JOINT DECODING FOR PHONEME-GRAPHEME CONTINUOUS SPEECH RECOGNITION

Mathew Magimai.-Doss, Samy Bengio, and Hervé Bourlard

Dalle Molle Institute for Artificial Intelligence, CH-1920, Martigny, Switzerland
Swiss Federal Institute of Technology (EPFL), CH-1015, Lausanne, Switzerland

ABSTRACT

Standard ASR systems typically use phoneme as the subword units. Preliminary studies have shown that the performance of the ASR system could be improved by using grapheme as additional subword units. In this paper, we investigate such a system where the word models are defined in terms of two different subword units, i.e., phoneme and grapheme. During training, models for both the subword units are trained, and then during recognition either both or just one subword unit is used. We have studied this system for a continuous speech recognition task in American English language. Our studies show that grapheme information used along with phoneme information improves the performance of ASR.

1. INTRODUCTION

State-of-the-art HMM-based Automatic Speech Recognition (ASR) systems model $p(Q,X)$, the evolution of the hidden space $Q = \{q_1, \cdots, q_n, \cdots q_N\}$ and the observed feature space $X = \{x_1, \cdots, x_n, \cdots x_N\}$ also denoted X_1^N over time frame $1, \cdots, N$ [1]. The states represent the subword units which describe the word model. Standard ASR typically use phoneme as subword units. In recent studies, good results have been reported using grapheme as subword units.

There are certain advantages in using graphemes as subword unit such as, the word models could be easily derived from the orthographic transcription of the word and it is relatively "noise free" as compared to word models based upon phoneme units, for e.g. the word $ZERO$ can be pronounced as /z/ /ih/ /r/ /ow/ or /z/ /iy/ /r/ /ow/; but the grapheme-based representation remains as $[Z][E][R][O]$. At the same time there are certain drawbacks, such as, acoustic feature vectors derived from the smoothed spectral envelope of the speech signal typically depict the characteristics of phonemes and there is a weak correspondence between the graphemes and the phonemes in languages such as English, e.g., the grapheme $[E]$ in word $ZERO$ associates itself to phoneme /ih/, where as, in word $EIGHT$ it associates itself to phoneme /ey/. In [2], 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, however, makes the modelling process complex.

In [3], the approach to model grapheme is similar to modelling auxiliary information [4, 5]. The grapheme is treated as an auxiliary information $L = \{l_1, \cdots, l_n, \cdots l_N\}$ and the evolution of the hidden spaces Q and L over X is modelled (i.e. $p(Q,X,L)$ instead of $p(Q,X)$). This system could be seen as a system where word models are described by two different subword units, the phonemes and the graphemes. During training, models are trained for both the subword units maximizing the likelihood of the training data. During recognition, decoding is performed using either one or both the subword units. This system is similar to factorial HMMs [6], where there are several chains of states as opposed to a single chain in standard HMMs. Each chain has its own states and dynamics; but the observation at any time depends upon the current state in all the chains. In [3], the preliminary studies conducted on isolated word recognition task showed that the performance of the ASR could be improved by using phoneme and grapheme subword units together.

In this paper, we further investigate the system proposed in [3] in the context of continuous speech recognition task. In Section 2, we briefly describe the system we are investigating. Section 3 presents the experimental studies. Finally in Section 4, we summarize and conclude with future work.

2. MODELLING PHONEME GRAPHEME

Standard ASR models $p(Q,X)$ as

$$p(Q,X) \approx \prod_{n=1}^{N} p(x_n|q_n) \cdot P(q_n|q_{n-1}) \qquad (1)$$

where $q_n \in \mathcal{Q} = \{1, \cdots, k, \cdots, K\}$, represents the phoneme.

Similarly for a system with L as the hidden space we model

$$p(L,X) \approx \prod_{n=1}^{N} p(x_n|l_n) \cdot P(l_n|l_{n-1}) \qquad (2)$$

where $l_n \in \mathcal{L} = \{1, \cdots, r, \cdots, R\}$, the grapheme set.

We are interested in an ASR where the word models are described by two different subword units, and hence, interested in modelling the evolution of two hidden spaces Q and L (instead of just one) and the observed space X over time i.e. $p(Q,L,X)$

$$p(Q,L,X) \approx \prod_{n=1}^{N} p(x_n|q_n,l_n) P(q_n|q_{n-1}) P(l_n|l_{n-1}) \qquad (3)$$

For such a system, the forward recurrence can be written as:

$$\begin{aligned}\alpha(n,k,r) &= p(q_n = k, l_n = r, X_1^n) \\ &= p(x_n|q_n = k, l_n = r) \sum_{i=1}^{K} P(q_n = k|q_{n-1} = i) \\ &\sum_{j=1}^{R} P(l_n = r|l_{n-1} = j) \;\; \alpha(n-1,i,j) \qquad (4)\end{aligned}$$

During recognition, we decode in the joint phoneme and grapheme spaces. The Viterbi decoding algorithm that gives the best sequence in the Q and L spaces, can be written as:

$$\begin{aligned}V(n,k,r) &= p(x_n|q_n = k, l_n = r) \max_i P(q_n = k|q_{n-1} = i) \\ &\max_j P(l_n = r|l_{n-1} = j) \;\; V(n-1,i,j) \qquad (5)\end{aligned}$$

For task such as isolated word recognition studied in [3], this decoding step could be reduced to two independent decoding steps in Q and L space, respectively; but for continuous speech recognition we need to perform 2D decoding as described above.

We are investigating the proposed system in the framework of hybrid HMM/ANN ASR [7]. In hybrid HMM/ANN ASR, during training a Multilayer Perceptron (MLP) is trained say, with K output units for system in (1). The likelihood estimate is replaced by the scaled-likelihood estimate which is computed from the output of the MLP (posterior estimates) and priors of the output units (hand counting). For instance, $p(x_n|q_n)$ in (1) is replaced by its scaled-likelihood estimate $p_{sl}(x_n|q_n)$, which is estimated as [7]:

$$p_{sl}(x_n|q_n) = \frac{p(x_n|q_n)}{p(x_n)} = \frac{P(q_n|x_n)}{P(q_n)} \qquad (6)$$

The emission distribution $p(x_n|q_n = k, l_n = r)$ of the phoneme-grapheme system could be estimated in different ways, such as, we could train an MLP with $K \times R$ output units and estimate the scaled-likelihood as

$$\frac{p(x_n|q_n = k, l_n = r)}{p(x_n)} = \frac{P(q_n = k, l_n = r|x_n)}{P(q_n = k, l_n = r)} \qquad (7)$$

Such a system, during training would automatically, model the association between the subword units in Q and L. This system has an added advantage that it could be reduced to a single hidden variable system by marginalizing any one of the hidden variables, yielding:

$$\frac{p(x_n|q_n = k)}{p(x_n)} = \frac{\sum_{j=1}^{R} P(q_n = k, l_n = j|x_n)}{P(q_n = k)} \qquad (8)$$

$$\frac{p(x_n|l_n = r)}{p(x_n)} = \frac{\sum_{i=1}^{K} P(q_n = i, l_n = r|x_n)}{P(l_n = r)} \qquad (9)$$

and using this scaled-likelihood estimate to decode according to (1) or (2), respectively.

Yet another approach would be to assume independence between the two hidden variables Q and L, train two separate systems one for each phoneme and grapheme, and estimate the scaled-likelihood as following:

$$\frac{p(x_n|q_n = k, l_n = r)}{p(x_n)} \approx p_{sl}(x_n|q_n = k) p_{sl}(x_n|l_n = r) \quad (10)$$

In [3], the phoneme-grapheme system studies were conducted in the lines of (10). In this paper, we will be investigating phoneme-grapheme systems in the lines of both (7) and (10).

3. EXPERIMENTAL STUDIES

Standard ASR typically use phonemes as subword units. The lexicon of an ASR contains the orthographic transcription of the word and its phonetic transcription. During decoding, standard ASR uses the phonetic transcription only, ignoring the orthographic transcription. In this paper, we are mainly interested in investigating the use of the orthographic information for automatic speech recognition.

We use OGI Numbers database for connected word recognition task [8]. 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 single pronunciation for each word.

The acoustic vector x_n is the PLP cepstral coefficients [9] 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 MLPs 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 $\mathcal{Q}$, each modelled by a single emitting state. We trained a phoneme baseline system (**System P**) via embedded Viterbi training [7] and performed recognition using single pronunciation of each word. The performance of the phoneme baseline system is given in Table 1.

There are 19 context-independent grapheme subword units including silence associated with $\mathcal{L}$ representing the characters in the orthographic transcription of the words. Similar to phonemes each of the grapheme units are modelled by a single emitting state. We trained a grapheme 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. The phoneme baseline system performs significantly better than the grapheme baseline system.

Table 1. Performance of phoneme and grapheme baseline systems. The performance is expressed in terms of Word Error Rate (WER).

System	# of output units	WER
System P	24	9.6%
System G	19	17.8%

As suggested in Section 2, we study the two approaches to model phoneme and grapheme subword units,

(a) Modelling the phoneme and grapheme subword units with a single MLP.

(b) Modelling the phoneme and grapheme subword units through separate MLPs.

For (a), we trained an MLP with $24 \times 19 = 456$ output units (**System PG**). During training, at each iteration, we marginalize out the phoneme information as per (9) and perform Viterbi decoding according to (2) to get the segmentation in terms of graphemes. We performed recognition experiments by marginalizing the grapheme subword units according to (8) and decoding according to (1), and similarly we performed recognition experiments by marginalizing the phoneme subword units according to (9) and decoding according to (2). The performances are given in Table 2. There is no improvement in the performance of the phoneme system; but there is a significant improvement in the performance of the grapheme system.

We also studied systems where the phoneme is reduced to its broad-phonetic-class representation. By broad-phonetic-class, we refer to the phonetic features, such as manner, place, height. In our studies, we use the phonetic feature values similar to the one used in [10, Chapter 7] and [3]. The mapping between the phonemes and the values of the broad-phonetic-class could be obtained from a *International Phonetic Alphabet (IPA) chart*.

Table 2. Performance of phoneme-only and grapheme-only system by marginalizing (hide) grapheme and phoneme, respectively, at the output of phoneme-grapheme MLP (**System PG**). The performance is expressed in terms of Word Error Rate (WER).

Subword Unit	Subword Unit Hidden	WER
Phoneme	Grapheme	9.6%
Grapheme	Phoneme	14.5%

We studied three different grapheme-broad-phonetic-class systems corresponding to the different broad-phonetic-classes, (1) manner (**System GBM**), (2) place (**System GBP**) and (3) height (**System GBH**). We train acoustic models for both grapheme units and values of the broad-phonetic-class by training a single MLP via embedded Viterbi training, similar to the phoneme-grapheme MLP. We performed recognition studies:

1. Marginalizing the broad-phonetic-class according to (9) and performing decoding just using grapheme transcription.

2. Performing 2D decoding in the grapheme and broad-phonetic-class space according to (5).

Table 3 presents the experimental results of this study. The grapheme systems perform significantly better than the grapheme baseline system; but the grapheme-broad-phonetic-class system performs significantly better than all the grapheme systems.

Table 3. Performance of grapheme-broad-phonetic-class based system. The performance is expressed in terms of Word Error Rate (WER). The results of the grapheme systems (Graph) are given in column 3 and of the grapheme-broad-phonetic-class systems (GB) in column 4.

System	Broad-phonetic class	WER Graph	WER GB
System G	-	17.8%	-
System GBM	Manner	15.3%	13.1%
System GBP	Place	14.4%	11.9%
System GBH	Height	15.0%	11.7%

Next, we study the performance of the phoneme-grapheme system. As mentioned earlier in this paper, we study two different kinds of system.

(a) Modelling the phoneme and grapheme subword units through single MLP. For such a system the scaled-likelihood is estimated as per (7) from the posterior output of the MLP and the decoding is performed according to (5) (**System PG**).

(b) Assuming independence between the phoneme units and the grapheme units, i.e., modelling them through different MLPs. The scaled-likelihood $p(x_n|q_n, l_n)$ is then obtained from the scaled-likelihood estimate of phoneme units and grapheme units according to (10) and the decoding is performed according to(5). In [3], the best results were obtained by weighting the $\log$ probability streams of phoneme and grapheme differently. However, in this paper we estimate $p(x_n|q_n, l_n)$ exactly according to (10).

The results of this study are given in Table 4. The first row contains the performance of the phoneme baseline system. The second row contains the performance of **System PG**. This system performs slightly poorer compared to the baseline system. The remaining rows are the results obtained for phoneme-grapheme system where the phoneme units and grapheme units are modelled by different MLPs. These systems perform better than the baseline system.

Table 4. Performance of phoneme-grapheme system. Columns 1 and 2 indicate from which of the MLPs the phoneme and grapheme scaled-likelihood estimates were estimated, respectively for the system where the independence between phoneme units and grapheme units is assumed. The performance is expressed in terms of Word Error Rate (WER).

Phoneme	Grapheme	WER
System P	-	9.6%
System PG	**System PG**	9.9%
System P	**System PG**	9.0%
System P	**System GBM**	9.0%
System P	**System GBP**	8.9%
System P	**System GBH**	9.2%

4. CONCLUSION AND FUTURE WORK

In this paper, we investigated a continuous speech recognizer which uses both phoneme and grapheme as subword units. ASR using just grapheme as subword unit yields acceptable performance, which could be further improved by introducing phonetic knowledge in it.

We studied two different phoneme-grapheme systems. The results obtained from phoneme-grapheme system studies suggest that modelling phoneme and grapheme subword units, and using them together during recognition could help in improving the performance of ASR. This has to be further studied for large vocabulary continuous speech recognition task.

In this paper, our primary focus was upon using the grapheme information at model level. In future work, we intend to investigate combining hypotheses generated by separate phoneme and grapheme recognizers.

In languages such as English, there is a weak correspondence between the graphemes and the phonemes. So, it would be worth investigating this approach for languages such as German which has strong correspondence between the graphemes and the phonemes.

5. 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.

6. REFERENCES

[1] L. R. Rabiner and H. W. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs New Jersey, 1993.

[2] S. Kanthak and H. Ney, “Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition,” in *ICASSP*, 2002, pp. 845–848.

[3] M. Magimai.-Doss, T. A. Stephenson, H. Bourlard, and S. Bengio, “Phoneme-Grapheme based automatic speech recognition system,” in *ASRU*, December 2003.

[4] T. A. Stephenson, M. Magimai.-Doss, and H. Bourlard, “Speech recognition with auxiliary information,” *To appear in IEEE Trans. Speech and Audio Processing*, 2003.

[5] M. Magimai.-Doss, T. A. Stephenson, and H. Bourlard, “Using pitch frequency information in speech recognition,” in *Eurospeech*, September 2003, pp. 2525–2528.

[6] Z. Ghahramani and M. I. Jordan, “Factorial hidden Markov models,” *Machine Learning*, vol. 29, pp. 245–273, 1997.

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

[8] R. A. Cole, M. Fanty, and T. Lander, “Telephone speech corpus development at CSLU,” in *ICLSP*, September 1994.

[9] H. Hermansky, “Perceptual linear predictive(PLP) analysis of speech,” *JASA*, vol. 87, no. 4, pp. 1738–1752, 1990.

[10] J.-P Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*, PhD dissertation, CSLU, OGI, USA, 2000.