FACTORIZED CONTEXT MODELLING FOR TEXT-TO-SPEECH SYNTHESIS

Heng Lu, Simon King

The Centre for Speech Technology Research, The University of Edinburgh, UK

hlu2@inf.ed.ac.uk, Simon.King@ed.ac.uk

ABSTRACT

Because speech units are so context-dependent, a large number of linguistic context features are generally used by HMM-based Text-to-Speech (TTS) speech synthesis systems, via context-dependent models. Since it is impossible to train separate models for every context, decision trees are used to discover the most important combinations of features that should be modelled. The task of the decision tree is very hard - to generalize from a very small observed part of the context feature space to the rest - and they have a major weakness: they cannot directly take advantage of factorial properties: they subdivide the model space based on one feature at a time. We propose a Dynamic Bayesian Network (DBN) based Mixed Memory Markov Model (MMMM) to provide factorization of the context space. The results of a listening test are provided as evidence that the model successfully learns the factorial nature of this space.

Index Terms— Text-To-Speech synthesis, Dynamic Bayesian Network, Mixed Memory Markov Model, factorized model, maximum likelihood parameter generation

1. INTRODUCTION

1.1. HMM-based statistical parametric models

The HMM-based statistical parametric speech synthesis method is now well-established [1, 2, 3, 4]. Compared with unit-selection and concatenation [5, 6], the HMM-based statistical parametric method has the advantages of small footprint, high intelligibility and more flexibility to transform the models (notably, for speaker adaptation). However, the naturalness obtained by the statistical parametric method is generally worse that for methods that concatenate recorded speech units, provided the training corpus is large enough and of good quality. One of the major weak points in HMM-based synthesis is the inaccuracy of its acoustic model. In particular, because the context-dependent state space is enormous, due to the richness of context information, model clustering (commonly in combination with Minimum Description Length (MDL) complexity control [7]) is necessary to avoid over-fitting and to create models for unseen contexts. This complexity control is non-trivial task and, although the MDL criterion itself might be based on information-theoretic principles, in practice, expert manual tuning of the MDL factor is performed to adjust the complexity of the clustered model and get the best performance.

1.2. Alternative statistical parametric models

In order to deal with the rich context-dependent model space, some attempts were made in statistical parametric models. Cluster Adaptive Training (CAT) was originally developed for speech recognition to enable rapid speaker adaptation [8]. And in [9] CAT has been extended for statistical parametric synthesis to perform the speaker and language factorization. The CAT model consists of cluster of models and transformation is employed to represent the specific target model.

Dynamic Bayesian Networks (DBNs) [10], a family of models of which the familiar HMM is one of the simplest members, offer a useful framework for representing structure in a set of variables. DBNs represent random variables as nodes, with dependencies between variables being represented by arcs between pairs of nodes, and missing arcs indicating conditional independence. DBNs are a type of Bayesian Network that include directed edges pointing in the direction of time.

Whilst the dependency structure between variables can in principle be learned automatically from data – as in our previous work where BN structure learning was applied to the problems of predicting phone duration [11] and of finding the most relevant context features for HMM-based speech synthesis [12] – it is more common to design the network structure by hand using expert intuition. This is the approach taken here.

In [13], Markov models whose state spaces arise from the Cartesian product of two or more discrete random variables are presented. The authors propose the Mixed Memory Markov Model to parameterize the transition matrices of these models as a convex combination or mixture of simpler dynamical models. In this paper, within the general framework of Dynamic Bayesian Networks, we propose to use a Mixed Memory Markov Model to realize a factorization of the linguistic context features in HMM-based speech synthesis. We implement this model using the Graphical Models Toolkit (GMTK) [14, 15].

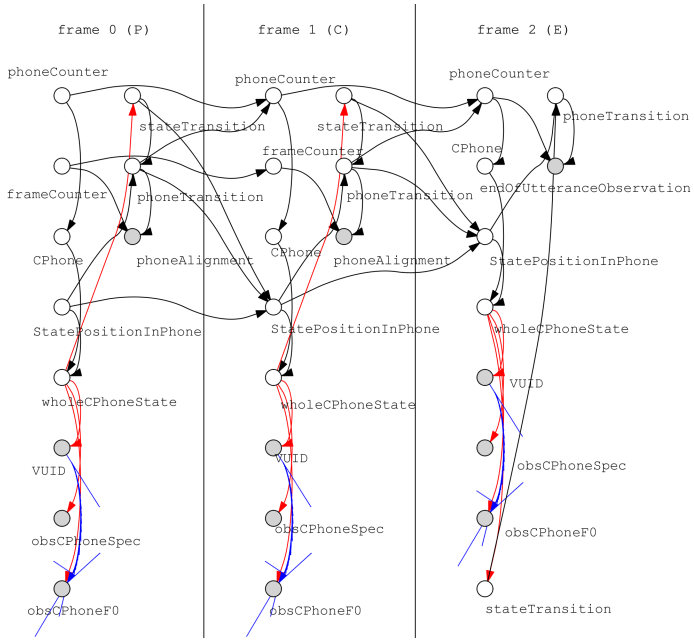

Fig. 1. Simple DBN structure for phone-based TTS (figure produced by the GMTK toolkit gmtkViz)

2. A DYNAMIC BAYESIAN NETWORK FACTORIZED MODEL FOR SPEECH SYNTHESIS

2.1. Dynamic Bayesian Networks

Let $U = \{x_1, ..., x_n\}, n > 1$ be a set of variables. A Bayesian network B over a set of variables U is a network structure B_S, which is a directed acyclic graph (DAG) over U and a set of probability tables $B_P = \{p(u|pa(u))|u \in U\}$ where $pa(u)$ is the set of parents of u in B_S. A Bayesian network represents the factorisation of the joint probability distribution $P(U) = \prod_{u \in U} p(u|pa(u))$. Since speech is a time signal, model for speech parameters will in fact be a Bayesian network that dynamically 'unrolls' to fit the observation sequences. This type of BN is known as a dynamic Bayesian network, and consists of instances of a Bayesian network repeated over time, with additional arcs added to join variables at differing times.

For background material on DBNs, we refer the reader to [16, 17, 18, 19]. By way of illustration, Figure 1 shows the structure for a simple phone-based model for speech synthesis. This figure is plotted by the GMTK toolkit gmtkViz. The structure is just a conventional phone-based HMM, but with all variables drawn out explicitly as nodes in the network: that it, the DBN shows each variable in the statistical model as a node and the relations between variables as arcs. In Figure 1, black arcs indicate that a parent and child have a direct relationship, a red arc means the child node variable is observed, and a blue arc means a switching parent. The acoustic observations for spectral and F0 features depend directly on the state of the current phone. F0 observations also depend on the Voice/Unvoiced indication variable (VUID) as a switching parent; this structure is similar to the MSD-HMM [2]. To fit a given training observation sequence, the middle frame (C = 'chunk') is unrolled and an EM algorithm is used to learn the model parameters.

2.2. Mixed Memory Markov Model

The Mixed Memory Markov Model [13, 20, 21] is an attractive method to deal with the key problem in speech synthesis case, which is that the context-dependent HMM is a model whose state space is equal in size to the product of the cardinalities of every category of context factor – that is, very large indeed!

For simplicity, suppose we have 2 context features, A and B. A has cardinality n (i.e., it can take on n different values) and B has cardinality j. In this case, the state space of the context-dependent model has size $j \times n$. To performance inference with this model, we could write:

$$P(\boldsymbol{O}|A, B, \lambda) = \sum_{\boldsymbol{m} \in \boldsymbol{M}} P(\boldsymbol{O}|A, B, \lambda_{\boldsymbol{m}}, \boldsymbol{m}) P(\boldsymbol{m}) \quad (1)$$

where $\boldsymbol{O}$ is the observation vector, λ contains the parameters of the statistical model, and $\boldsymbol{M} = [\boldsymbol{m}_1, \boldsymbol{m}_2, ..., \boldsymbol{m}_K]$ is a latent variable which can be regarded as a mixture component. K is the number of mixture components for the MMMM context dependent model, which we can set to any small integer.

Now, we make an assumption that, when the value of $\boldsymbol{m}$ belongs to a subset of the mixture components related to context A, say $\boldsymbol{m} \in \boldsymbol{M}_A$, then $P(\boldsymbol{O}|A, B, \boldsymbol{m})$ is independent of B. Likewise, we assume that when the value of $\boldsymbol{m}$ belongs to a subset of the mixture components related to context B, say $\boldsymbol{m} \in \boldsymbol{M}_B$, then $P(\boldsymbol{O}|A, B, m)$ is independent of variable A.

Further, enforcing the conditions $\boldsymbol{M}_A \cap \boldsymbol{M}_B = \emptyset$ and $\boldsymbol{M}_A \cup \boldsymbol{M}_B = \boldsymbol{M}$ allows us to rewrite Equation 1 as

$$\begin{aligned} P(\boldsymbol{O}|A, B, \lambda) &= \sum_{\boldsymbol{m} \in \boldsymbol{M}_A} P(\boldsymbol{O}|A, \lambda_{\boldsymbol{m}}, \boldsymbol{m}) P(\boldsymbol{m}|A, B) \\ &+ \sum_{\boldsymbol{m} \in \boldsymbol{M}_B} P(\boldsymbol{O}|B, \lambda_{\boldsymbol{m}}, \boldsymbol{m}) P(\boldsymbol{m}|A, B) \end{aligned} \quad (2)$$

where the term $P(\boldsymbol{m}|A, B)$ is a conditional probability. This is implemented in the model as a Conditional Probability Table (CPT) with two parents: context A and context B, with the child node indicating the mixture component index m at frame t. The EM algorithm is employed to learn a total of K models plus the conditional probability table holding the distribution $P(\boldsymbol{m}|A, B)$. We can see that, compared to a conventional context-dependent model which would use $j \times n$ models, we only need to train K models. For $P(\boldsymbol{m}|A, B)$

CPT, it is a sparse matrix with only small number of none zero values, but we still need to estimate $j \times n$ set of $P(\boldsymbol{m}|A, B)$.

2.3. Parameter generation

After obtaining the parameters for all the Gaussians and the conditional probability table from the DBN-based MMMM, we use the speech parameter generation algorithm method proposed by Tokuda [1] to generate smooth parameter trajectories to drive a vocoder. Given context factor sequence $\boldsymbol{A} = [a_1, a_2, ..., a_T]$ and $\boldsymbol{B} = [b_1, b_2, ..., b_T]$, we can simply look up the $P(\boldsymbol{m}|A, B)$ in the learned conditional probability table. This gives us the weights on a set of Gaussians. Note that in our implementation, the conditional probability table for $P(\boldsymbol{m}|A, B)$ is sparse: for each combination of A and B, the weights on only a few Gaussians are non-zero.

Currently, we then approximate a single Gaussian from the resulting mixture for each state in the sequence, and use this sequence of Gaussians with MLPG to generate a trajectory of observations, ready to pass to the vocoder. Future work will include more exact ways of treating the mixture distribution in MLPG itself.

3. EXPERIMENTS

3.1. Context factors

In order to confirm that our factorized model is behaving as expected, we need to select some suitable context factors A and B. We performed an experiment in which A was the clustered combination of 11 context features selected from amongst the usual full-context features, using the method in [12]. The cardinality of A was 2817 for spectral feature and 2986 for F0 feature. Factor B was an additional factor called 'emphasis', with a cardinality of 4: possible values range from -1 to 2 (de-emphasized to strong emphasis).

3.2. Database

We used speech recorded from a British English male speaker known as 'roger'[1]. For this speaker, we have data in which the speaker deliberately emphasised certain words according to simple markup in the text, plus additional standard read-text data of arctic sentences [22] and newspaper sentences. The corpus is described in [23]. We used 1631 utterances from the emphasis portion, 1132 utterances from the Arctic portion, and 925 utterances from newspaper sentences to compose a set with a total of 3688 training utterances. Emphasis is labeled by appending an additional feature to the usual HTS-labels

[1]Some data from this speaker was used in the Blizzard Challenge 2008,2009 and 2010

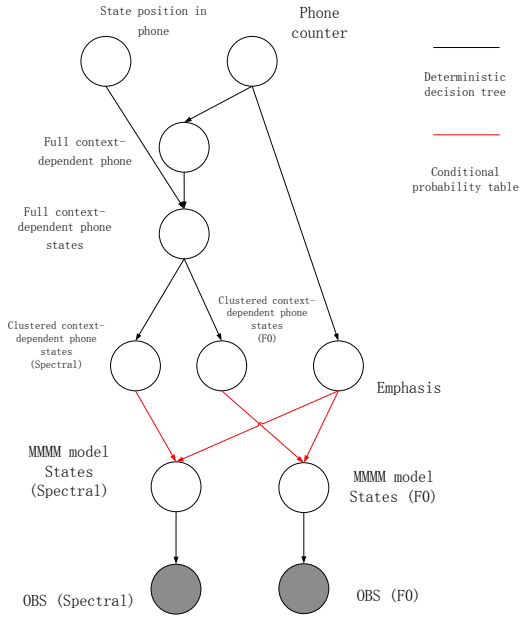

Fig. 2. The core part of the model structure used in our experiments

3.3. Conditional probability table and MMMM model initialization

The DBN-based MMMM model structure shown in Figure 2 was designed. In the figure, the context (not including emphasis) dependent phones are clustered using the gmtkTie tool, which is similar to the decision tree clustering in conventional HMM based method (although it does not use MDL). In this model, the factorization is simply into two factors: A is context (not including emphasis), and B is emphasis. Clustering and parameter tying is still used within factor A to control complexity. The clustered context variable and the emphasis variable are the two parents of the MMMM model state. Note that the model has parallel structures for spectral parameters and F0 respectively. A conditional probability table holds the distribution of the MMMM model state given its two parents.

For a total of 3688 training utterances with sampling rate 48KHz, there are 49696 types of full context-dependent phones; we are using 5 state per phone, so the cardinality of the (unclustered) full context-dependent phone variable (i.e., state) is $49696 \times 5 = 248480$. 59 order Mel-cepstral and F0 feature as well as their delta and delta delta are extracted for each utterance. After clustering, the cardinality of the clustered context-dependent variable is 2817 for the spectral state, and 2986 for the F0 state. We initialize the MMMM mixture size $K = 2817 + (5 * 10) = 2867$ (Spectral) and $K = 2986 + (5 * 10) = 3036$ (F0) which allows for 1 mixture component per clustered context-dependent phone plus 10 components for each value of the emphasis variable (4 + the null value). For comparison, in a conventional full context-dependent model, to realize the 5 level emphasis for

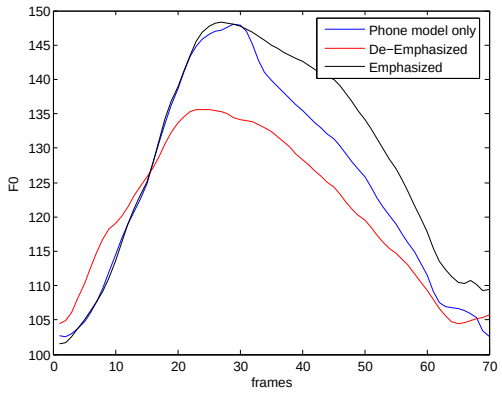

Fig. 3. F0 contour for "...RIO..." in the utterance "No, it was RIODAN who did it!" where uppercasing indicates emphasis.

each clustered model, one would need $2817 * 5$ models. The CPT size in the MMMM is $2817 \times 5 \times 11$ for spectral feature, and $2986 \times 5 \times 11$ for F0. A previously-clustered context-dependent phone model and a monophone emphasis model are used to initialize the parameters for the MMMM model.

3.4. Subjective Listening Test

Figure 3 is the F0 contour generated from the DBN-MMMM with the emphasis variable set to 'de-emphasized' (emphasis context label = -1) and set to 'strong emphasis' (emphasis context label = 2), alongside a contour generated from the DBN-MMMM using the emphasis-independent context-dependent phone mixture only. We change the Emphasis context label to -1 (de-emphasized) and 2 (strong emphasized) and to synthesize using the one Emphasis independent context phone mixture only. Every time, the same one Emphasis independent context phone mixture is chosen for the above three cases, but with different emphasis mixture and weight(conditional probability) given different Emphasis context. From the plots of F0 contour shown in Fig. 3, we can see clearly the effectiveness of proposed method.

Since the current model does not predict duration, HTS-generated phone durations were used when synthesising the materials for the subjective listening tests. Two listening tests were conducted. One tested the ability of the system to generate perceivable emphasis. The carrier sentences had pattern such as :"It was ELIZA, not Erwin!" where one or other of the two proper names was emphasised (strong emphasis) and the other was de-emphasised. 12 native English listeners took part in the test. For each of 30 presented utterances, they were asked to chose the name they thought had been emphasized. The accuracy for emphasis recognition was 62.78% (chance level = 50%).

The second subjective listening test was an A/B forced choice naturalness preference test. For each pair of presented utterances, participants were asked to choose the synthetic utterance that sounded most natural. One utterance in each pair was synthesised with strong emphasis in capital words part and one without. The sentences had patterns such as "It was ERWIN who did it!", "No, it was ELIZA who did it", "It was ELIZA, not ERWIN". 12 native English listener (the same people as the previous test) also took part in this test, and the result was a $55.83\% : 45.17\%$ preference for the speech synthesised with emphasis ,as shown in Figure 4. The preference was relatively small (as would be expected: emphasis is only one contributing factor to overal naturalness judgements), but significant ($p = 0.0059$).

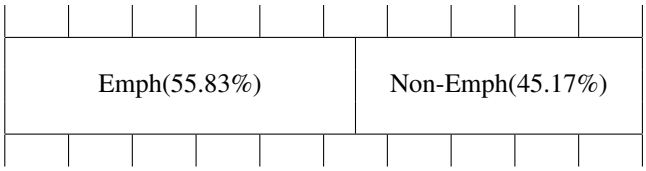

Fig. 4. Results of the forced-choice subjective preference listening test.

4. CONCLUSION

In this work, we proposed a DBN-MMMM model which to factorizes the context features used in speech synthesis. Compared with conventional HMM-based speech synthesis system using MDL based decision tree clustering method to reduce the context model space, our proposed DBN based MMMM method factorized context-dependent models into a series of mixture models that dependent only on one context information. Objective examination of the resulting F0 contours (Figure 3) and two subjective listening tests demonstrate the effectiveness of the method.

5. FUTURE WORK

Our future plan is to build a fully factorized model for context-dependent speech synthesis system, not only for the emphasis context but for other factors too, including the standard ones that in the current experiment were bundled together into factor A.

6. ACKNOWLEDGEMENTS

The research leading to these results was funded from EPSRC grant EP/I031022/1 (Natural Speech Technology).

7. REFERENCES

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

[2] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. and Syst.*, vol. E85-D, pp. 455–464, 2002.

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

[4] "HTS," http://hts.sp.nitech.ac.jp/.

[5] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *In Proc. ICASSP*, vol. 1, pp. 373–376, 1996.

[6] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[7] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Oxford University Press*, vol. 21, pp. 79–86, 2000.

[8] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Transactions on Audio Speech and Language Processing*, vol. 8, pp. 417C428, 2000.

[9] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, pp. 1713–1724, 2012.

[10] T. Dean and K. Kanazawa, "Probabilistic temporal reasoning," *AAAI*, p. 524C528, 1988.

[11] O. Goubanova and S. King, "Bayesian networks for phone duration prediction," *Speech Communication*, vol. 50, pp. 301–311, 2008.

[12] H. Lu and S. King, "Using bayesian networks to find relevant context features for hmm-based speech synthesis," *Proc Interspeech 2012*, 2012.

[13] L. K. Saul and M. I. Jordan, "Mixed memory Markov models: Decomposing complex stochastic processes as mixtures of simpler ones," *Machine Learning*, vol. 37, 1999.

[14] "GMTK," http://ssli.ee.washington.edu/ bilmes/gmtk/.

[15] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," *Proc. of ICASSP*, vol. 4, pp. 3916–3919, 2002, May.

[16] Livescu, Karen, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic bayesian networks," *8th European Conference on Speech Communication and Technology (Eurospeech)*, 2003.

[17] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic bayesian networks," *Computer Speech and Language*, vol. 21(4), pp. 620–640, 2007.

[18] K. P. Murphy, "Dynamic bayesian networks: representation, inference and learning," *Doctoral dissertation, University of California*, 2002.

[19] F. V. Jensen, "An introduction to bayesian networks," *UCL press*, vol. 74, 1996.

[20] T. Choudhury and S. Basu, "Modeling conversational dynamics as a mixed memory Markov process," *In Proc. NIPS*, 2004, December.

[21] K. Kirchhoff, S. Parandekar, and J. Bilmes, "Mixed-memory Markov models for automatic language identification," *Proc. of ICASSP*, vol. 1, pp. 761–764, 2002, May.

[22] J. Kominek and A. Black, "The CMU ARCTIC speech databases," in *SSW5, Pittsburgh, PA.*, 2004.

[23] V. Strom et al, "Modelling prominence and emphasis improves unit-selection synthesis," *in Proc. Interspeech*, 2007.