IDIAP RESEARCH REPORT

Improving Continuous Speech Recognition System Performance with Grapheme Modelling

Mathew Magimai.-Doss [a b] John Dines [a]
Hervé Bourlard [a b] Hynek Hermansky [a]

IDIAP–RR 05-16

April 2005

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[a] IDIAP Research Institute, CH-1920 Martigny, Switzerland
[b] École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

IDIAP Research Report 05-16

Improving Continuous Speech Recognition System Performance with Grapheme Modelling

Mathew Magimai.-Doss John Dines Hervé Bourlard
Hynek Hermansky

April 2005

submitted for publication

Abstract. This paper investigates automatic speech recognition system using context-dependent graphemes as subword units based on the conventional HMM/GMM system as well as TANDEM system. Experimental studies conducted on two different continuous speech recognition tasks show that systems using only context-dependent graphemes can yield competitive performance when compared to state-of-the-art context-dependent phoneme-based automatic speech recognition system. We further demonstrate that a system using both context-dependent phoneme and grapheme subword units can out perform either of these systems alone.

1 Introduction

State-of-the art automatic speech recognition (ASR) systems represent words as a sequence of subword units, typically phonemes which are more strongly correlated with the acoustic observation. In recent studies, attention has been drawn toward speech recognition systems using grapheme as subword units [STNE+93, KN02, KSS03, MDSBB03]. The main advantages of using grapheme as subword units are (1) the definition of lexicon is easy (orthographic transcription), (2) the pronunciation models are relatively noise free. The main drawback of using graphemes as subword units is that a single grapheme can map on to different phonemes, i.e. there is a weak correspondence between graphemes and phonemes, particularly in English language.

Schukat-Talamazzaini et al. were one of the first who presented results in speech recognition based on graphemes [STNE+93]. They used "polygraph" as subword 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 subword units.

In a recent study, the approach of mapping orthographic transcription to a phonetic one has been investigated in the context of speech recognition [KN02]. 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 [KSS03]. 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 [MDSBB03, MDBB04], we proposed a phoneme-grapheme based system that jointly models the both phoneme and grapheme subword units during training. During decoding, recognition is done either using one or both subword 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 subword units in recognition on two different tasks isolated word recognition task [MDSBB03] and recognition of numbers task [MDBB04].

In later work we studied the use of context-dependent graphemes for the numbers task (for which there are only a limited number of word internal contexts) [MDDBH04]. We investigated a system with context-independent phonemes and context-dependent graphemes. During recognition, the phoneme information was marginalized out and the decoding was performed in the context-dependent grapheme space. This system performed similar to the context-independent phoneme based system. This motivated further investigation into using context-dependent graphemes as subword units in state-of-the-art systems.

In this paper, we present our studies using context-dependent graphemes as subword units on OGI Numbers95 task and DARPA resource management (RM) task. The paper is organized as follows. In Section 2, we describe our studies on OGI Numbers95 task. In Section 3, we present our analysis on the behaviour of context-dependent graphemes. Section 4 describes the studies on DARPA RM corpus. Finally, in Section 5 we conclude with discussion and future work.

2 Context-Dependent Grapheme Studies on OGI Numbers95 Corpus

We use OGI Numbers95 database for connected word recognition task [CFNL94]. The training set contains 3233 utterances spoken by different speakers and the validation set consists of 357 utterances. The test set contains 1206 utterances. The vocabulary consists of 31 words with a single pronunciation for each word.

The acoustic vector comprises of PLP cepstral coefficients [Her90] extracted from the speech signal using a window of 25 ms with a shift of 12.5 ms, followed by cepstral mean subtraction. At each time frame, 13 PLP cepstral coefficients $c_0 \cdots c_{12}$ and their first-order and second-order derivatives are extracted, resulting in 39 dimensional acoustic vector.

There are 24 context-independent phonemes and 80 context-dependent phonemes including silence and, 19 context-independent graphemes and 85 context-dependent graphemes including silence representing the characters in the orthographic transcription of the words.

We trained HMM/Gaussian mixture models (GMMs) system with 80 context-dependent phonemes, 3 emitting states per phoneme and 12 mixtures per state with PLP feature vectors using HTK toolkit [YOO+97] (*GMM-P*). We also trained HMM/GMM system with 85 context-dependent graphemes, 3 emitting states per phoneme and 12 mixtures per state with PLP feature vectors (*GMM-G*). In addition to this, we trained context-dependent phoneme based hybrid HMM/ANN system (*ANN-P*) and context-dependent grapheme based hybrid HMM/ANN system (*ANN-G*). The performances of these systems are given in Table 2. The results show that the systems using context-dependent graphemes perform better than their context-dependent phoneme counterparts.

System	Subword Unit	WER
GMM-P	Phoneme	6.8
GMM-G	Grapheme	**6.0**
ANN-P	Phoneme	6.9
ANN-G	Grapheme	6.3

Table 1: Performance of different context-dependent subword units systems on OGI Numbers95 Database. The performance is expressed in terms of word error rate (WER) and expressed in %.

TANDEM systems have been shown to yield state-of-the-art performance [HES00]. A TANDEM system combines the discriminative feature of an ANN with Gaussian mixture modelling by using the processed posterior probabilities as the input feature for the HMM/GMMs-based system. Hence, to further validate our results, we obtained tandem features from a trained MLP (used earlier in our context-independent phoneme studies) with 24 context-independent phonemes as output and trained two TANDEM systems, one with context-dependent grapheme units (*Tandem-CD-G*) and the second with context-dependent phoneme units (*Tandem-CD-P*), with the same configurations of *GMM-CD-G* and *GMM-CD-P*, respectively. The results are given in Table 2. It can be seen from the results that both systems yield similar performance.

System	Subword Unit	WER
Tandem-P	Phoneme	**4.9**
Tandem-G	Grapheme	5.1

Table 2: Performance of TANDEM system using different context-dependent subword units on OGI Numbers 95 database. The performance is expressed in terms of word error rate (WER) and expressed in %.

3 Analysis

In the previous section we noted that the context-dependent grapheme-based ASR system performance is similar to the performance of state-of-the-art context-dependent phoneme-based ASR system. However, this not true with ASR system using context-independent graphemes as subword units (17% WER with PLP features). Hence, in order to further understand the effect of contextual modelling in grapheme-based ASR system we performed some contextual modelling studies.

We trained systems with only preceding context and only following context. The number of preceding-context and following-context phonemes were 81 and 71 (including short pause model in HTK), respectively. The number of preceding-context and following-context graphemes were 75 and 68, respectively. All the systems were trained using HTK toolkit with 3 emitting states per subword unit and 12 mixtures per state. The results of this study are given in Table 3. The results indicate that the effect of modelling context in grapheme-based system is similar to that of modelling context in phoneme-based system. Moreover, the results also suggest that context-dependent grapheme units behave like phoneme units.

Subword unit	Context	Feature	WER
Phoneme	Following	PLP	9.1
Phoneme	Preceding	PLP	13.5
Grapheme	Following	PLP	9.6
Grapheme	Preceding	PLP	14.1
Phoneme	Following	TANDEM	5.2
Phoneme	Preceding	TANDEM	6.8
Grapheme	Following	TANDEM	6.6
Grapheme	Preceding	TANDEM	9.5

Table 3: Results of contextual modelling studies on OGI Numbers95 database. The performance is expressed in terms of Word Error Rate (WER) and expressed in %.

The main idea behind modelling context in phoneme-based ASR is to capture coarticulation effects; where as in grapheme-based system, our studies suggest that by modelling context we can expect to jointly model co-articulatory effects and pronunciation variation. However, the numbers task has only 31 words and it can be expected that there is a singular mapping between the context-dependent grapheme and context-dependent phoneme. Hence, we extended this study the context-dependent grapheme-based ASR studies to DARPA RM corpus which has a vocabulary size of 997 words.

4 Context-Dependent Grapheme Studies on DARPA RM Corpus

The DARPA RM corpus consists of read queries on the status of Naval resources [PFB88]. The task is artificial in many aspects such as speech type, range of vocabulary and grammatical constraint. The training set consists of 3,990 utterances spoken by 109 speakers corresponding to approximately 3.8 hours of speech. We use 2,880 utterances for training and 1,100 for cross validation and development. The test set contains 1,200 utterances amounting to 1.1 hours in total. The test set is evaluated using a word pair grammar which is included in the task specification. There are 44 phonemes and 29 graphemes including silence. The feature vector comprises 13th order PLP cepstral coefficients, their deltas and delta-deltas. The features were computed every 10ms over a window of 30 ms. In the grapheme dictionary, the numbers and abbreviated words were replaced by their graphemic representation.

We trained a HMM/GMM system with context-dependent phoneme acoustic models and a HMM/GMM system with context-dependent graphemes acoustic models. The system was trained

using HTK toolkit [YOO+97]. The acoustic models were trained through: 8 iterations of reestimation on context-independent models, 2 iterations of reestimation on context-dependent models followed by model tying, 7 iterations of reestimation on tied context-dependent models and finally increment of mixtures from 1 to 8 in multiples of two with 3 iterations of reestimation at each increment step. The question set for tying consisted of singleton questions about left and right contexts. The number of context-dependent phoneme and grapheme models after tying are 2294 and 1912, respectively. The recognition results are given in Table 4.

Subword Unit	WER
Phoneme	**7.6**
Grapheme	10.2

Table 4: Recognition performance of HMM/GMM system trained on DARPA resource management corpus with context-dependent phoneme acoustic models and context-dependent grapheme acoustic models. The performance is measured in terms of word error rate (WER) expressed in %.

We further studied the above systems in the framework of TANDEM systems. We trained an MLP with 44 output units corresponding to the context-independent phonemes. We extracted the tandem-features using the phoneme posterior estimates generated by this MLP followed by Karhunen-Loeve transformation and, then trained HMM/GMM system with these features with the same configuration as that of PLP HMM/GMM system. The recognition results for the different systems are given in Table 5. We observe that TANDEM system performs better than the PLP-based HMM/GMM system for both type of subword units. Also, the amount of gain for grapheme-based system (2.8% absolute) is much greater than the phoneme-based system (0.8% absolute) making the two systems more comparable.

Subword Unit	WER
Phoneme	**6.8**
Grapheme	7.4

Table 5: Recognition performance of TANDEM system trained on DARPA resource management corpus with context-dependent phoneme acoustic models and context-dependent grapheme acoustic models. Performance is measured in terms of word error rate (WER) and expressed in %.

In our earlier phoneme-grapheme studies, the decoding is done in the 2D state space [MDSBB03, MDBB04], but in large vocabulary systems with context-dependent acoustic models this is an expensive computation. One way to combine the information from these two different subword units would be to decode in each individual space and then combine the recognized word sequences. Another way would be to merge the two acoustic models and dictionaries and perform decoding in a standard way. In this way the best acoustic model representation of the word is chosen at the decoding time. In this paper the later approach was taken. We performed recognition studies by merging the acoustic models (without retraining) and the two dictionaries of the phoneme and grapheme units. The results of this study are given in Table 6, showing that the merged model system performs the best compared to the systems reported earlier in this paper.

System	WER
HMM/GMM	7.4
TANDEM	**6.4**

Table 6: Recognition performance of HMM/GMM system using PLP features and TANDEM system trained on DARPA resource management corpus with merged acoustic models and dictionaries. The performance is measured in terms of word error rate (WER) and expressed in %.

Further analysis performed using the merged models and dictionaries on the development set of DARPA RM task showed that grapheme is more preferred when the word is a function word and short in terms of length (number of graphemes).

5 Summary and Conclusion

In this paper we have studied the use of context-dependent graphemes as subword units for automatic speech recognition. ASR studies conducted on different tasks show that by using context-dependent graphemes as subword units, performance similar to the state-of-the-art context-dependent phoneme based ASR system can be achieved. Analysis demonstrates that the contextual modelling of grapheme units gives behaviour similar to phonemes.

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 through discriminative tandem features while still having no requirement for an explicit phonetic lexicon. Moreover, we also observed that the performance of the the ASR system can be further improved by merging the acoustic models and dictionaries of phoneme and grapheme units.

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 otherwords, there were no unseen contexts. Hence, in future work we are interested in extending these studies to less constrained task such as switchboard conversational telephone speech to study how the grapheme-based system generalizes to unseen contexts.

6 Acknowledgment

This work was supported by the Swiss National Science Foundation (NSF) under grant MULTI (2000-068231.02/1) and Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. The NCCR is managed by the Swiss NSF on behalf of the federal authorities. This work is also supported by EU 6th FWPIST integrated project AMI (FP6-506811).

References

[CFNL94] R. A. Cole, M. Fanty, M. Noel, and T. Lander. Telephone speech corpus development at CSLU. In *ICSLP*, 1994.

[Her90] H. Hermansky. Perceptual linear predictive(PLP) analysis of speech. *JASA*, 87(4):1738–1752, 1990.

[HES00] H. Hermansky, D. Ellis, and S. Sharma. Tandem connectionist feature stream extraction for conventional HMM systems. In *ICASSP*, pages III–1635–1638, 2000.

[KN02] S. Kanthak and H. Ney. Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In *ICASSP*, pages 845–848, 2002.

[KSS03] M. Killer, S. Stüker, and T. Schultz. Grapheme based speech recognition. In *Eurospeech*, pages 3141–3144, 2003.

[MDBB04] M. Magimai.-Doss, S. Bengio, and H. Bourlard. Joint decoding for phoneme-grapheme continuous speech recognition. In *ICASSP*, pages I–177–I–180, 2004.

[MDDBH04] M. Magimai.-Doss, J. Dines, H. Bourlard, and H. Hermansky. Phoneme vs grapheme based automatic speech recognition. IDIAP-RR 48, IDIAP, 2004.

[MDSBB03] M. Magimai.-Doss, T. A. Stephenson, H. Bourlard, and S. Bengio. Phoneme-Grapheme based automatic speech recognition system. In *ASRU*, pages 94–98, 2003.

[PFB88] P. J. Price, W. Fisher, and J. Bernstein. A database for continuous speech recognition in a 1000 word domain. In *ICASSP*, pages 1:651–654, 1988.

[STNE+93] E. G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck. Automatic speech recognition without phonemes. In *Eurospeech*, pages 129–132, 1993.

[YOO+97] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. Hidden Markov model toolkit V2.1 reference manual. Technical report, Speech group, Engineering Department, Cambridge University, UK, March 1997.