IDIAP RESEARCH REPORT

Phoneme vs Grapheme Based Automatic Speech Recognition

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

IDIAP–RR 04-48

September 2004

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] Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland

IDIAP Research Report 04-48

Phoneme vs Grapheme Based Automatic Speech Recognition

Mathew Magimai.-Doss John Dines Hervé Bourlard
Hynek Hermansky

September 2004

submitted for publication

Abstract. In recent literature, different approaches have been proposed to use graphemes as subword units with implicit source of phoneme information for automatic speech recognition. The major advantage of using graphemes as subword units is that the definition of lexicon is easy. In previous studies, results comparable to phoneme-based automatic speech recognition systems have been reported using context-independent graphemes or context-dependent graphemes with decision trees. In this paper, we study both context-independent and context-dependent grapheme-based automatic speech recognition systems. Experimental studies conducted on American English continuous speech recognition task show that systems using context-independent grapheme units perform fairly poor, while their performance can be improved by incorporating phonetic knowledge. However, systems using only context-dependent graphemes can yield competitive performance (even better) when compared to state-of-the-art phoneme-based automatic speech recognition.

1 Introduction

State-of-the art automatic speech recognition (ASR) systems represent words as a sequence of subword units, typically phonemes. 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 associate itself 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 HMM state models using phonetically motivated decision tree questions, for instance, a grapheme is assigned to a phonetic question if the grapheme is part of the phoneme. This approach however, makes the modelling process complex.

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 English language.

In [MDSBB03, MDBB04], we proposed a phoneme-grapheme based system that jointly models the phoneme subword units 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. One of the primary difficulty with this approach is training an ANN with KxR (≈ 1200) output units where K is the number of context-independent phoneme units and R is the number of context-independent grapheme units. In this paper, we first reformulate this approach in Section 2, where instead of training one ANN with KxR output units we train two ANNs with K and R output units, respectively (as used in [BMWR92] for context-dependent phoneme modelling). We have tested this on OGI Numbers database and have obtained performance similar compared to those reported in [MDBB04]. Since OGI Numbers corpus has a relatively smaller number of context-dependent grapheme units, we studied context-dependent grapheme system with-in the frame work of the new formulation. This system, using only context-dependent grapheme units, performs better than the context-independent phoneme system.

This motivated us to look into using context-dependent grapheme units in a standard way similar to context-dependent phoneme units for ASR. Our experimental studies show that systems using context-dependent graphemes as subword units can yield competitive performance (even better) when compared to state-of-the-art phoneme-based automatic speech recognition system. We describe these studies in Section 3. Section 4 provides a discussion. Finally, Section 5 concludes with future work.

2 Phoneme-Grapheme Based ASR System

The phoneme-grapheme system models the evolution of hidden phoneme space $Q = \{q_1, \cdots, q_n, \cdots q_N\}$ and grapheme space $L = \{l_1, \cdots, l_n, \cdots l_N\}$ and the observed space $X = \{x_1, \cdots, x_n, \cdots x_N\}$ as

$$p(Q, L, X) \approx \prod_{n=1}^{N} p(x_n|q_n, l_n)P(q_n|q_{n-1})P(l_n|l_{n-1}) \tag{1}$$

where $q_n \in \mathcal{Q} = \{1, \cdots, k, \cdots, K\}$ and $l_n \in \mathcal{L} = \{1, \cdots, r, \cdots, R\}$. In the system proposed in [MDSBB03, MDBB04], $p(x_n|q_n, l_n)$ is estimated as

$$\frac{p(x_n|q_n = k, l_n = r)}{p(x_n)} = \frac{P(q_n = k, l_n = r|x_n)}{P(q_n = k, l_n = r)} \tag{2}$$

where $P(q_n = k, l_n = r|x_n)$ is the output of ANN with KxR output units. Since Kx$R \approx 1200$ or more, training an ANN with such large number of output units poses a problem. Instead the $P(q_n = k, l_n = r|x_n)$ can be estimated in the following manner

$$P(q_n = k, l_n = r|x_n) = P(l_n = r|x_n, q_n = k)P(q_n = k|x_n) \tag{3}$$

i.e. training an ANN with R output units to estimate $P(l_n = r|x_n, q_n = k)$ and an ANN with K units to estimate $P(q_n = k|x_n)$. Equation (3) can be marginalized to estimate $P(l_n = r|x_n)$

$$P(l_n = r|x_n) = \sum_{k=1}^{k=K} P(q_n = k, l_n = r|x_n) \tag{4}$$

which can be then scaled by its prior to obtain the scaled-likelihood [BM94], and used as the emission probability to decode in the grapheme space.

2.1 Experimental Setup

We use OGI Numbers 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 x_n is the 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 $c_0 \cdots c_{12}$, their first-order and second-order derivatives are extracted, resulting in 39 dimensional acoustic vector. All the ANNs trained in our studies take nine frames input feature (4 frames of left and right context, each) and have the same number of parameters.

There are 24 context-independent phonemes including silence associated with $\mathcal{Q}$. There are 19 context-independent grapheme subword units including silence associated with $\mathcal{L}$ representing the characters in the orthographic transcription of the words.

2.2 Experimental Studies

We trained a context-independent phoneme hybrid HMM/ANN baseline system (*System P*) via embedded Viterbi training [BM94] and performed recognition using a single pronunciation of each word. The performance of the phoneme baseline system is given in Table 1. We trained a context-independent grapheme hybrid HMM/ANN baseline system (*System G*) via embedded Viterbi training and performed recognition experiments using the orthographic transcription of the words. The performance of the grapheme baseline system is given in Table 1. As reported earlier, this system performs considerably poor compared to *System P*.

We trained an ANN with R output units to estimate $P(l_n = r|x_n, q_n = k)$ in (3). In addition to PLP features this ANN has phoneme information as inputs. We provide posteriors obtained from *System P* as phoneme information for the contextual frames, except for the center frame. During training, the center frame information is defined based on the knowledge of phoneme segmentation and during recognition, we define the center frame for all possible phonemes i.e. perform K forward passes and sum all the probabilities as in (4) to obtain $P(l_n|x_n)$. The $P(l_n|x_n)$ is then transformed into scaled-likelihood and used as emission probability to perform decoding in the context-independent grapheme space. We refer to this system as *System CI-G*. The performance of this system is given in Table 1. The performance of grapheme-based recognizer improves; but still it is fairly poor compared to *System P*. Similar to previously reported studies [MDBB04], introduction of phonetic knowledge improves the performance of grapheme-based ASR.

We also performed decoding in context-independent phoneme-grapheme space with the emission probability $p(x_n|q_n, l_n)$ estimated from $P(q_n|x_n)$ (obtained from *System P*) and $P(l_n|x_n)$ (obtained from *System CI-G*) according to Equation (10) in [MDBB04]. We refer to this system as *System CI-PG*. The results of this study is given in Table 1. This system performs better than *System P*. A similar trend was observed in our earlier studies [MDBB04].

The OGI Numbers lexicon contains only 80 context-dependent[1] phonemes and 85 context-dependent graphemes, hence it is feasible to train hybrid HMM/ANN systems with outputs corresponding to each context dependent target. We trained an hybrid HMM/ANN with 80 output units corresponding to context-dependent phonemes (*ANN-CD-P*). This system performs better than any system given in Table 1.

We trained an hybrid HMM/ANN system in the lines of *System CI-G*; but with 85 context-dependent grapheme units as the output of ANN. We refer to this system as *System CD-G*. We performed recognition just using grapheme subword units similar to *System CI-G*. The performance of this system is given in Table 1. This system yields performance comparative to (*System P*). This is quite interesting as this recognizer is *purely* grapheme based. In order to understand the contextual modelling in grapheme-based ASR system, we conducted further studies using context-dependent grapheme units. This study is reported in the next section.

Table 1: Performance of different systems as described in Section 2.2. The performance is expressed in terms of Word Error Rate (WER).

System	WER
System P	9.1%
System G	17.3%
System CI-G	13.7%
System CI-PG	8.3%
ANN-CD-P	7.7%
System CD-G	8.9%

3 Context-Dependent Grapheme Based ASR System

We trained HMM/Gaussian mixture models (GMMs) system with 80 context-dependent phonemes, 3 emitting states per phoneme and 12 mixtures per state with 39 dimensional PLP feature vector using HTK toolkit [YOO+97] (*GMM-CD-P*). We trained HMM/GMMs system with 85 context-dependent graphemes, 3 emitting states per phoneme and 12 mixtures per state with 39 dimensional PLP feature vector (*GMM-CD-G*). In addition to this, we trained a context-dependent grapheme hybrid

[1] unless specified it is both preceding and following context

HMM/ANN system with 39 dimensional PLP feature vector (*ANN-CD-G*). The performances of these systems are given in Table 2.

Table 2: Performance of different context-dependent subword units systems. The performance is expressed in terms of Word Error Rate (WER).

System	WER
GMM-CD-P	7.3%
GMM CD-G	6.0%
ANN-CD-G	6.3%
Tandem-CD-G	5.1%
Tandem-CD-P	5.1%

As it could be seen from Table 2 that both HMM/GMMs system and hybrid HMM/ANN system using only context-dependent grapheme units perform significantly better (McNemar's test) than their context-dependent phoneme counterparts *GMM-CD-P* and *ANN-CD-P*.

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 using *System P* 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 is quite interesting to note that the context-dependent grapheme-based ASR system using PLP features yields close to state-of-the-art performance, where as, the context-independent grapheme system (*System G*) performs very poor.

In order to further understand the effect of contextual modelling in grapheme based ASR, we trained systems with only preceeding context and only following context. The number of preceeding-context and following-context phonemes were 81 and 71 (including short pause model in HTK), respectively. The number of preceeding-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.

Table 3: Results of contextual modelling studies. The performance is expressed in terms of Word Error Rate (WER).

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%

4 Discussion

This work began with a reformulation of the approach proposed in [MDBB04] to jointly model context-independent phoneme units and grapheme units in ASR system; but became a question of comparing pure modelling systems when it became apparent that modelling context-dependent grapheme (*System CD-G*) leads to improvement in the performance of ASR. One of the key difference between context-dependent grapheme and context-dependent phoneme is that noisy phoneme transcription is relied upon for the phoneme-based system. Also, the main idea behind modelling context in phoneme-based ASR is to capture the influence of phonemes on each other; where as in grapheme-based system, our studies suggest that by modelling context we jointly model co-articulatory effects and pronunciation variation. This could be the possible reasons why there is a significant difference between the performance of systems *GMM-CD-P* and *GMM-CD-G*, and systems *ANN-CD-P* and *ANN-CD-G*. The tandem system is able to handle the noise in the phoneme transcriptions (possibly due to discriminative features) and yields state-of-the-art performance for both type of context-dependent subword units.

Further experiments conducted to understand the effect of contextual modelling in grapheme-based ASR shows that the context-dependent graphemes have similar behavior compared to context-dependent phonemes. This could be possibly the reason why *System CD-G* yields relatively lower performance compared to *ANN-CD-G*; as we are feeding the phoneme information from *System P* as additional input to *System CD-G*, which could be noisy. We have restricted the context in our studies to maximum one preceeding grapheme and one following grapheme. In [STNE+93], it has been observed that increasing the context window helps in improving the performance; but on the contrary in [KSS03], it has been observed that longer context windows does not leads to improvement in the performance. The issue of context window length will be part of our future work.

5 Conclusion and Future Work

In this paper, we have carried out a detailed analysis of phoneme and grapheme modelling schemes in ASR and show that within an appropriate modelling framework a pure grapheme-based ASR approach has the potential to outperform its phoneme-based counter part. We have also demonstrated state-of-the art performance in context-dependent phoneme and context-dependent grapheme systems on OGI Numbers task.

Motivated by this encouraging result, future work will attempt to apply this approach to tasks of increasing vocabulary size. It is expected that for such tasks, non-singular mappings from context-dependent grapheme targets to phoneme-targets together with ever increasing number of these targets will pose a great challenge; but the advantages of implicit pronunciation modelling will also become more apparent.

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

[BM94] H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach.* Kluwer Academic Publishers, 1994.

[BMWR92] H. Bourlard, N. Morgan, C. Wooters, and S. Renals. CDNN: A context dependent neural network for continuous speech recognition. In *ICASSP*, pages II–349–II–352, 1992.

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

[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.

[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.