THE UNIVERSITY of EDINBURGH

Deliverable D3.2

Report describing initial version of deep layered models

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287678.

Participant no.	Participant organisation name	Part. short name	Country
1 (Coordinator)	University of Edinburgh	UEDIN	UK
2	Aalto University	AALTO	Finland
3	University of Helsinki	UH	Finland
4	Universidad Politécnica de Madrid	UPM	Spain
5	Technical University of Cluj-Napoca	UTCN	Romania

Project reference number	FP7-287678
Proposal acronym	SIMPLE4ALL
Status and Version	Complete, proofread, ready for delivery: version 1
Deliverable title	Report describing initial version of deep layered models
Nature of the Deliverable	Report (R)
Dissemination Level	Public (PU)
This document is available from	`http://simple4all.org/publications`
WP contributing to the deliverable	WP3
WP / Task responsible	WP3 / Task T3.2
Editor	Oliver Watts (UEDIN)
Editor address	owatts@staffmail.ed.ac.uk
Author(s), in alphabetical order	Rob Clark, Simon King, Oliver Watts
EC Project Officer	Pierre Paul Sondag

Abstract

The most common approach to integrating linguistic and textual contexts into acoustic modelling involves constructing highly context-dependent models of phone-sized units, then solving the resulting data sparsity problem by sharing parameters between related models, using decision tree-based state-tying. Decision trees may not be the best choice for this task, because of the factorial nature of context-dependency.

This report describes initial work on deep layered models for TTS, which offer an alternative way to integrate linguistic and textual contexts into acoustic modelling. Work on three different approaches to overcoming the problems of the established approach is described.

Contents

1 Introduction

Currently, the most widely-used method in statistical parametric TTS for incorporating linguistic context into acoustic models is to use highly context-dependent model names. The name of a model will typically include the identity of that model's phoneme, neighbouring phonemes' identities, the position of the phoneme in a syllable, word, intonational phrase etc., and whether the syllable of which the unit is part is lexically stressed, pitch-accented, etc. In effect, the hierarchical textual/linguistic specifications constructed by a TTS text analyser (see Deliverable 2.1) are flattened to a sequence of highly context-dependent model names. The combination of many features in a model name results in a such a vast number of possible model names that most seen at training time will be of a unique type. In order to map to a smaller number of models whose parameters can be estimated from available data and which will generalise to the many new, unseen model types encountered at synthesis time, decision tree context clustering of submodel states and subsequent sharing of state parameters is typically used.

This conventional strategy of flattening linguistic specifications into a vast number of models which are then pooled in the leaf nodes of a decision tree has drawbacks. Firstly, it is scientifically unsatisfying: it seems clear that e.g. the hierarchical phrase- and syllable-structures inherent in utterances should be exploited directly in the estimation of acoustic models.

Secondly, the 'divide-and-conquer' technique of decision trees fails to make best use of data for modelling factorial phenomena. To take one example, both phoneme identity and whether or not a phoneme is in a phrase-final syllable might both be expected to contribute to the length of an acoustic segment corresponding to that phoneme. If a question querying phoneme identity occurs high up a decision tree for predicting a distribution over state duration, for example, then a question querying phrase-finality might occur independently in both resulting subtrees. In this case, the contribution of phrase-finality is modelled independently for different phoneme types, and thus using smaller subsets of the data.

Thirdly, the traditional TTS paradigm where text analysis and waveform generation are performed in a stage-wise fashion means that:

1. A lot of expert knowledge concerning the types of linguistic features that are expected to affect the acoustic realisation of an utterance is required to construct a text analyser

2. There is no guarantee that the features assigned by a text analyser are the most relevant to synthesising the acoustic signal.

These objections could be overcome in a framework where linguistic and acoustic representations are learned jointly from data; initial steps in this direction were taken in [10] and this work is continuing at present, as outlined in Section 3 of the current document.

The original plan for this part of Work Package 3 was to use Multiple Regression HMMs (MR-HMMs) and HMMs with stream dependencies to integrate numerical linguistic representations of the sort discussed in Deliverable D2.1 into acoustic modelling. This strategy was built upon previous work by Junichi Yamagishi who has used these models to incorporate stylistic and articulatory parameters into TTS and who was expected to be a member of SIMPLE4ALL . Since then, circumstances have changed because Dr. Yamagishi has obtained independent sources of funding (including a prestigious EPSRC Career Acceleration Fellowship) to pursue his work on the MR-HMM and related models for speech synthesis. The benefit of this development for SIMPLE4ALL is clear: although Dr. Yamagishi is no longer funded by SIMPLE4ALL , he continues to be actively engaged with and interested in this project, and the work being conducted in his other projects is well-aligned with that of SIMPLE4ALL . Consequently, the results of much of the work that was planned to be conducted under SIMPLE4ALL funding have become available to the project at no cost. For example, the codebase (a modified version of HTS) that has been developed in external projects has become stable enough to be usable by a wider community, and is now due to be publicly released as part of HTS 2.3 in December 2012. The creation and use of such freely-available resources is consistent with the SIMPLE4ALL commitment to the free public availability of the software developed.

In light of Dr. Yamagishi's parallel research projects, we have modified the plan of work within SIMPLE4ALL – we still plan to employ MR-HMMs as a mechanism to incorporate linguistic context into the acoustic model, but will delay our work on this until after the public release of HTS 2.3 so that we can build upon this stable codebase. In the meantime, time has been dedicated to two alternative approaches to the same problem. The future planned work using MR-HMM will be outlined in Section 4, but first the alternative approaches (on which project time has actually been spent) will be reported in Sections 2 and 3. At this stage, we are reporting only preliminary experiments and will report the later planned work in a future deliverable.

2 Explicit modelling of syllable structure

An additional problem with the traditional HMM approach to TTS is that the prosodic modelling of F_0 to synthesise the resulting pitch movement in the speech naively follows the same structure as modelling the frame-level spectral characteristics of speech. The models that generate F_0 are trained as if pitch is a property of the segment in the same way that spectral information is, whereas in practice most aspects of prosody are syllable- or utterance-level effects. We have started to investigate ways in which to provide a more appropriate framework to model prosody. We are in the process of building comparable voices with conventional models and with models that use syllable-sized units. The aim is first to make a baseline comparison between models which have the different underlying structures, and then to compare the the use of appropriate beyond-the-utterance level linguistic features (e.g. position in paragraph, type of utterance) within these frameworks. Initial exploratory experiments (conducted by a Masters student) using beyond-the-utterance level linguistic features [4] suggest that they could be useful in producing better prosody, but at the current time more substantial amounts of speech are required to obtain reasonable coverage.

Our future plans are to investigate ways of jointly training F_0 models that have a different underlying structure to the segment models, such as multi-rate and variable-rate modelling [1] which has been applied to ASR, to investigate ways to reduce the amount of data required to enable the technology to be easily adapted to new languages, and to incorporate automatically generated beyond-the-utterance level features. The hope is that some of this research can be usefully carried out jointly with a recently-funded SNSF (Swiss National Science Foundation) project in which UEDIN is a partner, and is looking at similar issues. UEDIN plans to take an intern to work on this topic, starting early in 2013.

3 Deep modelling of linguistic representations

3.1 Vector space models for TTS

Introduced in [10] (PhD work which was a precursor to SIMPLE4ALL and was completed in the early stages of this project) and discussed in Deliverable D2.1, is an approach to the unsupervised construction of representations for context modelling in TTS, using vector space models (VSMs) at various levels of analysis (e.g. letter, word) which are built on large amounts of unlabelled text. The use of these models has met with considerable success on various tasks, such as standing in for a set of phonetic features during acoustic state tying [10, §5.3.1] and standing in for part of speech tags for the task of phrase-break prediction [10, §6.3, §7.1]. In both cases, the incorporation of vector space models closes much of the gap in performance between baseline systems and toplines where expert knowledge is included.

The technique is attractive because it incorporates unsupervised learning and thus reduces the amount of expert time and knowledge needed for system construction. Unlabelled text data can typically be collected in larger amounts than labelled data, enabling the use of larger training sets.

In the distributional phase (i.e., unsupervised acquisition of the model from unlabelled data), vector space models discover a continuous space and do not partition this space into categories. Rather, it is left to subsequent models to either explicitly or implicitly infer boundaries in this space in order to form categories that are actually relevant for the task at hand to be created.

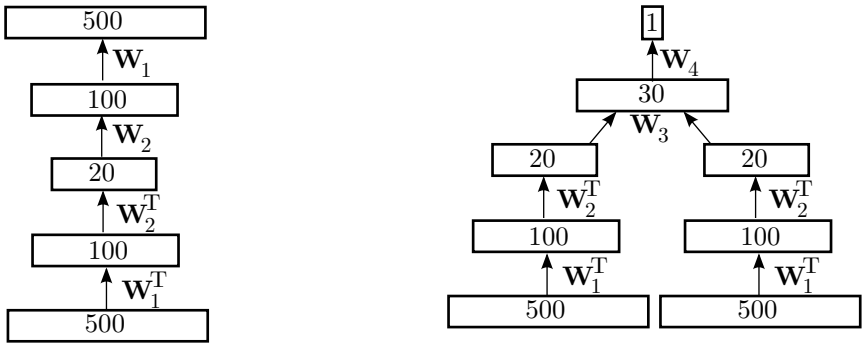

Figure 3.2a: ***Left:*** *A deep autoencoder can be used to derive a low-dimensional representation from a vector of word counts; each set of weights is pretrained to encode activations of the previous layer.* ***Right:*** *weights discovered in an autoencoder can be incorporated into an MLP which is trained on a supervised task such as phrase-break or acoustic prediction, using left- and right-context word counts as input.*

3.2 Shortcomings of VSMs

The technique as presented in [10] and Deliverable 2.1 has weak points; for example, there is still a reliance on decision trees for model clustering: in other words, decision trees are used to infer boundaries in the continuous vector space and thus to form categories in that space. Decision trees are susceptible to noisy and numerous independent variables where irrelevant features are present. In [10, §7.2] a method of feature selection using ensembles of trees and artificial contrast variables is developed; this feature selection turned out to be a necessary step before the features derived from the word-level VSM could be demonstrated to have benefits for acoustic state-tying. In later experiments which simultaneously used VSM features from multiple levels of textual analysis, however, use of this feature selection method did not lead to stable behaviour [10, §8.6.2].

Also in the initial distributional phrase, the techniques used are most likely suboptimal. Use of the singular value decomposition for reducing the dimensionality of the raw word-count matrix is borrowed from the *Latent Semantic Indexing* approach to Information Retrieval [5], and has the advantage of the ready availability of computationally feasible implementations. However, it also has clear limitations, employing only a single layer of hidden features, which are simply linear combinations of the observed counts. This lead us to consider the use of more powerful methods for discovering latent structure in unlabelled data.

3.3 Connectionist Methods

Connectionist methods have already been applied to the problem of mapping from vectors of word counts to low-dimensional representations. [9] outlines a technique for training a deep autoencoder for reducing the dimensionality of word count observations. An experiment in Information Retrieval is reported where the autoencoder outperforms singular value decomposition, and it is supposed that similar benefits might also be obtained by using such deep autoencoders for inducing representations of linguistic objects for TTS. The shortcomings of SVD mentioned above might be overcome through the use of such techniques.

The goal of this part of our work is to implement the concepts outlined in [10] using neural networks. The use of connectionist methods introduces the possibility of a model which can be learned in both unsupervised and supervised fashion. First, unsupervised learning is used to learn a latent space; this is done by training an autoencoder network from word counts to word counts, and which has a relatively small middle layer. This layer is a lower-dimensional representation of the space of word counts. A multilayer perceptron (MLP) is then created, with weights initialised from this mapping to the lower-dimensional representation; the MLP is then further trained in a supervised fashion, e.g., for the prediction of acoustics.

In contrast to decision trees, an MLP simultaneously considers all predictor features for each training example. Consequently, it does not suffer from the problem of data-fragmentation, and is less adversely affected by the presence of irrelevant attributes.

A major benefit of the combination of unsupervised and supervised learning is that the latent representations acquired by unsupervised learning on (potentially very large amounts of) unlabelled dat can be tailored towards the end task through supervised learning on (much smaller amounts of) labelled data. This is possible to some extent in [10], where decision trees select relevant dimensions of the discovered spaces, and partition those dimensions in a way that is relevant to the task. However, arbitrary transformations of the discovered representations are out of scope for that method. Connectionist methods, on the other hand, have already used on letter [6] and word [2, 3] representations, which are updated in a way that is optimal for certain NLP tasks of interest.

One ultimate destination for this work is to apply deep representations of TTS linguistic and textual context directly to the prediction of the parameters of states of an acoustic model for novel contexts. This work also benefits from overlapping interests with another project in speech technology ongoing at UEDIN (EPSRC-funded "Natural Speech Technology"), in which preliminary experiments on replacing state-clustering decision trees with a MLP have already been conducted in collaboration with SIMPLE4ALL . Work in SIMPLE4ALL will focus on inducing word representations with autoencoders in an unsupervised stage (see Figure 3.2a, left), and then fine-tuning these discovered representations to be optimal for predicting e.g. distributions over acoustics (see Figure 3.2a, right).

Before attempting this, we wish to evaluate the latent representation induced by the auto-encoder. As in [10], we are doing this by using the representation to perform a task that is well-understood and simpler than the task of acoustic prediction: phrase-break prediction using a decision-tree classifier.

4 Multiple Regression HMM

[7] presents several models for the introduction of continuous-valued features into an HMM-based speech synthesis system in such a way that the system is made controllable. For example, in the Multiple Regression HMM (MRHMM) model presented there (and depicted in Figure 4.0a), the state sequence $\boldsymbol{q}$ is supplemented by an auxiliary feature sequence $\boldsymbol{Y}$. In [7], $\boldsymbol{Y}$ is a sequence of feature vectors derived from measured articulator positions which were recorded from a speaker alongside the acoustic signal $\boldsymbol{X}$; previous work in TTS has integrated utterance-level style control vectors using the same model [8]. The likelihood of acoustic feature sequence $\boldsymbol{X}$ is conditioned on $\boldsymbol{Y}$ via a state-dependent linear transform $\boldsymbol{A}_j$ for state j: the mean of the conditional probability density function $b_j(\boldsymbol{x}_t|\boldsymbol{y}_t)$ at time t is based on the linearly-transformed articulatory vector observed at that time.

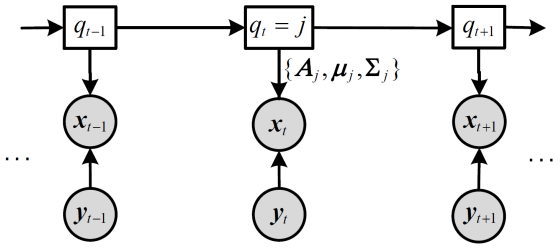

Figure 4.0a: *MRHMM, from [7].*

In [7] this and related models are presented as a way of controlling an HMM-based synthesiser's hard-to-interpret output acoustic parameters $\boldsymbol{X}$ via the easier-to-interpret articulatory features $\boldsymbol{Y}$, but without losing the high quality of resynthesis that can be achieved by a state-of-the-art vocoder driven by $\boldsymbol{X}$.

But even if we are not able to assign an intuitively-appealing interpretation to $\boldsymbol{Y}$, these models still offer a novel way to integrate generic continuous features into an HMM-based synthesiser. The work planned within SIMPLE4ALL is to introduce vector space representations of linguistic and textual context (cf. Deliverable D2.1)

in place of articulatory features. It is expected that introducing features in this way via a linear transform instead of via a context clustering decision tree will alleviate the problems associated with data fragmentation during decision tree training.

Context-dependent state-clustering can be combined with the MRHMM. For example, the MRHMM will allow us to have states that are dependent only on short-range graphemic or phonetic contexts, with prosodic contexts being integrated via the stream of auxiliary feature vectors.

Continuously-valued features derived from text are fundamentally different from the type of auxiliary features that have been previously used with the MRHMM: unlike articulatory or formant features which generally evolve smoothly, and unlike style control vectors which remain fixed over an utterance, numerical text-derived features remain steady over multiple consecutive states and jump at some state boundaries. It is anticipated that some modification of the textual features (e.g. smoothing over state-boundaries) might be necessary for their successful incorporation in a system using MRHMM.

References

[1] O. Cetin and M. Ostendorf. Multi-rate and variable-rate modeling of speech at phone and syllable time scales. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 665 – 668, 18-23, 2005.

[2] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning, ICML*, 2008.

[3] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *CoRR*, abs/1103.0398, 2011.

[4] Benjamin Dawson. Expressive speech synthesis, 2012.

[5] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[6] Kåre Jean Jensen and Søren Riis. Self-organizing letter code-book for text-to-phoneme neural network model. In *INTERSPEECH*, pages 318–321, 2000.

[7] Z. Ling, K. Richmond, and J. Yamagishi. Articulatory control of hmm-based parametric speech synthesis using feature-space-switched multiple regression. *Audio, Speech, and Language Processing, IEEE Transactions on*, PP(99):1, 2012.

[8] Takashi Nose, Junichi Yamagishi, and Takao Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE Trans. Information and Systems*, E90-D(9):1406–1413, September 2007.

[9] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969 – 978, 2009. Special Section on Graphical Models and Information Retrieval.

[10] Oliver Watts. *Unsupervised Learning for Text-to-Speech Synthesis*. PhD thesis, University of Edinburgh, 2012.