_{r_e}\boldsymbol{\Sigma}_{r_p}\mathbf{A}_{r_e}^{\top} \tag{5}$$

One advantage of using CMLLR is that CMLLR can be rewritten as a feature transform. The update formulae for $\boldsymbol{\mu}_{r_p}$ and $\boldsymbol{\Sigma}_{r_p}$ then take the standard form except that the adapted observation must be used. For more details about the update of CMLLR transforms, refer to [7].

The third implementation uses cluster adaptive training [8]. As multiple-cluster HMMs are used here, the general form of equation (3) is used. The adapted Gaussian parameters are represented as

$$\hat{\boldsymbol{\mu}}_{r_c} = \lambda_1\boldsymbol{\mu}_{r_p} + \lambda_2\boldsymbol{\mu}_{r_e}$$
$$\hat{\boldsymbol{\Sigma}}_{r_c} = \boldsymbol{\Sigma}_{r_p} \tag{6}$$

where λ_1 and λ_2 are global weights for interpolating $\boldsymbol{\mu}_{r_p}$ and $\boldsymbol{\mu}_{r_e}$. The update formulae of the interpolation weights can be found in [8]. However, due to the intersection of $\boldsymbol{\mu}_{r_p}$ and $\boldsymbol{\mu}_{r_e}$, the update formulae for multiple-cluster HMMs in [8] must be modified as explained in [11].

The above describes three different models for context adaptive training. The training procedure itself is an iterative process of interleaved update of the two sets of parameters [5, 11].

4. Experiments

4.1. Experimental conditions

The context adaptive training techniques described in Sections 2 and 3 were evaluated in a *natural emphasis* synthesis task described in [5]. The training data is a subset of the male English voice with a Scottish accent (awb) in the CMU ARCTIC speech database. A human judge annotated the 597 utterances of set A of the dataset, by labelling the word(s) that were perceived as the focus of the utterance based on the natural emphasis of the speaker. The emphasis labels were given to the naturally emphasized words (e.g., content words) as well as involuntary fluctuations of the speaker[2].

Altogether four systems were built, three context adaptive training systems as described in section 3 and a standard full context-dependent HMM system which uses both normal and emphasis contexts in state clustering[3]. All systems were built using a modified version of the HMM-based speech synthesis system (HTS) [14]. The HMM-GTD technique for $\log F_0$ modelling [15] was used as it yielded better speech quality. Six emphasis contexts were used to form questions for emphasis decision tree construction[4]. The static feature set comprised 25 mel-cepstral coefficients [16] including the zero-th coefficients, $\log F_0$ and aperiodic energy components in five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 kHz). All features were extracted using STRAIGHT [17]. The MLLR and the CMLLR systems used block diagonal transforms for the spectrum and full transforms for $\log F_0$ and aperiodic component features. The CAT system used fixed global weights (1.0) to interpolate mean vectors of normal and emphasis contexts. In this paper, context adaptive training techniques were only applied to the spectrum, $\log F_0$ and aperiodic components, while duration was still modelled using standard full context-dependent HMMs. The speech parameter generation algorithm considering global variance [18] was used during synthesis.

4.2. Experimental results

A subjective listening test was performed to measure the ability to convey emphasis. For each system, 10 utterances in the tourist information domain were generated without any emphasized word. The same utterances were generated again but with *one* word emphasized, forming 10 contrast pairs (e.g., 'Char Sue is an *expensive* Chinese restaurant'). The contrasting waveform pairs from the four systems were then provided to listeners. When perceiving a difference of emphasis, the listener was asked to select the word that carried the emphasis, otherwise to indicate that there is no perceivable emphasis . Altogether 14 listeners, 7 native and 7 non-native, participated in the test. The performance of emphasized word detection is shown in table 1.

System	# Det	Rec (%)	Pre (%)	F-measure
GMM	2.0	20.0	53.8	0.29
MLLR	4.7	47.1	68.0	0.56
CMLLR	3.2	32.1	68.2	0.44
CAT	2.9	28.6	69.0	0.40

Table 1: Average number of correctly detected emphasised words and Recall, Precision and F-measure of emphasis detection.

The row labelled GMM in Table 1 is the standard full context-dependent system. It can be observed that all context adaptive training systems obtained better emphasis detection performance than the standard full context-dependent HMMs. A pair-wise two-tailed Student's t-test was performed to evaluate the statistical difference of the average number of correctly detected emphasized words. It was found that the improvements of context adaptive training systems from the standard system were all significant at the 95% confidence level. Amongst the different forms of context adaptive training, the MLLR system

---

[2]Available at http://mi.eng.cam.ac.uk/~farm2/emphasis. It is worth noting that the emphasis labels are rough due to the often vague acoustic cues, as shown in a labelling agreement check in [5].

[3]The purpose of this experiment is to compare different approaches for complex contexts modelling. Though emphasis word adaptation can be an alternative approach for emphasis synthesis as shown in [5], it is not a generally applicable approach for modelling complex contexts, such as phone positions. Hence, it is not considered in this paper.

[4]Each emphasis related question consists of one emphasis context feature and one normal context feature. This will lead to powerful transforms as the number of transforms is large.

achieved the best performance while the CMLLR and CAT systems were similar.

Model complexity may be one reason for the performance difference between context adaptive training systems. Emphasis is mainly carried by $\log F_0$ features and the number of $\log F_0$ states in the factorized decision trees are shown in Table 2[5].

System	#$\mathtt{r_p}$	#$\mathtt{r_e}$	#$\mathtt{r_c} = \mathtt{r_p} \cap \mathtt{r_e}$	#para ($\times 10^3$)
MLLR	2445	2988	17739	48.7
CMLLR	1959	2458	13361	40.1
CAT	2402	2894	16307	23.1

Table 2: Number of $\log F_0$ states in factorized decision trees.

In table 2, $\mathtt{r_p}$ and $\mathtt{r_e}$ are the numbers of clustered states in the normal and emphasis decision trees of the $\log F_0$ stream, respectively. $\mathtt{r_c}$ is the set of leaf nodes of the intersection of the two trees, which is the number of atomic units for adaptation. The last column gives the total number of free parameters, consisting of both HMM and transform parameters. It can be observed that the CAT system has far fewer parameters than the MLLR system. This is mainly because global weights are used and there is only one vector associated with each emphasis state $\mathtt{r_e}$. This significant reduction of parameters may then limit its power to transform normal contexts to full contexts. In contrast, the CMLLR system has powerful emphasis transforms, but less powerful HMMs (small #$\mathtt{r_p}$) for normal contexts. However, this may not be the main reason for the performance degradation compared to the MLLR system. During training, there are two decision tree based state clustering stages. The first one uses roughly estimated standard HMMs as the base model for clustering, while the second one uses further refined adapted HMMs. For the MLLR and the CAT systems, the adapted parameters were explicitly saved and can then be safely used for state clustering [6]. On the contrary, for the CMLLR system, due to the current implementation of using CMLLR as feature transforms, the adapted model parameters were not explicitly saved. Therefore, in the second state clustering stage, the base model parameters were HMMs for normal contexts rather than full contexts, which will consequently affect the quality of state clustering. It was found that after the second state clustering stage, the MLLR and the CAT systems both received an increased number of clustered states, while the CMLLR system received a decrease. This is probably the main reason for the performance difference between the CMLLR and the MLLR system. The effect of state clustering in context adaptive training systems will be investigated in future work.

5. Conclusions

This paper has described a context adaptive training framework to model weak contexts in HMM-based speech synthesis. Two sets of parameters are constructed to represent the different context groups and are estimated inter-dependently. In contrast to adaptive training for speech recognition, decision tree clustering must be modified for context adaptive training. The factorized decision tree approach is used here, where two independent decision trees are built for normal and weak contexts respectively.

----

[5]The complexity difference of spectrum and aperiodic component features is small, hence not shown here.

[6]Strictly speaking, for adaptive HMMs, the state clustering should also follow an adaptive fashion. Using adapted parameters and a standard state clustering technique in [3] is just an approximation.

Context adaptation is then performed for the intersections of the two trees. Three forms of context adaptive training systems, MLLR, CMLLR and CAT, are investigated in this paper. Experiments on a word-level emphasis synthesis task shows that context adaptive training significantly outperformed standard full context HMMs with the MLLR system showing the best overall performance.

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

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

[4] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL principle for speech recognition," in *Proc. EUROSPEECH*, 1997, pp. 99–102.

[5] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *Proc. ICASSP*, 2010, pp. 4238–4241.

[6] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[7] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[8] ——, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[9] N. Iwahashi and Y. Sagisaka, "Statistical modelling of speech segment duration by constrained tree regression," *Trans. of IEICE*, vol. E83-D, pp. 1550–1559, 2000.

[10] Y. Nankaku, K. Nakamura, H. Zen, and K. Tokuda, "Acoustic modeling with contextual additive structure for HMM-based speech recognition," in *Proc. ICASSP*, 2008, pp. 4469–4472.

[11] H. Zen and N. Braunschweiler, "Context-dependent additive $\log F_0$ model for HMM-based speech synthesis," in *Proc. of Interspeech*, 2009, pp. 2091–2094.

[12] K. Saino, "A clustering technique for factor analyzed voice models," Master thesis, Nagoya Institute of Technology, 2008, (in Japanese).

[13] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, pp. 171–185, 1995.

[14] "HMM-based Speech Synthesis System (HTS)," http://hts.sp.nitech.ac.jp.

[15] K. Yu, T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson, and S. Young, "Probablistic modelling of F0 in unvoiced regions in HMM based speech synthesis," in *Proc. ICASSP*, 2009.

[16] T. Fukada, K. Tokuda, K. T., and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP*, 1992, pp. 137–140.

[17] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[18] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.