Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models

Matthew Gibson

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, U.K.

mg366@cam.ac.uk

Abstract

Hidden Markov model (HMM) -based speech synthesis systems possess several advantages over concatenative synthesis systems. One such advantage is the relative ease with which HMM-based systems are adapted to speakers not present in the training dataset. Speaker adaptation methods used in the field of HMM-based automatic speech recognition (ASR) are adopted for this task. In the case of unsupervised speaker adaptation, previous work has used a supplementary set of acoustic models to firstly estimate the transcription of the adaptation data. By defining a mapping between HMM-based synthesis models and ASR-style models, this paper introduces an approach to the unsupervised speaker adaptation task for HMM-based speech synthesis models which avoids the need for supplementary acoustic models. Further, this enables unsupervised adaptation of HMM-based speech synthesis models without the need to perform linguistic analysis of the estimated transcription of the adaptation data.

Index Terms: speech synthesis, unsupervised speaker adaptation

1. Introduction

Hidden Markov model-based systems have delivered synthetic speech of comparable quality to that of concatenative (or unit selection) synthesis systems [1]. Additionally, HMM-based systems possess several advantages over unit selection systems. These advantages include simple ways to interpolate between speakers and synthesise speech of varying styles or emotions [2, 3] . Perhaps the most significant advantage is the ability to adapt to new speakers using a relatively small amount of adaptation data [4].

Most research into speaker adaptation for HMM-based speech synthesis (or text-to-speech, TTS) has focussed upon the supervised scenario, where transcribed adaptation data is available. More recent work has tackled the challenge of adaptation of HMM-based synthesis models using unlabelled adaptation data [5]. As will be explained in due course, unsupervised adaptation of HMM-based synthesis models is problematic for two reasons. Firstly, the modelling of suprasegmental contextual information renders the synthesis models unsuitable for ASR purposes. Therefore a supplementary set of e.g. triphone acoustic models are typically used to estimate a transcription of the unlabelled adaptation data [5]. Secondly, linguistic analysis is required to transform word-level transcriptions into transcriptions containing suprasegmental contextual information. In the case of unsupervised adaptation, it is feasible that the linguistic analysis step exacerbates errors present in the estimated word-level transcription.

This paper presents an alternative to the unsupervised adaptation approach described in [5]. A two-stage decision tree construction method is introduced, which enables a single set of acoustic model components to be used for both ASR and TTS. This method is then used to circumvent the need for supplementary ASR acoustic models and linguistic analysis of estimated transcriptions during unsupervised adaptation. The goal of this work is to demonstrate two results, numbered below.

1. In the case of unsupervised adaptation, performance can be achieved which is indistinguishable to that yielded by fully supervised adaptation without the need for
   (a) supplementary ASR acoustic models or
   (b) linguistic analysis of estimated transcriptions.
2. Two-stage decision tree clustering does not compromise the quality of the resulting speech synthesis models.

The paper is structured as follows. Section 2 provides a brief introduction to HMM-based speech synthesis models and explains why unsupervised adaptation of such models is problematic. Section 3 explains the two-pass decision tree construction technique, and how this enables unsupervised adaptation of HMM-based synthesis models. Sections 4 and 5 respectively describe the experimental setup and results. Lastly, Section 6 summarises the contributions of this work.

2. Unsupervised adaptation of speech synthesis models

In the domain of ASR, unsupervised adaptation is usually conducted by firstly estimating a transcription of the adaptation data using a speech recogniser. This speech recogniser often deploys the same models which are subsequently adapted, thus avoiding the need for multiple sets of acoustic models.

In the domain of HMM-based synthesis, use of the same unsupervised adaptation framework is problematic. This is because the context-dependent acoustic models typically used in state-of the art HMM-based speech synthesis [6] are unsuitable for ASR. These contexts, henceforth referred to as full contexts, are based on segmental (e.g. context-sensitive phoneme) and suprasegmental (e.g. stress, total number of syllables in utterance) information. As is the case for ASR acoustic modelling, decision tree clustering of the full contexts is used to enable robust estimation of the model parameters.

Although theoretically possible to recognise unlabelled data using full context models, the presence of suprasegmental contextual information adds a prohibitive amount of complexity to the construction of recognition networks. The approach described in Section 3 avoids this complexity without introducing the need for a supplementary set of acoustic models to estimate the transcription of unlabelled adaptation data.

3. Two-pass decision tree construction

Multiple-component triphone mixture models may be derived from single-component full context models by imposing constraints upon the decision tree structure when constructing full context models. This constrained decision tree construction process is illustrated in Figure 1.

The first stage, indicated as Pass 1 in Figure 1, uses only questions relating to left, right and central phonemes to construct a phonetic decision tree. This decision tree is used to generate a set of tied triphone contexts, which are easily integrated into the ASR search. No state output distributions are estimated at this stage.

Pass 2 extends the decision tree constructed in Pass 1 by introducing additional questions relating to suprasegmental information. The output of Pass 2 is an extended decision tree which defines a set of tied full contexts. Each leaf node of the extended decision tree has an ancestor node which coincides with a leaf node of the Pass 1 decision tree. This is called the 'triphone ancestor'.

After this two-pass decision tree construction, single component Gaussian state output distributions are estimated to model the tied full contexts associated with each leaf node of the extended decision tree. These models are easily integrated into TTS synthesisers.

A mapping from the single-component full context models to multiple-component triphone models is defined as follows. Each set of Gaussian components associated with the same 'triphone ancestor' are grouped as components of a multiple component mixture distribution to model the triphone context defined by the 'triphone ancestor'. The derived triphone models are illustrated at the bottom of Figure 1. The mixture weight c_m of a mixture component m is calculated from the occupancies associated with components of the Pass 2 leaf node contexts as described by Equation 1. The sum in the denominator is over each component associated with the same 'triphone ancestor' and the symbol γ_k is the occupancy of component k.

$$c_m = \frac{\gamma_m}{\sum_k \gamma_k} \tag{1}$$

The inverse mapping from triphone models to full context models is obtained by associating each Gaussian component with its original full context. This is achieved by associating a unique full context identifier to each component as illustrated in Figure 1. Given this mapping between full context and triphone models, unsupervised adaptation of full context acoustic models is straightforward.

3.1. Unsupervised adaptation

As illustrated in Figure 2, triphone models derived from estimated full context models (as described in Section 3) are used to transcribe unlabelled adaptation data. Once word and triphone-level transcriptions of the adaptation data are available, the full context models may be adapted in two different ways.

The first adaptation method, labelled as 'Triphone adaptation' in Figure 2, uses the estimated triphone-level transcription to adapt the triphone models. The adapted triphone models are then mapped back to full context models using the inverse mapping described in Section 3.

The second adaptation method, labelled as 'Full adaptation' in Figure 2, firstly analyses the estimated word-level transcription to produce an estimated full context labelling of the adaptation data. The full context models are then adapted directly using this labelling.

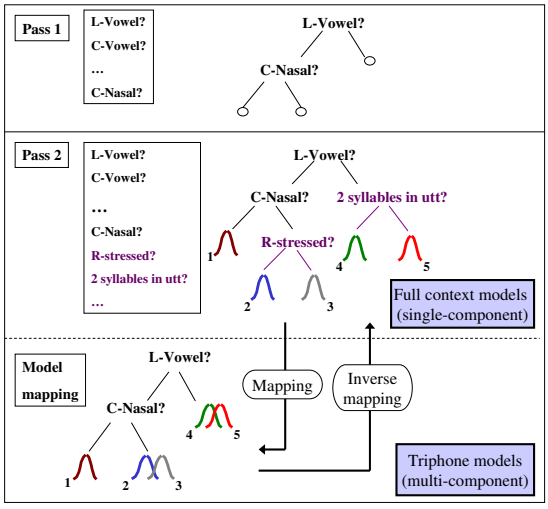

Figure 1: Two-pass decision tree construction. Mapping functions permit sharing of full context models for TTS and triphone models for ASR.

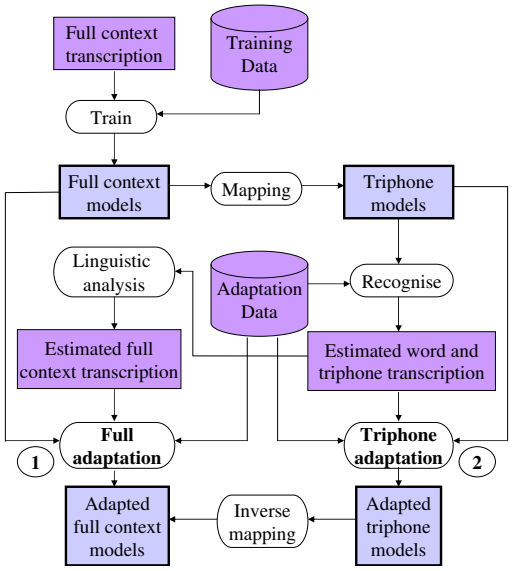

Figure 2: Unsupervised adaptation of full context models via (1) full adaptation or (2) triphone adaptation.

Note that linguistic analysis may exacerbate errors present in the estimated word-level transcription. It is therefore feasible that the triphone adaptation technique is more robust than full context adaptation in the unsupervised case. This hypothesis is tested in the experimental sections which follow.

4. Experiments

Full context average voice models are estimated using the Wall Street Journal (WSJ) SI84 dataset and maximum likeli-

hood estimation. The acoustic features comprise the following: STRAIGHT-analysed Mel-cepstral coefficients [7] (40 dimensions), fundamental frequency ($F0$), and measurements which quantify the aperiodicity of the speech (5 dimensions). The first and second order temporal derivatives of all of these coefficients are appended to yield a feature vector of dimension 138. The feature vector is split into three streams: cepstral coefficients, $F0$ and the aperiodicity measures. Multi-space probability distributions are used to model observations of varying dimension, namely the $F0$ observation [8]. Explicit duration models are integrated (hidden semi-Markov models) to improve the quality of synthesised speech [9]. One decision tree per state and stream combination is used, with an additional decision tree to cluster contexts of the duration model. A speech utterance is generated from models via feature sequence generation [10] and resynthesis of a waveform from the feature sequence [7].

Average voice models corresponding to standard, unconstrained decision tree construction (system A of Figure 3) are estimated for comparison with those corresponding to two-pass decision tree construction (system B).

Adapted systems are derived from System B using either the triphone or full adaptation method described in Section 3.1. Constrained maximum likelihood linear regression adaptation is used, and the adaptation data corresponds to spoke 4 of the 1993 ARPA evaluation (speaker 440M). The adaptation techniques are evaluated using two different volumes of adaptation data (10 and 40 utterances), and in the supervised and unsupervised cases, resulting in eight adapted model sets corresponding to systems C through J in Figure 3. System K corresponds to vocoded natural speech, analysed and resynthesised using the STRAIGHT technique [7]. The synthesised test utterances are a subset of the 1992 ARPA speaker independent read 5k test dataset with no verbal pronunciation.

In the case of unsupervised adaptation, triphone models derived from the estimated full context average voice models are used for the recognition step, in conjunction with the closed vocabulary 20k bigram language model provided with the WSJ0 corpus. A set of state transition probabilities are estimated from the SI84 dataset for use with the triphone models during recognition. A phoneme error rate of 47.1% (word error rate 72.5%) is observed for the unsupervised transcriptions.

A total of eleven systems, A through K in Figure 3, were evaluated by listening to synthesised utterances via a web browser interface closely resembling that used in the Blizzard Challenge 2007. The evaluation comprised two sections. In the first section, listeners judged the naturalness of an initial set of synthesised utterances. In the second section, listeners judged the similarity of a second set of synthesised utterances to a target speaker's (speaker 440M) speech. Four of the target speaker's natural utterances were available for comparison. No utterances from the initial set were present in the second set. Each synthetic utterance was judged using a five point Likert-type psychometric response scale [11], where '5' is the most favourable response and '1' is the least favourable.

Twenty two native English speakers conducted the evaluation, and were divided into eleven blocks of two listeners. Two different Latin squares of order eleven were used (one for each section of the evaluation) to define the order in which systems were judged. Each listener block was assigned a row of each Latin square, and judged eleven different utterances per section, each synthesised by a different system.

Significant differences between systems are detected using a Bonferoni-corrected pairwise Wilcoxon signed rank test [12]. In the discussion which follows, a difference is deemed significant if this test discovers significance at the 99% confidence level.

5. Results

Figure 3 summarises listener judgements of 'naturalness' and 'similarity to target speaker' [12]. The system definitions are specified in the table above the boxplots, which also displays the mean opinion scores (MOS) for naturalness and similarity for each system. Since no significant differences were observed between the systems adapted using 10 utterances (C through F) and those adapted using 40 utterances (G through J), the discussion which follows concerns only the latter set of systems. The same discussion holds for the former set.

5.1. Similarity to target speaker

With regard to similarity to the target speaker, significant improvements over the average voice system B are observed in the case of all adapted systems (G through J). Moreover, no significant difference is found between any pair of these systems. This demonstrates that unsupervised adaptation of TTS models achieves performance indistinguishable to that of supervised adaptation (systems G and I) without use of supplementary acoustic models (systems H and J) or linguistic analysis of the adaptation data (system J). Result 1, as stated in Section 1, has therefore been demonstrated.

One additional result, in response to the hypothesis mentioned at the end of Section 3.1, is that no significant performance degradation or improvement is observed when using full adaptation (system H) instead of triphone adaptation (system J) in the unsupervised case. These results suggest that linguistic analysis does not improve or adversely affect the unsupervised adaptation procedure. Therefore the linguistic analysis stage may be omitted for the sake of efficiency.

5.2. Naturalness

With regard to naturalness, no significant differences are found between any pair selected from systems A through J. The volumes of adaptation data used are insufficient to significantly improve the naturalness of speech synthesised using the average voice models. Importantly, however, note that no significant difference in naturalness is observed between system A (standard decision tree construction) and system B (two-pass decision tree construction). So constraining decision tree construction using the two-pass technique has not compromised the naturalness of the resulting synthetic speech. Result 2, as stated in Section 1, has therefore been demonstrated.

6. Conclusions

A two-pass decision tree construction method has been introduced. This method enables sharing between full context models used for HMM-based speech synthesis and triphone models used for HMM-based ASR via a simple mapping between these models. This enables unsupervised adaptation of full context models without a separately estimated set of components. Further, the technique enables the components to be adapted without the use of linguistic analysis. These refinements have been introduced without any perceived degradation to the quality of the speech synthesis models.

System	Clustering	# utterances adaptation data	Adaptation method	Supervised?	MOS naturalness	MOS similarity
A	Standard	0			2.0	1.0
B	Two-pass	0			1.8	1.0
C	Two-pass	10	Full	Y	2.0	2.9
D	Two-pass	10	Full	N	2.1	2.5
E	Two-pass	10	Triphone	Y	2.0	2.9
F	Two-pass	10	Triphone	N	1.9	2.4
G	Two-pass	40	Full	Y	2.1	3.3
H	Two-pass	40	Full	N	1.9	2.8
I	Two-pass	40	Triphone	Y	2.1	2.9
J	Two-pass	40	Triphone	N	2.0	2.9
K					3.8	4.9

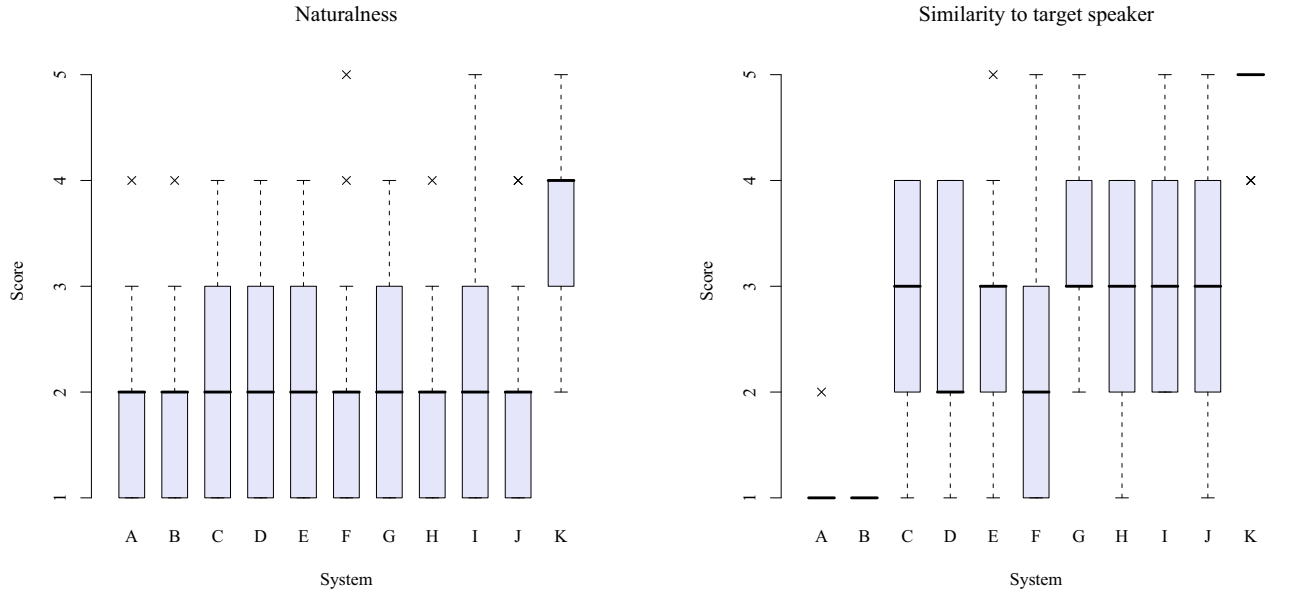

Figure 3: Boxplots of listener opinion scores for naturalness and similarity to target speaker.

7. 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 Communitys Seventh Framework Programme (FP7/2007-2013), grant agreement 213845 (EMIME).

8. References

[1] Karaiskos, V., King, S., Clark, R. and Mayo, C., "The Blizzard Challenge 2008", Proc. Blizzard 2008, 2008.

[2] Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., Kitamura, T., "Speaker interpolation in HMM-based speech synthesis system", Proc. Eurospeech, 1997

[3] Yamagishi, J., Onishi, K., Masuko, T. and Kobayashi, T., "Modeling of various speaking styles and emotions for HMM-based speech synthesis, Proc. Eurospeech, 2003.

[4] Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K. and Isogai, J., "Analysis of Speaker Adaptation Algorihms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm" IEEE Audio, Speech & Language Processing, 17(1):66–83, 2009.

[5] King, S., Tokuda, K., Zen, H. and Yamagishi, J., "Unsupervised adaptation for HMM-based speech synthesis", Proc. Interspeech, 2008.

[6] Yamagishi, J. Zen, H. Wu, Y.-J., Toda, T. and Tokuda, T., "The HTS-2008 System: Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System in The 2008 Blizzard Challenge", Proc. Blizzard 2008, 2008.

[7] Kawahara, H., Masuda-Katsuse, I. and Cheveigne, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds", Speech Communication, 27:187-207, 1999.

[8] Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., "Multi-space probability distribution HMM", IEICE Trans. Inf. & Syst., E85-D(3):455–464, 2002.

[9] Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., "Hidden semi-Markov model based speech synthesis", Proc. ICSLP, 2004.

[10] Tokuda, K., Yoshimura, T. Masuko, T., Kobayashi, T. and Kitamura, T., "Speech parameter generation algorithms for HMM-based speech synthesis", Proc. ICASSP, 2000.

[11] Likert, R., "A technique for the measurement of attitudes", Archives of Psychology, 140:1–55, 1932.

[12] Clark, R., Podsiadlo, M., Fraser, M., Mayo, C. and King, S., "Statistical analysis of the Blizzard Challenge 2007 listening test results", Proc. Blizzard 2007, 2007.