

# Hidden Markov Models with Context-Sensitive Observations for Grapheme-to-Phoneme Conversion

Kalu U. Ogbureke, Peter Cahill, Julie Carson-Berndsen

CNGL, School of Computer Science and Informatics,  
University College Dublin, Belfield, Dublin 4, Ireland

kalu@ucdconnect.ie, peter.cahill@ucd.ie, julie.berndsen@ucd.ie

## Abstract

Hidden Markov models (HMMs) have proven useful in various aspects of speech technology from automatic speech recognition through speech synthesis, speech segmentation and grapheme-to-phoneme conversion to part-of-speech tagging. Traditionally, context is modelled at the hidden states in the form of context-dependent models. This paper constitutes an extension to this approach; the underlying concept is to model context at the observations for HMMs with discrete observations and discrete probability distributions. The HMMs emit context-sensitive discrete observations and are evaluated with a grapheme-to-phoneme conversion system.

**Index Terms:** hidden Markov models, grapheme-to-phoneme conversion, machine learning

## 1. Introduction

This paper presents an approach to grapheme-to-phoneme conversion using hidden Markov models (HMMs) with context-sensitive observations. HMMs are generators of observations and hidden states [1] as well as statistical models for a sequence of observation vectors [2]. In grapheme-to-phoneme conversion, the observations are the letters or alphabet of a language and the phonemes are the hidden states. Conventionally, automatic grapheme-to-phoneme methods are used when the dictionary does not contain a particular word. This enables speech systems to estimate the pronunciation of any word and not be limited to the entries in the dictionary.

Every HMM is characterised by a number of parameters as denoted in [1]. In this paper, the number of states is represented by  $N$ ; while  $\Sigma$  is a set of distinct observation symbols and  $M$  is the number of distinct observation symbols. The codebook (parametrisation of observations) indices are  $V$ .  $S$  and  $O$  are the state and observation sequences respectively, while each observation symbol is emitted at a time  $t$ . Furthermore,  $A$  is a matrix of state transition probabilities;  $B$  is a set of state observation probabilities;  $\pi$  is a set of initial state probabilities and  $\lambda$  denotes the whole model.

$$S = \{S_1, S_2, S_3, \dots, S_N\} \quad (1)$$

$$t = \{1, 2, 3, \dots, T\} \quad (2)$$

$$\Sigma = \{O_1, O_2, O_3, \dots, O_M\} \quad (3)$$

$$V = \{V_1, V_2, V_3, \dots, V_M\} \quad (4)$$

$$O = \{O_1 O_2 O_3 \dots O_T\} \quad (5)$$

$$\lambda = (A, B, \pi) \quad (6)$$

The motivation for the approach presented in this paper includes modelling context at the observations in order to address

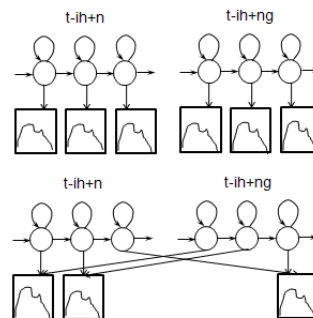


Figure 1: Data-driven tied-state triphones.

context effects arising from sequences of graphemes [3]. The context of graphemes affects the pronunciation of words, for example, the pronunciation for the English word ‘phone’ is /f oʊ n/; the grapheme ‘p’ is realised as the phone ‘f’, before the letter ‘h’, because of the aspirated ‘h’. Current HMM framework is unable to use the grapheme context directly [4]. The technique presented represents an augmentation to HMMs where the underlying concept is to model context with respect to graphemes (i.e the observations).

The remainder of this paper is organised as follows. Section 2 gives an overview of context modelling at the hidden states of the model. Section 3 describes the proposed discrete HMMs that emit context-sensitive observations. Section 4 gives an overview of grapheme-to-phoneme conversion using HMMs. The proposed approach and data are presented in section 5, while section 6 presents the results of an evaluation. Section 7 discusses the results and section 8 concludes with a summary of the contribution and highlights the future direction of this research.

## 2. Context modelling at the hidden states

This section reviews the traditional paradigm of context modelling at hidden states for HMMs with phones as hidden states. Context-dependent models represent each phone based on their right and left contexts. They are context-dependent phones of the parent phones [5, 6]. The upper part of figure 1 shows 2 context-dependent phones of the phone ‘ih’ from [7]. The triphone ‘t-ih+n’ represents the phone ‘ih’ with the phone ‘t’ as the left context and the phone ‘n’ as the right context. The upper part of the figure shows the grouping of the context-dependent phones and the lower part shows the sharing of distribution between similar states. However, one of the drawbacks of context-

dependent models is insufficient training data for the large number of contexts generated due to a large number of parameters associated with each model. The effect of this is greatly reduced by state tying, which balances model complexity against the amount of available training data [5, 6]. This is achieved by clustering and tying acoustically similar states within each triphone set [5]. States that are tied share common parameters which could be the means, variances, or transition probabilities. There are two methods currently used to cluster and tie acoustically similar states, namely, data-driven and tree-based clustering [6, 7]. An example of data-driven clustering is shown in the lower part of figure 1.

### 3. HMMs with context-sensitive observations

This section presents discrete HMMs that emit context-sensitive observations. The underlying concept is to model contexts with respect to graphemes (i.e. at the observations). This addresses the problem of modelling context at the observations. Previous research has shown the effect of contexts between observations. In [8], the emission probability is modified to account for the correlation between successive feature vectors. The work described in this paper is an extension to modelling output probability distribution  $B$  in HMMs with a discrete probability distribution.

#### 3.1. Context-sensitive observations

The first step in this extension is to generate the context-sensitive observations. This involves transforming each observation sequence with respect to their left and right contexts. The observation sequence in equation 5 is thus transformed to:

$$O' = \{\#O_1O_2 O_1O_2O_3 \dots O_{T-1}O_T\# \}, \quad (7)$$

where  $O'$  is the new context-sensitive observation sequence,  $\#$  is the sequence boundary, and  $O_1O_2O_3$  is a short form for  $O_1 - O_2 + O_3$ , the second observation in the context of the first and third observation symbols. This is necessary in order to differentiate, for example, the 'p' in the word 'phone' which is realised as the phoneme /f/, from 'pen' where it is realised as /p/. For example, in the grapheme-to-phoneme conversion of the word 'book',  $O = \{b o o k\}$  and  $O' = \{\#bo boo ook ok\# \}$ . The effect of this transformation is an increase in the number of distinct observation symbols  $M$  which requires much data for the training. Furthermore, there is the problem of new contexts not seen in the training examples.

One of the ways of dealing with insufficient training data is to transform observations with respect to either the left or right context, thereby reducing  $M$ . Another option is to define contexts with respect to broad classes if they exist. In grapheme-to-phoneme conversion, such classes could be vowels and consonants as the class of contexts influences the realisation of the sound. For example, the letter 'x' is almost always realised as the phonemes /k/ /s/ when the left context is a vowel.

The unseen contexts problem can be dealt with by clustering and tying similar contexts. If the observations have natural groupings as in the grapheme-to-phoneme problem, then any unseen context is tied to any observed context where the left and right contexts belong to the same class. The second option is to tie unseen context with observed context based on a distance measure if there are no natural groupings in the observations. Tied contexts share the same codebook (parametrisation)

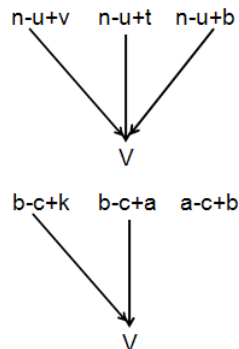


Figure 2: Clustering and tying similar contexts.

index  $V_i \in V$  shown in equation 4. This is illustrated in figure 2. In the upper figure, the three context-sensitive observations are tied because their right contexts  $v, t$  and  $b$  are consonants. If 'n-u+v' was not observed in the training data, it would share the same codebook index with 'n-u+t' and 'n-u+b'. In the lower figure, 'b-c+k' and 'b-c+a' are tied based on a distance measure and is useful when there are no natural groupings or classes in the observations.

### 4. Grapheme-to-phoneme using HMMs

One application of HMMs with context-sensitive observations is grapheme-to-phoneme conversion. While decision trees are the most common method for grapheme-to-phoneme conversion, HMMs have also been used with some success [3]. The major difference between decision tree and HMM approaches concerns how context is modelled. Decision trees commonly use 3 graphemes either side of the grapheme being classified (i.e. 7-gram) as attributes. In the case of the previous HMM approach, context is modelled with respect to phonemes rather than graphemes. In the case of the methods presented in the literature, only first-order HMMs have been used for this problem, where the only context modelled is a single preceding phoneme. This minimal amount of context modelling is somewhat compensated by the language model, where a 4-gram model can achieve about 40% word accuracy when compared to a uni-gram model which achieves 0.5% word accuracy in the same experiment [3].

In grapheme-to-phoneme conversion using HMMs, the graphemes are the observations while the phonemes are the hidden states. The grapheme-to-phoneme system searches for the most likely sequences of phonemes  $s$  out of all phoneme sequences in the language given some input graphemes  $O$ . This can be expressed as:

$$S = \underset{s}{\operatorname{argmax}} P(S|O) \quad (8)$$

$$\begin{aligned} &= \underset{s}{\operatorname{argmax}} \frac{P(O|S)P(S)}{P(O)} \\ &= \underset{s}{\operatorname{argmax}} \underbrace{P(O|S)}_{\text{likelihood}} \underbrace{P(S)}_{\text{language model}} \end{aligned}$$

The observation  $O$  is the concatenation of a successive number of graphemes and  $S$  is the concatenation of successive

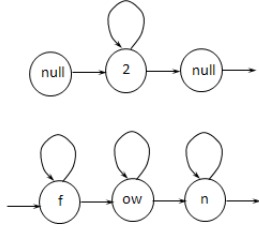


Figure 3: The prototype HMM and a composite model for the word ‘phone’.

phonemes. Since equation 8 is being maximised,  $P(O)$  is constant and the same for all  $s$ . Thus, the problem reduces to finding the phoneme sequences  $s$  with the highest  $P(O|S)P(S)$ .  $P(O|S)$  is the observation likelihood. It is computed using HMMs.  $P(S)$  is the prior probability of phoneme sequences commonly realised using  $n$ -gram language models.

## 5. The proposed approach

This section presents the proposed approach in the grapheme-to-phoneme conversion problem. Two dictionaries were used in this paper: the Carnegie Mellon University (CMU) pronunciation dictionary (CMUDICT) of North American English and the Unilex dictionary of received pronunciation UK English. Both the CMUDICT and the Unilex dictionaries contain over 100,000 words along with their phonetic transcriptions. After splitting the dictionaries into training and test sets, the CMUDICT sets contain 92,474 train and 10,274 test entries respectively. These entries are phonetically represented by a phone set of 40 phones. The Unilex training and test sets contain 101,414 and 11,268 entries respectively. Entries in the Unilex sets are phonetically represented by 50 phone labels.

The approach presented in this paper differs from [3] in the following ways. In [3], each phoneme model contains 4 states. Context is modelled with respect to phonemes, which is the standard technique. Furthermore, there is a pre-processing stage which rewrites some words to generate new grapheme sequences. Finally, there is the stress adjustment stage whereby a separate stress prediction model is trained. In the approach presented in this paper, each phoneme model contains 1 state. Context is modelled with respect to graphemes and phonemes using context-sensitive graphemes and context-dependent phoneme models. Finally, stress information is not included yet.

The upper part of figure 3 shows the prototype model while the lower part shows a composite model for the word ‘phone’. The prototype model for all phonemes is defined as a 1-state left-right topology whereby each state could go to itself or the end state. This is based on the topology presented in [3] as each phoneme can generate one or more letters, for example, in the word ‘bought’ which is realised as ‘/b aa t/’, the phoneme ‘aa’ generates the letters ‘ough’. The null states do not emit observations.

Six different systems are presented in section 6. The proposed system is the sixth, which combines context-sensitive (CS) and context-dependent triphone (CD) models with a 4-gram language model. The other systems and configurations have been included to show the effects of contexts and language models. The first system is the baseline with  $M$  equal to 26,

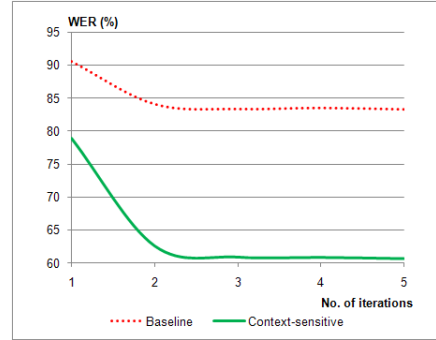


Figure 4: WER as a function of the number of training iterations for the baseline and context-sensitive models.

which represents the 26 distinct letters in English alphabet. The second is the baseline with CD models and a bigram language model. The third is the baseline with CD models and an  $n$ -best list rescored with a 4-gram language model. The fourth is the model with CS observations and a bigram language model.  $M$  equal to 8, 582 and 7,477 (the number of context-sensitive graphemes in the CMU and Unilex dictionaries respectively). The fifth combines contexts at the observation as well as the hidden states and a bigram language model while the sixth rescoring an  $n$ -best list with a 4-gram language model. During rescoring, 20 best hypotheses are generated for each word and are rescored with a 4-gram language model to generate a single best hypothesis. All experiments are carried out with the HTK toolkit [7].

## 6. Evaluation results

This section presents the result of the evaluation of the grapheme-to-phoneme system using HMMs. The criterion used is the word accuracy.

Table 1: Performance on CMUDICT dictionary in %.

Type	M	Grammar	WA
Baseline	26	bigram	19.86
Baseline + CD	26	bigram	24.90
Baseline + CD	26	4-gram	37.93
CS	8582	bigram	40.71
CS + CD	8582	bigram	50.27
CS + CD	8582	4-gram	57.85

Table 2: Performance on Unilex dictionary in %.

Type	M	Grammar	WA
Baseline	26	bigram	16.65
Baseline + CD	26	bigram	30.89
Baseline + CD	26	4-gram	53.12
CS	7477	bigram	39.38
CS + CD	7477	bigram	68.38
CS + CD	7477	4-gram	79.19

Tables 1 and 2 show the performance of the grapheme-to-phoneme system on both dictionaries. **WA** is the word accuracy which marks a word as correct if all the phonemes are correctly recognised. Word accuracy is more important than phoneme accuracy for this task. Figure 4 shows the word error rate (WER)

on the Unilex test set as a function of the number of training iterations for the baseline ( $M$  equal 26) and context-sensitive models ( $M$  equal 7477). The short dash line and the straight line are the baseline and context-sensitive models respectively. The WER decreases rapidly with the number of iterations for the context-sensitive models as compared with the baseline.

## 7. Discussion

Context can be modelled at the observations or the hidden states or both, depending on the availability of training data. Modelling at the observations is less sensitive to the effect of data scarcity as only the number of codebook indices,  $M$ , increases as compared with the large number of parameters to be estimated with context-dependent phoneme models. Furthermore, it is easier to model contexts at the observations for discrete HMMs as the knowledge of different classes of the hidden states is not required. For example, in part-of-speech tagging, there are no natural classes to tie together unlike in grapheme-to-phoneme conversion or speech recognition where the hidden states (phonemes) can be classified based on distinctive phonetic features (articulation, voicing, etc). Context modelling at the observations is closely related to increasing the number of Gaussian mixtures in continuous density HMMs; Gaussian models with more mixtures perform better than single mixture models. The number of codebook indices in discrete HMMs is analogous to the number of Gaussian mixtures in continuous density HMMs.

An increase in  $M$  improves the performance of the model because this takes into account the different contexts of each grapheme. The rescored triphone model with  $M$  equal to 8582 and 7477 give the best performance. This is because 20 hypotheses generated with a bigram language model are rescored with a 4-gram language model to select the best hypothesis. A 4-gram language model is more powerful than a bigram language model. Furthermore, the performance on Unilex is far better than the CMUDICT dictionary. This is because Unilex is more consistent than CMUDICT; CMUDICT contains many foreign words as well as errors.

One of the drawbacks of using HMMs for grapheme-to-phoneme conversion is that HMMs require the number of graphemes to be equal to or greater than the number of phonemes [3]. However, some graphemes are generated by two phoneme sequences. For example, the grapheme 'x' is almost always realised as the phoneme pairs /k/ /s/, /k/ /sh/, /g/ /z/, /g/ /zh/, when the left context is a vowel. A suggested method of dealing with this is to merge the phoneme pairs to form a new phoneme, for example, /k/ /s/ are merged to give /ks/.

The number of graphemes used to model the context can be increased which also increases  $M$ . This depends on the amount of training data available.

In order to train higher-order language models, larger dictionaries than used in this paper are required. Even when using linguistic classes to reduce the number of models (i.e. vowel/consonant information), typical HMM approaches treat each attribute of context (be it in language model or observation labels) as equally significant. Decision trees do not have this issue as trees' attributes are questioned individually, such that at any point only the attributes deemed relevant during the training process have some influence on the outcome. One disadvantage to requiring a large dictionary is that they do not exist for many languages, and in situations where they do, they reduce the importance of automatic grapheme-to-phoneme approaches as systems can use the large dictionary for most words.

## 8. Conclusions and future work

This paper presented HMMs that emit context-sensitive observations. The underlying concept was modelling context at the observations for discrete HMMs. This involved modelling the different contexts in which each observation symbol could be. The effect of this was an increase in the number of distinct observation symbols. Context modelling at both domains (hidden states and observations) is data intensive but the former is more sensitive to the amount of training data because a large number of parameters need to be estimated. Furthermore, the problem of data scarcity and unseen contexts were addressed by tying, whereby contexts that were tied shared the same codebook index (parametrisation).

An experiment on using HMMs with context-sensitive observations for grapheme-to-phoneme conversion yielded a result of 57.85% and 79.19% on a held-out test set of CMUDICT and Unilex respectively.

Future work will address the development of an efficient tying algorithm as well as adaptation of this work to deal with long distance dependences in symbol sequence recognition.

## 9. Acknowledgments

This material is based upon works supported by the Science Foundation Ireland under Grant No. 07/CE/I1142 as part of the Centre for Next Generation Localisation ([www.cngl.ie](http://www.cngl.ie)) at University College Dublin. The opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Science Foundation Ireland.

## 10. References

- [1] Rabiner, L., "A tutorial on hidden Markov models and selected applications in speech recognition", In Proceedings of the IEEE(77), 257-286, 1989.
- [2] Bilmes, J., "What HMMs can do", Journal of the IEICE Transactions on Information and Systems, 89-D(3) 257-286, 2006.
- [3] Taylor, P., "Hidden Markov model for grapheme to phoneme conversion", In Proceedings of INTERSPEECH, 1973-1976, 2005.
- [4] Jiampojarn, S. and Kondrak, G., "Online discriminative training for grapheme-to-phoneme conversion", In Proceedings of INTERSPEECH, 1303-1306, 2009.
- [5] Young, S. J. and Philip, C. W., "The use of state tying in continuous speech recognition", In Proceedings of EUROSPEECH, 2203-2206, 1993.
- [6] Young, S. J., James J. O. and Philip C. W., "Tree-based state tying for high accuracy acoustic modelling", In Proceedings of the workshop on Human Language Technology, 307-312, 1994.
- [7] Young, S. J. et al., "The HTK Book for HTK V3.4.", Cambridge University Press, Cambridge UK, 2006.
- [8] Wellekens, C., "Explicit time correlation in hidden Markov models for speech recognition", In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 384-386, 1987.
- [9] Carnegie Mellon University, "Carnegie Mellon University pronouncing dictionary CMUdict 0.7a", <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- [10] Polyakova, T. and Antonio, B., "Using error-driven approach to improve automatic grapheme-to-phoneme conversion accuracy", In Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, Barcelona Spain.