IDIAP RESEARCH REPORT

A study of phoneme and grapheme based context-dependent ASR systems

John Dines [a] Mathew Magimai Doss [a]

IDIAP–RR 07-12

March 2007

[a] IDIAP Research Institute and Ecole Polytechnique Federale de Lausanne (EPFL), Martigny, Switzerland

IDIAP Research Institute www.idiap.ch
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 07-12

A study of phoneme and grapheme based context-dependent ASR systems

John Dines Mathew Magimai Doss

March 2007

Abstract. In this paper we present a study of automatic speech recognition systems using context-dependent phonemes and graphemes as sub-word units based on the conventional HMM/GMM system as well as tandem system. Experimental studies conducted on three different continuous speech recognition tasks show that systems using only context-dependent graphemes can yield competitive performance on small to medium vocabulary tasks when compared to a context-dependent phoneme-based automatic speech recognition system. In particular, we demonstrate the utility of tandem features that use an MLP trained to estimate phoneme posterior probabilities in improving grapheme based recognition system performance by incorporating phonemic knowledge into the system without having to explicitly define a phonetically transcribed lexicon.

1 Introduction

State-of-the art automatic speech recognition (ASR) systems represent words as a sequence of sub-word units, typically phonemes which have a strong correlation with the acoustic observations. In recent studies, attention has been drawn toward speech recognition systems using grapheme as sub-word units [14, 7, 8, 10]. The main advantages of using grapheme as sub-word units are (1) the definition of lexicon is easy (orthographic transcription), and (2) the pronunciation models are relatively noise free. The main drawback of using graphemes as sub-word units is that a single grapheme can map onto many different phonemes, i.e. there is often a weak correspondence between graphemes and phonemes, particularly in the English language.

Schukat-Talamazzaini et al. were one of the first to present results in speech recognition based on graphemes [14]. They used "polygraph" sub-word units for word modelling, which is essentially letters-in-context similar to polyphones (phonemic units allowing preceding and following context of arbitrary length). Experimental studies conducted on continuous speech recognition task and isolated word recognition showed that good results (better than context-independent phone) can be obtained using "polygraph" as sub-word units.

In a recent study, the approach of mapping orthographic transcription to a phonetic one has been investigated in the context of speech recognition [7]. In this approach, the orthographic transcription of the words are used to map them onto acoustic hidden Markov model (HMM) state models using phonetically motivated decision tree questions. For instance, a grapheme is assigned to a phonetic question if the grapheme maps to the phoneme. Recognition studies performed on Dutch, German and English yielded performances comparable to phoneme-based ASR system for languages Dutch and German and, fairly poor performance for English language.

Killer et al. have investigated a context dependent grapheme based speech recognition, where the context is modelled through a decision tree based clustering procedure [8]. Experimental studies conducted on English, German and Spanish languages yielded competitive results compared to phoneme-based system for German and Spanish languages, but fairly poor performance for the English language.

In [10, 9], we proposed a phoneme-grapheme based system that jointly models the both phoneme and grapheme sub-word units during training. During decoding, recognition is done either using one or both sub-word units. This system was investigated in the framework of hybrid hidden Markov model/artificial neural network (HMM/ANN) system and improvements were obtained over a context-independent phoneme based system using both sub-word units in recognition on two different tasks isolated word recognition task [10] and recognition of numbers task [9].

In this paper, we present a study of context-dependent phonemes and graphemes as sub-word for English ASR systems. On three tasks of increasingly complexity: OGI Numbers95 (NU95) [4], DARPA resource management (RM) [13] and continuous telephone speech (CTS) [2] and using different features (standard PLP cepstral feature and tandem feature), we analyse the use of grapheme as sub-word units for English ASR by comparing it with the standard phoneme based system. Our studies show that on tasks of smaller complexity such as NU95 the grapheme based ASR system can perform as good as the phoneme based ASR system. At the same time, on tasks of increased complexity such as RM and CTS the performance difference between the two systems, phoneme based system and grapheme based system, becomes more pronounced with the phoneme based system being the better one. Our studies also show that on these tasks of increased complexity the difference between the two systems is greatly reduced when using tandem features.

2 Background

Lexical representations play a critical role in ASR. In all but the most constrained tasks, it is necessary to represent words by a sequence of sub-word units (the so called 'beads-on-a-string' paradigm), in order to give a compact representation of the lexicon that still provides good correspondence between the words and acoustical observations. Most commonly, sub-word units take the form of phonemes, as they are limited in number (of the order of 45 for English) and show good correspondence with the acoustic observations. One disadvantage of the use of phonemes is that mapping from words to phonemes is generally a knowledge driven process, which is difficult to automate with a high level of fidelity, thus making it an expensive process in terms of development time and effort. Automatic means for deriving pronunciations in text-to-speech synthesis exist in order to enable such systems to handle out-of-vocabulary text, but generally such mechanisms are not employed in ASR. Interestingly enough, if we examine two of the most commonly used techniques in letter-to-sound mapping and in ASR lexicon representations, we see that context-dependency plays a critical role. In this section, we describe the use of context dependent sub-word units in letter-to-sound mapping and acoustic modelling for ASR, drawing attention to the similarities between the two. We also briefly describe the tandem acoustic features, which feature significantly in our studies.

2.1 Letter-to-sound mapping using decision trees

In text-to-speech synthesis it is often necessary to produce pronunciations for words that lie outside the pronunciation dictionary of the system. Such systems employ letter-to-sound mapping techniques to automatically generate pronunciations. A commonly used approach to this problem is to use decision trees [1]. The decision tree approach is carried out by first aligning grapheme and phoneme symbols from a pronunciation dictionary that is to be used for training[1]. For each grapheme occurrence the graphemes surrounding it (a context window of N to the left and right) are recorded as well as the phoneme which has been aligned to the grapheme. The decision tree is trained from this data by pooling all of the instances of a particular grapheme together then successively splitting the data according to the grapheme context that gives rise to the largest decrease in leaf node impurity (entropy times number of sample points). By building a decision tree in this manner a set of rules is derived that use a grapheme's context to determine its pronunciation.

2.2 Context dependent modelling of sub-word units

Word pronunciations can differ greatly from their lexical form, in particular due to the effects of coarticulation, making it common practice to explicitly model each sub-word unit according to the context in which it occurs. Due to the limitations of data coverage and decoding complexity, a single phone context to the left and right (the 'triphone') is generally used. Even then, a large quantity of data is required in order to independently learn the statistics of each context dependent unit, hence, a parameter sharing scheme is needed. The most commonly employed parameter sharing scheme is the decision tree-tying approach [12], which pools all of the data for a particular sub-word unit into a single root node and performs tree growth by selecting questions at each split that maximise the increase in likelihood of the acoustic models over the training data. The decision tree approach not only achieves more robust modelling of seen contexts, but also enables the synthesis of unseen contexts.

The questions used to split the data may be singleton (each question relates to only a single sub-word unit), knowledge based (eg. phonemic: "is the left phone context a VOWEL") or data driven [3]. In general, the knowledge based approach is used as it gives both good data utilisation and generalisation, in particular for unseen contexts, but clearly, for grapheme based systems, only singleton and data driven can be used. Killer et. al. [8] explored different approaches to question set derivation for context-dependent grapheme based speech recognition and demonstrated that, in fact, the singleton questions sets gave best results, though with the disadvantage of more inefficient data utilisation compared to data-driven approaches. It should also be noted that context dependent modelling of grapheme-based sub-word units displays strong similarities with the letter-to-sound mapping described in the previous section, since we are learning a mapping from the graphemic representation to the acoustic feature space, which is much more strongly correlated with the phonemic representation.

[1] Extra steps need to be taken to deal with words which have fewer/more phonemes than graphemes.

2.3 Tandem acoustic features

Tandem systems have been shown to yield state-of-the-art performance [5]. A tandem system combines the discriminative feature of an ANN with Gaussian mixture modelling by using the processed posterior probabilities generated by the MLP as the input feature for the HMM/GMM-based system. It has been demonstrated that tandem features exhibit greater robustness to unwanted variabilities [16, 6]. This is due to the ability of the ANN to project the standard acoustic feature on dimensions carrying information most pertinent to the speech recognition task.

A tandem based system can also be viewed as a cascade of classifiers, thus, permitting the integration of decisions made in an earlier classification stage into later stages. Tandem acoustic features are of interest in this study as they present a means of introducing phonetic knowledge into a grapheme based system through the use of an MLP trained on phonemic targets without the need for explicit specification of a phonemic pronunciation dictionary (though phonemic targets are still required for the training of the MLP, we can assume that these can be obtained from a corpus where phonetic transcriptions are available, rather than the target training corpus).

3 Empirical studies

3.1 Experimental setup

Our studies were conducted on three well known speech corpora that comprise tasks of varying complexity with regard to training data, lexicon and language model. The major features of each corpora are listed in Table 1, highlighting their respective.

Acoustic models were trained for the three corpora using the hidden Markov model toolkit (HTK) from both PLP and tandem-features [15] . In each case, the acoustic models were trained through: 8 iterations of re-estimation on context-independent models, 2 iterations of re-estimation on context-dependent models followed by model tying, 7 iterations of re-estimation on tied context-dependent models and finally increment of mixtures from 1 to 8 in multiples of two with 3 iterations of re-estimation at each increment step. In these studies we investigated singleton, knowledge-based and data driven question sets for state tying. We used a fixed log-likelihood threshold to control decision tree growth, thus models were allowed to achieve differing levels of complexity based on the sub-word units, features, and question sets used. In comparing two systems, this enabled us to consider how the choice of sub-word units, features, and question sets influenced model complexity and ultimately ASR performance.

PLP feature extraction comprised 13th order PLP cepstral coefficients and their deltas and delta-deltas. The features were computed every 10ms over a window of 30 ms. For the tandem-features, an MLP was trained on the PLP features with output units corresponding to context-independent phonemes. The phoneme targets for MLP training were derived from a forced alignment of the training data using the PLP based acoustic models. We extracted the tandem-features using the MLP's phoneme log-posterior estimates followed by Karhunen-Loeve transformation. In the grapheme dictionary, the numbers and abbreviated words were replaced by their graphemic representation eg. 45 $\Rightarrow$ FOURTY FIVE.

3.2 OGI numbers95

The OGI numbers95 (NU95) database comprises a limited vocabulary task that employs a word-loop language model, for which we used the definition of the training set, validation set, and test set is similar to the one defined in [11]. For the purposes of investigating different lexical representations, this is a very simple task. In comparing the ASR systems produced from context dependent phoneme and grapheme models shown in Table 2 we can see that the complexity of the acoustic models is

[2]Meaning that the same words appear in train and test data.
[3]This is provided with the preamble of the DARPA Resource Management word-pair grammar.

Table 1: Summary of the three corpora used in our studies. CI: context independent, CD: content dependent

Name	Component	Description	Statistic
OGI Numbers95	Audio data	Quantity of data	
		Train:	90 mins
		Test:	30 mins
	Lexicon	Closed[2]	
		Words:	31
		Phoneme (CI/CD):	24/81
		Graphemes (CI/CD):	19/85
	Acoustic model	Word internal, context dependent	
	Language model	Wordloop	
DARPA RM	Audio data	Quantity of data	
		Train:	3.8 hrs
		Test:	1.1 hrs
	Lexicon	Closed	
		Words:	991
		Phonemes (CI/CD):	42/2269
		Graphemes (CI/CD):	29/1912
	Acoustic model	Word internal, context dependent	
	Language model	Wordpair	
CTS	Audio data	Quantity of data	
		Train:	32 hrs
		Test:	1.3 hrs
	Lexicon	Open	
		Words:	1000
		Phonemes (CI/CD):	47/20k
		Graphemes (CI/CD):	36/9k
	Acoustic model	Cross-word, context dependent	
	Language model	Bigram	

quite similar with the grapheme system having slightly more models/states than its phoneme based counterpart. This is reflected in the overall performance of the grapheme system, which has slightly lower error rates than the phoneme system. The tandem based systems had the same performance on this task, this being significantly better than that obtained from PLP features. While these results suggest that phoneme and grapheme system can achieve equivalent performance, it is clear that this is because both the context dependent grapheme and phoneme acoustic models have an almost one-is-to-one mapping to their corresponding lexical entry.

Table 2: ASR results on OGI Numbers95 task

Unit	Feature	Quest	Log. Models	Phy. Models	Log. States	Phys. States	WER (in %)
Phoneme	PLP	Phonemic	81	74	241	191	6.3
	Tandem	Phonemic	81	74	241	193	4.4
Grapheme	PLP	Singleton	85	79	256	198	5.9
	Tandem	Singleton	85	78	256	196	4.4

3.3 DARPA resource management

We next performed ASR evaluations on the DARPA resource management (RM) corpus. This corpus is also of relatively low complexity compared to state-of-the-art evaluation tasks, but is still quite a step up from the OGI numbers task. In particular, the lexicon is greatly increased from 31 to almost 1000, thus context dependent models may no longer have a unique mapping to a single word. The lexicon is still closed, thus it is not necessary for the acoustic models to generalise to words not seen in training, nor is it necessary to synthesise unseen contexts.

The results from the experiments on the RM corpus are shown in Table 3. We extended analysis on the RM corpus in order to better compare the systems by building systems using both singleton and data driven questions sets (according to [3]). We only report the results for singleton questions sets here as the data driven approach was not found to provide any more useful insight for this study.

Table 3: ASR results on DARPA resource management task

Unit	Feature	Quest	Log. Models	Phy. Models	Log. States	Phys. States	WER (in %)
Phoneme	PLP	Singleton	2269	1501	6729	1477	5.7
	Tandem	Singleton	2269	1628	6729	2013	5.7
Grapheme	PLP	Singleton	1912	1298	5727	1369	7.3
	Tandem	Singleton	1912	1360	5727	1985	6.3
Merged	PLP	Singleton	4181	2799	12456	2846	5.5
	Tandem	Singleton	4181	2988	12456	3998	5.1

A number of observations can be made from these results. In particular we can note that for both PLP and tandem systems the number of physical states in the grapheme and phoneme systems is roughly equivalent, despite there being fewer actual (logical) states for the grapheme system. This demonstrating that the decision tree growth for grapheme based models needs to be deeper (more questions) in order to disambiguate the one-to-many mapping associated with graphemes.

In comparing the PLP and tandem feature based systems we see that tandem features provide a significant improvement for the grapheme based system, although for this task it still remains to some degree behind that of the phoneme based system. We also observe that tandem based features lead to a greater number of states, mostly likely due to there being less unwanted variability in the tandem features. This is particularly important for the grapheme system where co-articulatory effects further complicate the task of learning the feature space relationship with the context dependent grapheme models.

In a last test we also merged phoneme and grapheme acoustic models and lexica (without retraining), thus enabling a mixture of grapheme and phoneme based models to be used during recognition. We see that this gives a slight improvement over both phoneme and grapheme systems, suggesting that the grapheme models, while giving overall inferior performance to the phoneme system, still manage to achieve some degree of complementarity. I.e. grapheme modelling is not just an inferior alternative to phoneme modelling. Further analysis performed using the merged models and dictionaries on the development set of DARPA RM task showed that grapheme models were more preferred for function words which short in terms of length (number of graphemes).

3.4 Conversational telephone speech

The final evaluation carried out as part of this study was with the conversational telephone speech (CTS) corpus. This corpus is significantly more complex than those previously described in that although the lexicon is of similar size to that used in RM, it is now open (meaning that words may appear in testing that do not appear during training). Furthermore, the acoustic conditions are significantly more challenging as the audio is taken from a telephone channel. In training the context-dependent models on the CTS corpus, we made one change to the training procedure, which was to allow for cross-word context dependency. Due to the increased complexity of the task we have only conducted limited investigations on the CTS task, namely the male part of the corpus. The results are detailed in Table 4.

Table 4: Preliminary ASR results on male part of the CTS task

Unit	Feature	Quest	Log. Models	Phy. Models	Log. States	Phys. States	WER (in %)
Phoneme	PLP	Phonemic	20810	5601	62430	1325	45.7
	Tandem	Phonemic	20640	7370	61920	1786	45.3
Grapheme	PLP	Singleton	9309	4435	27927	2602	53.0
	Tandem	Singleton	9278	4125	27834	2885	50.3

One of the first points that stands out from these results is the discrepancy between the number of logical models and physical states in the phoneme and grapheme systems. The phoneme system has twice the number of logical models (by virtue of the fact that there are more graphemes than phonemes), but conversely half as many physical states. This is partly due to the fact that the singleton question set will naturally lead to deeper decision trees, but can also be attributed to the greater complexity required in modelling context-dependent grapheme models. This is consistent with the findings in Black et. al. [1], who demonstrated that using an early stopping criterion to prevent over-fitting of decision tree learning of letter-to-sound mappings was actually detrimental to performance.

Further observations from these results may also be noted. First of all, once again the tandem features appear to provide some improvement in both phoneme and grapheme systems, particularly in the grapheme case. Unfortunately though, the grapheme tandem system still lags significantly behind the phoneme system. This can be attributed to a number of factors. The use of cross-word context dependent models made the grapheme based system significantly disadvantaged in that cross-word contexts are likely to be counter-productive for letter-to-sound mapping. In addition the open nature of the vocabulary demands that the grapheme based system be able to generalise to unseen words and contexts, which is likely to be considerably more challenging than for the phoneme system. While these issues could be addressed to some extent by (for example) the use of special symbols to disambiguate word internal and cross-word contexts the problem of generalisation may not be easily solved (and at the least may require significantly more training data than for the phoneme based system).

4 Conclusions

In this paper we have studied the use of context-dependent phonemes and graphemes as sub-word units for automatic speech recognition. ASR studies conducted on different tasks show that by using context-dependent graphemes as sub-word units, performance similar to the state-of-the-art context-dependent phoneme based ASR system can be achieved on constrained tasks. Analysis demonstrates that the contextual modelling of grapheme units gives behaviour similar to phonemes and is achieved in a similar fashion to that observed in letter-to-sound mapping techniques.

In OGI Numbers95 studies we obtained better performance using graphemes when the acoustic models were trained with PLP features and similar performance when trained with tandem features. In the DARPA RM task studies we observed a marked difference between ASR systems using phoneme and grapheme when trained with PLP features. However, this difference is reduced when using tandem features. An explanation for this can be that the tandem system is able to incorporate phonetic knowledge while still having no requirement for an explicit phonetic lexicon. In the much more complex CTS task we also observed improvements thanks to tandem features, though not to the same extent to that observed on the simpler tasks. These observations are summarised in Table 5

In both OGI Numbers95 task and DARPA RM task the words that are present in the dictionary are present in both training data and test data. In other words, there were no unseen contexts unlike

Table 5: Summary of findings from our studies. $\approx$ means comparable, $\uparrow/\downarrow$ means somewhat greater/reduced, $\Uparrow/\Downarrow$ means significantly greater/reduced

Lexicon	Cross-word Modelling	System Complexity	Phoneme-Grapheme Correspondence	Performance (grapheme)	Tandem vs PLP
small (closed)	no	$\approx$	$\approx$	$\approx$	$\uparrow$
medium (closed)	no	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$
medium (open)	yes	$\Uparrow$	$\Downarrow$	$\Downarrow$	$\uparrow$

in the CTS task. It is likely that this played a large role in the significantly reduced performance of the grapheme based CTS system compared with the phoneme based system. Further research will need to look at how to overcome this either through improved parameter sharing approaches or by drawing upon non-acoustical data such as existing pronunciation lexica (which may not provide full coverage of the acoustic training data). It may also be interesting to look a wider sub-word unit context in the framework of either WFST based decoding or lattice rescoring.

We also carried out an experiment on the RM corpus in which we merged grapheme and phoneme models and lexica and showed improved performance over either system alone. This suggests that the grapheme based models are complimentary to the phoneme models. In order further validate this hypothesis on a more challenging task such as the CTS, is clear that there are a number of hurdles that would first need to be overcome. Firstly, the use of cross-word models would require that we merge models in a less naive fashion as the current approach does not support cross-word contexts between phoneme and grapheme systems.

5 Acknowledgements

This work was supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811) and the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2).

References

[1] Alan W Black, Kevin Lenzo, and Vincent Pagel. Issues in building general letter to sound rules. In *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, pages 77–80, Jenolan Caves, Australia, 1998.

[2] B. Chen, Ö. Çetin, G. Doddington, N. Morgan, M. Ostendorf, T. Shinozaki, and Q. Zhu. A CTS task for meaningful fast-turnaround experiments. In *Proceedings of Rich Transcription Fall Workshop*, Palisades, NY, 2004.

[3] C. Ciprian and R. Morton. Mutual information phone clustering for decision tree induction. In *ICSLP*, Denver, Collorado, 2002.

[4] R. A. Cole, M. Fanty, M. Noel, and T. Lander. Telephone speech corpus development at CSLU. In *Proceedings of Int. Conf. Spoken Language Processing (ICSLP)*, 1994.

[5] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature stream extraction for conventional HMM systems. In *Proceedings of Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages III–1635–1638, 2000.

[6] S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, and H. Bourlard. Entropy based combination of tandem representations for robust speech recognition. In *Proceedings of Int. Conf. Spoken Language Processing (ICSLP)*, Korea, October 2004.

[7] S. Kanthak and H. Ney. Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In *Proceedings of Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 845–848, 2002.

[8] M. Killer, S. Stüker, and T. Schultz. Grapheme based speech recognition. In *Proceedings of Eurospeech*, pages 3141–3144, 2003.

[9] M. Magimai.-Doss, S. Bengio, and H. Bourlard. Joint decoding for phoneme-grapheme continuous speech recognition. In *Proceedings of Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages I–177–I–180, 2004.

[10] M. Magimai.-Doss, T. A. Stephenson, H. Bourlard, and S. Bengio. Phoneme-Grapheme based automatic speech recognition system. In *Proceedings of Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 94–98, 2003.

[11] N. Mirghafori and N. Morgan. Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers. In *Proceedings of Int. Conf. Spoken Language Processing*, pages 743–746, 1998.

[12] Julian James Odell. *The use of context in large vocabulary continuous speech recognition.* PhD thesis, Queens College, University of Cambridge, 1995.

[13] P. J. Price, W. Fisher, and J. Bernstein. A database for continuous speech recognition in a 1000 word domain. In *Proceedings of Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 1:651–654, 1988.

[14] E. G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Automatic speech recognition without phonemes. In *Eurospeech*, pages 129–132, 1993.

[15] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. Hidden Markov model toolkit V3.2.1 reference manual. Technical report, Speech group, Engineering Department, Cambridge University, UK, March 2002.

[16] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke. On using MLP features in lvcsr. In *Proceedings of Int. Conf. Spoken Language Processing (ICSLP)*, Korea, October 2004.