Speech recognition with speech synthesis models by marginalising over decision tree leaves

John Dines[1], *Lakshmi Saheer*[1,2], *Hui Liang*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Federale, Lausanne (EPFL), Switzerland
john.dines@idiap.ch, lsaheer@idiap.ch, hui.liang@idiap.ch

Abstract

There has been increasing interest in the use of unsupervised adaptation for the personalisation of text-to-speech (TTS) voices, particularly in the context of speech-to-speech translation. This requires that we are able to generate adaptation transforms from the output of an automatic speech recognition (ASR) system. An approach that utilises unified ASR and TTS models would seem to offer an ideal mechanism for the application of unsupervised adaptation to TTS since transforms could be shared between ASR and TTS. Such unified models should use a common set of parameters. A major barrier to such parameter sharing is the use of differing contexts in ASR and TTS. In this paper we propose a simple approach that generates ASR models from a trained set of TTS models by marginalising over the TTS contexts that are not used by ASR. We present preliminary results of our proposed method on a large vocabulary speech recognition task and provide insights into future directions of this work.

Index Terms: speech synthesis, speech recognition, decision trees, unified models

1. Introduction

Automatic speech recognition (ASR) and Text-to-Speech Synthesis (TTS) have long histories as relatively independent domains of research. In spite of this, the development of ASR and TTS technologies have similar stories, with the earliest systems relying on rule-based paradigms, followed by template based approaches and, most recently, statistical approaches. Though certain techniques and methodologies have managed to bridge the gap between these domains their relatively disparate objectives have kept the two largely apart. Such models have the potential to make significant contributions to fundamental scientific enquiries into speech production and perception and the links between the two, but also has the potential to reap more immediate benefits, most notably in the domain of speech-to-speech translation (ST), thus the development of unified frameworks for ASR and TTS remains a desirable goal for many researchers.

The EMIME[1] project has been conducting research in ST using a common hidden Markov model (HMM) statistical framework for both ASR and TTS. We envisage that a system using a common set of models for ASR and TTS would perform ST in the following manner: automatic speech recognition is first performed on input speech in the source language. As part of this ASR processing chain the ASR models are adapted to provide improved speech recognition performance. ASR output is processed by the translation engine before being passed onto the TTS system for synthesis in the target language. As part of this synthesis process the adaptation transforms that have been generated during ASR can also be applied to the TTS models, thus the identity of the speaker will be preserved in the ST output [1].

Development of such an ST system is possible using a common HMM based statistical framework, but several technical barriers still separate the ASR and TTS systems from being fully unified. One of the first challenges that must be overcome is that of how to share parameters between ASR and TTS models. The focus of this paper is the investigation of a *decision tree marginalisation approach*, a simple scheme that allows ASR contextual models to be derived from TTS contextual models while providing a seamless framework for applying adaptation transforms generated by the ASR to TTS models. The TTS models are not modified by this process.

The paper is organised as follows: in Section 2 we present a brief discussion on the convergence of ASR and TTS technology and in Section 3 we present the problem of contextual modelling in ASR and TTS systems and propose our approach that reconciles the inherent differences in ASR and TTS contextual modelling. In Section 4 we evaluate our proposed approach for ASR performance, demonstrating it's potential to enable both ASR and TTS within a common set of HMM parameters. In Section 5 we present a summary of our work and future directions.

2. Convergence of ASR and TTS

There has long been interest in building joint models for ASR and TTS. Arguably, a contributing factor to increased interest in this field is the convergence of ASR and TTS technologies in the form of statistical parametric models of speech. While the ASR community has used such models for some time, the TTS community has only more recently shown growing interest in such approaches, with the current state-of-the-art demonstrating the potential to challenge traditional concatenative approaches in terms of speech quality and naturalness [2]. In the context of ST systems, unified models are of particular interest as they would enable both ASR and TTS components to leverage from the same speaker adaptation algorithms.

Despite the adoption of the HMM as the common basic building block. there still remain many barriers to the successful development of models that can effectively perform both ASR and TTS. We have listed the major characteristics of ASR and TTS systems in Table 1, where it is evident that there still exist numerous discrepancies between the underlying modelling approaches. In addition to the features listed in this table, there

[1]Effective Multilingual Interaction in Mobile Environments: http://www.emime.org

Configuration	ASR	TTS
Acoustic parameterization		
Spectral analysis	fixed size window	STRAIGHT (pitch adaptive window)
Feature analysis	filter-bank cepstrum ($\Delta + \Delta^2$)	mel-generalised cepstrum ($+ \Delta + \Delta^2$)
Feature dimensionality	39	120
Frame shift	10ms	5ms
Acoustic modelling		
Number of states in HMM	3	5
Duration modelling	transition matrix	explicit duration distribution (HSMM)
Parameter tying	phonetic decision tree	shared decision tree
State emission distribution	multiple component Gaussian	single component Gaussian
Context	*triphone*	*full*

Table 1: Comparison of main components of ASR and TTS systems with respect to acoustic front-end and acoustic modelling.

may also exist discrepancies related to lexicon and phone set; speaker adaptation; and acoustic model training criteria. This paper is concerned with one of the most fundamental differences between ASR and TTS systems, that of the modelling of context.

3. Modelling context in ASR and TTS

The focus of this work is to provide a framework for performing ASR with TTS models, more specifically by addressing the differences between context dependent models used in ASR and TTS systems. In TTS, a broad range of contexts (so-called *full context*) are necessary, most importantly for the correct synthesis of prosodic features (duration, pitch etc.). Since such features are normally correlated with supra-segmental information, phonetic context is considered insufficient. Such contextual information can be predicted from text and provided as input to the HMM synthesis system. By contrast, ASR systems rely on relatively constrained context, most commonly *triphone context* is used since it is necessary to limit the search space – conducting a search over the full set of contexts used in TTS systems would be impossible.

In this section we first outline the basic approach for context dependent modelling in HMMs, in particular, we describe the decision tree-based parameter sharing scheme that is used in most modern ASR and TTS systems. This parameter sharing scheme also forms the basis of our proposed approach, which enables the generation of acoustic models conditioned on both ASR and TTS contexts.

3.1. Parameter sharing with decision trees

In HMM acoustic models it is usual that the acoustic units (such as phonemes) are modelled in context. Such contexts may include neighbouring phonemes as well as prosodic information (eg. phoneme stress) and supra-segmental information (phrase position,syllable position, etc.). Conditioning models on such a large set of variables inevitably results in sparsity of coverage within the training data, requiring that we use smoothing techniques for the estimation of state emission probability density functions. The most commonly used smoothing technique (and the defacto standard in ASR and TTS) is the decision tree clustering approach [3], which provides a many-to-one mapping between HMM states and state-clusters. Typically, in ASR one decision tree is constructed per state per base phone (phonetic trees) whereas in TTS one tree is constructed per state (shared decision tree). Emission pdfs are thus modelled at the state-cluster level.

We can consider that the decision tree maps the set of states to a set state-clusters where the tree is structured such that, during training, data at each branch is split in a greedy fashion according to questions pertaining to natural context groupings (ie. is the left context phoneme a vowel, is the centre phoneme stressed). The criterion used for splitting nodes is typically maximum likelihood and tree growth is controlled either by a likelihood threshold combined with leaf occupancy or minimum description length [3, 4]. Thus, the acoustic model become a pool of state cluster distributions, where each cluster is shared by one or more acoustic contexts. The parameters for each state cluster distribution are estimated from the acoustic observations associated to the contexts sharing that cluster.

We can thus define our notation for the decision tree clustering framework as follows. Firstly we define the set of states and state-clusters:

- $\mathcal{S}$ is the set of states for the HMMs
- $\mathcal{R}$ is the set of state-clusters for the HMMs

Each state-cluster models the conditional pdf of acoustic observations using a Gaussian mixture model (GMM):

$$p(\boldsymbol{o}|r) = \sum_{m=1}^{M^r} P(m|r)\mathcal{N}(\boldsymbol{o}|\mu_m^r, \Sigma_m^r) \tag{1}$$

where $r \in \mathcal{R}$, number of mixture components M^r, mixture component prior probabilities $P(m|r)$ and corresponding normal distribution, $\mathcal{N}(\boldsymbol{o}|\mu_m^r, \Sigma_m^r)$, with mean and covariance μ_m^r and Σ_m^r.

We can consider the decision tree as a function mapping states to clusters such that $f : \mathcal{S} \rightarrow \mathcal{R}$. We can write this as:

$$P(r|s) = \begin{cases} 1, & f(s) = r \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $s \in \mathcal{S}$ and $r \in \mathcal{R}$.

Thus, using Equations 1 and 2 it follows that each state emission density may be represented as:

$$\begin{aligned} p(\boldsymbol{o}|s) &= \sum_r P(r|s)p(\boldsymbol{o}|r) \\ &= p(\boldsymbol{o}|f(s)) \end{aligned}$$

3.2. Marginalising TTS contexts via the decision tree

Our goal is to derive an acoustic model of both ASR and TTS contexts in which the underlying parameters are shared between ASR and TTS contextual models. First of all we define the following notation for our ASR and TTS models:

- $\mathcal{S}^{TTS}$ is the set of states for the TTS HMMs
- $\mathcal{S}^{ASR}$ is the set of states for the ASR HMMs
- $\mathcal{R}^{TTS}$ is the set of state-clusters for the TTS HMMs
- $\mathcal{R}^{ASR}$ is the set of state-clusters for the ASR HMMs

Since the ASR context is part of the TTS contextual representation, we may generate an ASR contextual model by marginalising over the set of non-ASR contexts:

$$p(\boldsymbol{o}|s^{ASR}) = \sum_{\{s^{TTS}:s^{TTS}\mapsto s^{ASR}\}} P(s^{TTS}|s^{ASR})p(\boldsymbol{o}|s^{TTS}) \tag{3}$$

where $s^{TTS} \mapsto s^{ASR}$ means that s^{TTS} belongs to the set of TTS contexts of which ASR context s^{ASR} is a member (ie. disregarding contexts that are not relevant to ASR, the TTS context is equivalent to the ASR context). It is also understood that $s^{ASR} \in \mathcal{S}^{ASR}$ and $s^{TTS} \in \mathcal{S}^{TTS}$.

We know that, in practice, HMM states are modelled by state-cluster distributions, thus, we can simplify Equation 3 according to our knowledge of the clustering. More specifically we can marginalise the TTS HMM set over the set of TTS state-clusters in order to obtain the ASR state emission pdf. In mapping ASR states using the TTS decision tree, we ignore irrelevant questions at tree branches, thus arriving at multiple leaf nodes for a single ASR context. The ASR contextual model becomes a mixture distribution of the TTS decision tree leaves, which can be likened to a tied mixture or semi-continuous system in which the pool of Gaussians is determined by the TTS decision tree leaf nodes.

Firstly, the TTS state-clusters for each ASR context comprise a subset, $\mathcal{R}^{s^{ASR}} \subset \mathcal{R}^{TTS}$, as determined by the TTS decision tree:

$$\mathcal{R}^{s^{ASR}} = \{r^{TTS} \in \mathcal{R}^{TTS} : f^{TTS}(s^{ASR}) = r^{TTS}\}$$
$$\mathcal{R}^{ASR} = \{\mathcal{R}^{s^{ASR}} : s^{ASR} \in \mathcal{S}^{ASR}\}$$

We define a set of mixture weights for the ASR states by:

$$P(r^{TTS}|s^{ASR}) = \begin{cases} \frac{N(r^{TTS})}{\sum_{\hat{r}^{TTS}\in\mathcal{R}^{s^{ASR}}} N(\hat{r}^{TTS})}, & r^{TTS} \in \mathcal{R}^{s^{ASR}} \\ 0, & r^{TTS} \notin \mathcal{R}^{s^{ASR}} \end{cases} \tag{4}$$

where we define $N(r^{TTS})$ as the occupation count for state-cluster r^{TTS} during model training.

The ASR state emission pdf is thus approximated by:

$$\begin{aligned} p(\boldsymbol{o}|s^{ASR}) &\approx \sum_{r^{TTS}} P(r^{TTS}|s^{ASR})p(\boldsymbol{o}|r^{TTS}) \\ &= \sum_{r^{TTS}} P(r^{TTS}|s^{ASR}) \\ &\times \sum_{m=1}^{M^{r^{TTS}}} P(m|r^{TTS})\mathcal{N}(\boldsymbol{o}|\mu_m^{r^{TTS}}, \Sigma_m^{r^{TTS}}) \end{aligned} \tag{5}$$

We note that our proposed approach not only provides a framework for modelling ASR and TTS contexts using a common set of acoustic model parameters, but is also consistent with the desire to use multiple Gaussian and single Gaussian state emission distributions in ASR and TTS acoustic models, respectively.

4. Evaluation and analysis

Preliminary evaluation of the was carried out by comparing ASR performance of a conventional triphone context ASR models with that of triphone context models derived from TTS full-context using our proposed approach. We do not evaluate TTS performance as the proposed approach does not modify the TTS models.

4.1. Wall Street Journal system

We built several systems based on the HTS entry to the 2007 Blizzard Challenge [5]. Thus, models are trained using maximum likelihood speaker-adaptive training (ML-SAT)[6]. We also trained speaker independent models (ML-SI) for first-pass ASR decoding. For our initial studies we made several changes to the HTS training scripts. Firstly, we train conventional hidden Markov models instead of hidden semi-Markov models (HSMM) in order to avoid difficulties associated with decoding using explicit duration distributions. Secondly, we use conventional ASR features, 13$^{\text{th}}$ order perceptual linear prediction coefficients (PLP) [7], in order that our system is comparable to published results for similar systems[2]. We also omit the $\log F_0$ and aperiodicity features typically employed in HMM-based speech synthesis. Thirdly, we use a flat-start training regime which performs realignment of the training data with the word-level transcripts at several stages during training.

Training data comprised the Wall Street Journal (WSJ0) short term speaker training data (SI84)[9]. Full context labels are generated using a TTS front-end based on the Unisyn lexicon[10]. We are obliged to discard any transcripts for which there is insufficient agreement between the forced alignments and the full context labels that were generated by the TTS front-end. Of the original 7240 training utterances from the WSJ0 corpus we discarded approximately 10% of training data due to this alignment issue. In the future we plan to use a more sophisticated approach for generating full-context labels that should avoid this issue. For further details of the WSJ TTS system see [11].

Decoding is carried out using a two-pass system. The first pass uses a set of speaker independent models from which adaptation transforms are estimated and applied to the SAT models in the second decoding pass. For the models using 'decision tree marginalisation' it is first necessary to convert the full-context models to triphone context models and likewise the base classes for the adaptation transforms must be mapped from the full context models to the triphone context models.

4.2. ASR results

For the evaluation of ASR we use the primary condition (P0) of the 5k vocabulary hub task (H2) of the November 93 CSR evaluations. Decoding employs the 5k closed bigram language model that is distributed with the corpus and is carried out using speaker independent models for the first pass (ML-SI) and

[2]For a comparison of systems trained using conventional ASR features and those using TTS features see [8]

speaker adaptive models (ML-SAT) in the second pass. The results of comparisons of several systems are shown in Table 2. The table shows results for four systems, where the first two systems have been trained using standard triphone context. The last two models have been trained using full-context, thus the marginalisation approach has been used to convert the full context models to triphone context models. This results in multiple state-clusters in the TTS models being mapped to each ASR state-cluster and, consequently, the ASR state-clusters have a greater average number of mixture components per state-cluster.

Context	# mixtures	# mixes / state		WER (%)	
		ASR	TTS	ML-SI	ML-SAT
Triphone	3,148	1	–	15.7	12.1
	50,368	16	–	10.9	7.1
Full	3,792	9.4	1	16.7	14.1
	45,504	112.4	12	12.5	8.6

Table 2: Comparison of ASR performance for standard and proposed triphone models.

Inspection of these results reveals several observations of interest. First of all, it is evident that the marginalised triphone models have, on average, almost ten mixture components per state-cluster even when the TTS models from which they are only composed of a single Gaussian per-state pdf. In reality, the capacity of these models is not significantly greater than the standard triphone models as evidenced by the total number of mixture components in these systems. Thus, we can conclude that the bulk of the mixture components in the marginalised system are being shared between multiple state-clusters. A comparison between standard triphone models and the proposed triphone models show this is detrimental ASR performance, which we hypothesise is due to two factors: a greater partitioning of the data by the decision tree (compared to usual GMM-based state emission pdfs) and increased confusion between models due to the high degree of parameter sharing.

We trained a second pair of models with a greater number of mixture components such that the standard triphone and proposed triphone system have approximately the same total number of parameters. We set the number of mixture components to match that which we have used in complimentary WSJ ASR studies [8]. The performance of the standard triphone system increases by around 40% relative to the single mixture component system. The marginalised triphone system also improves substantially, but results in models with a very large number of mixture components per state. Additionally, the consequences of using multiple mixtures per state in a TTS system have not been extensively reported, though anecdotal evidence suggests that this can be detrimental to TTS quality.

From these results we can conclude that future research needs to focus on two directions: firstly we need to find a means to increase the capacity of the ASR models while not impacting on TTS performance and not resulting in an excessive number of mixture components in the ASR states. Secondly, we should further analyse the effects on ASR performance when partitioning the data using non-triphone questions and if necessary devise means to avoid any undesirable consequences of this. Further research directions should also include the investigation of TTS labels in ASR that include prosodic and supra-segmental contexts. This may be combined with TTS features ($\log F_0$ and aperiodicity features).

5. Conclusions and future work

We have presented a method for generating ASR models from TTS models by marginalising out the contexts of the TTS models that are not relevant to the ASR models. This is done by approximating the full summation over non-ASR contexts by that of a summation over the leaves of the decision tree used for clustering states. The proposed approach will enable the use of TTS models directly in ASR such that unsupervised adaptation of the TTS models can be applied in a principled fashion by directly using adaptation transforms generated during ASR. Our initial experiments show that the proposed approach is feasible, though it leads to degradation of ASR performance compared to standard triphone ASR models. Our future goal will be to investigate means for alleviating the drawbacks of the current approach and to extend our studies to include analysis of TTS performance in the context of unsupervised adaptation.

6. Acknowledgements

The authors would like to thank the Junichi Yamagishi for his assistance with the HTS system training scripts.

The research leading to these results was partly funded from the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement 213845 (the EMIME project).

7. References

[1] S. King, K. Tokuda, H. Zen, and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," in *Proc. Interspeech 2008*, Sep. 2008, pp. 1869–1872.

[2] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge Workshop*, Brisbane, Australia, September 2008.

[3] J. J. Odell, "The use of context in large vocabulary continuous speech recognition," Ph.D. dissertation, Queens College, University of Cambridge, 1995.

[4] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. Eurospeech*, vol. 1, Rhodes, Greece, 1997, pp. 99–102.

[5] J. Yamagishi, H. Zen, T. Toda, and K. Tokuda, "HTS-2007 system for the Blizzard challenge 2007," in *Proc. of Blizzard Challenge 2007 workshop)*, Bonn, Germany, August 2007.

[6] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP-96*, Oct. 1996, pp. 1137–1140.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[8] J. Dines, J. Yamagishi, and S. King, "Measuring the gap between HMM-based ASR and TTS," in *Proc. Interspeech*, vol. Submitted, Brighton, UK, September 2009.

[9] D. Pallet, "DARPA February 1992 pilot corpus CSR "dry run" benchmark test results," in *Proceedings of the workshop on Speech and Natural Language*, Harriman, USA, February 1992, pp. 382–386.

[10] S. Fitt and S. Isard, "Synthesis of regional English using a keyword lexicon," in *Proc. Eurospeech*, vol. 2, Sep. 1999, pp. 823–826.

[11] J. Yamagishi et. al., "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech*, vol. Submitted, Brighton, UK, September 2009.