PHONEME-GRAPHEME BASED SPEECH RECOGNITION SYSTEM

Mathew Magimai.-Doss, Todd A. Stephenson, Hervé Bourlard, and Samy Bengio

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

ABSTRACT

State-of-the-art ASR systems typically use phoneme as the subword units. In this paper, we investigate a system where the word models are defined in-terms of two different subword units, i.e., phonemes and graphemes. We train models for both the subword units, and then perform decoding using either both or just one subword unit. We have studied this system for American English language where there is weak correspondence between the grapheme and phoneme. The results from our studies show that there is good potential in using grapheme as auxiliary subword units.

1. INTRODUCTION

State-of-the-art HMM-based 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 $X = \{x_1, \cdots, x_n, \cdots x_N\}$ over time frame $1, \cdots, N$ [1]. The states represent the subword units (typically, phonemes) which describe the word model. The feature vectors are typically derived from the smoothed spectral envelope of the speech signal. In recent studies, it has been proposed that modelling the evolution of auxiliary information $L = \{l_1, \cdots, l_n, \cdots l_N\}$ along with Q and X (i.e. $p(Q, X, L)$ instead of $p(Q, X)$) could improve the performance of ASR [2]. The auxiliary information that were mainly investigated in the past are the additional features obtained from the speech signal such as pitch frequency, short-time energy, rate-of-speech etc [3]. In these studies, the auxiliary information has been observed throughout the training similar to X; but during recognition it has been either observed or hidden.

In this paper, we extend this strategy of modelling auxiliary information to model an information which is hidden both during training and recognition similar to Q. Basically, this system could be seen as a system where word models are described by two different subword units, the phonemes and the graphemes[1]. During training, we train models for both the subword units maximizing the likelihood of the training data. During recognition, we perform decoding using either one or both the subword units. This system is similar to factorial HMMs [4], 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. One of the first attempts in this direction has focussed upon dividing the states itself into chains for task such as phoneme recognition, which did not yield significant results [5]. In our case instead of dividing states representing the same subword units into chains, there are two chains corresponding to each of the subword units.

In literature, good results have been reported using graphemes as subword units [6]. The main advantage of using graphemes is that the word models could be defined easily (orthographic transcription) and it is relatively "noise free" as compared to word models based upon phoneme units, for e.g. the word COW can be pronounced as /k/ /o/ /v/ or /k/ /ae/ /v/; but the grapheme-based representation remains as $[C][O][W]$. At the same time, there are drawbacks in using graphemes too, such as, there is a weak correspondence between the graphemes and the phonemes in languages such as English, e.g., the grapheme $[C]$ in the case of the word CAT associates itself to phoneme /k/, where as, in the case of the word $CHURCH$ it associates itself to phoneme /C/. Furthermore, the acoustic feature vectors typically depict the characteristics of phonemes. In [6], this problem was handled by using a decision tree based, graphemic acoustic subword units with phonetic questions. This, however, makes the acoustic modelling process complex. As we will see in the later sections, the proposed system provides an easy approach to model relationship between two different subword units automatically from the data.

We study the proposed system in the framework of state-of-the-art hybrid HMM/ANN system [7], which provides some additional flexibility in modelling and estimation. 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.

[1] a written symbol that is used to represent speech. For e.g. alphabets in English language.

2. MODELLING AUXILIARY INFORMATION

Standard ASR models $p(Q, X)$[2] as

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

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

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

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

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

In this paper, we are 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. $p(Q, L, X)$. For such a system, the forward recurrence can be written as:

$$\begin{aligned}\alpha(n, k, r) &= p(q_n = k, l_n = r, x_n) \\ &= p(x_n|q_n{=}k, l_n{=}r) \sum_{i=1}^{K} P(q_n{=}k|q_{n-1}{=}i) \\ &\sum_{j=1}^{R} P(l_n = r|l_{n-1} = j) \;\; \alpha(n-1, i, j) \qquad (3)\end{aligned}$$

assuming conditional independence between Q and L given x_n.

The likelihood of the data can then be estimated as

$$p(X) = \sum_{k=1}^{K} \sum_{r=1}^{R} \alpha(N, k, r) \qquad (4)$$

Finally, the Viterbi decoding algorithm that gives the best sequence in the Q and L spaces, can be written as

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

In state-of-the-art ASR, the emission distribution could be modelled by Gaussian Mixture Models (GMM) or Artificial Neural Network (ANN). In case of 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]:

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

We are investigating the proposed system in the framework of hybrid HMM/ANN ASR, where the emission distribution $p(x_n|q_n = k, l_n = r)$ 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 and estimating the scaled-likelihood as following:

$$\begin{aligned}\frac{p(x_n|q_n = k, l_n = r)}{p(x_n)} &\approx \frac{P(q_n = k|x_n)P(l_n = r|x_n)}{P(q_n = k)P(l_n = r)} \\ &\approx p_{sl}(x_n|q_n{=}k)p_{sl}(x_n|l_n{=}r) (10)\end{aligned}$$

This would mean training two separate systems based upon (1) and (2), estimating the scaled-likelihood as in (10) and performing decoding according to (5).

3. EXPERIMENTAL SETUP AND STUDIES

The system proposed in Section 2 is applicable to any two kinds of subword units, e.g., phonemes and graphemes or phonemes and automatically derived subword units. 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 particularly interested in investigating the use of the orthographic information for automatic speech recognition.

We use PhoneBook database for task-independent speaker-independent isolated word recognition [8]. The training set consists of 5 hrs of isolated words spoken by different speakers. The test set contains 8 different sets of 75 word vocabulary. The words and speakers present in the training set, do not appear in either validation set or test set [9].

The acoustic vector x_n is the MFCCs extracted from the speech signal using a window of 25 ms with a shift of 8.3

[2] for all paths Q, if path unknown

ms. Cepstral mean subtraction and energy normalization are performed. At each time frame, 10 Mel frequency cepstral coefficients (MFCCs) $c_1 \cdots c_{10}$, the first-order derivatives (delta) of $c_0 \cdots c_{10}$ (c_0 is the energy coefficient) are extracted, resulting in a 21 dimensional acoustic vector. All the MLPs trained in our studies have the same 189 dimension (4 frames of left and right context, each) input layer.

There are 42 context-independent phonemes including silence associated with $\mathcal{Q}$, each modelled by a single emitting state. We trained a phoneme baseline system 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 28 context-independent grapheme subword units associated with $\mathcal{L}$ representing the 26 characters in English, silence and $+$ symbol present in the orthographic transcription of certain words in the lexicon. Similar to phonemes each of the grapheme units are modelled by a single emitting state. We trained a grapheme baseline system 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.

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

Subword Unit	# of output units	WER
Phoneme	42	4.7%
Grapheme	28	43.0%

It could be observed from the results that the grapheme-based system performs significantly poorer as compared to the phoneme-based system. In [6], similar trend was observed for the context-independent case of monophone and monograph. In [6], they generated phonetic questions (both manually and automatically) for each grapheme and modelled it through decision tree, which resulted in improvement. In our case, instead of generating such questions, we could model the relation between the phoneme and grapheme automatically from the data by training a single MLP with $42 \times 28 = 1176$ output units. However, training such a large network is a difficult task (still training). Hence, we take an alternate approach where we reduce the phoneme set to broad-phonetic-class representation. By broad-phonetic-class, we refer to the phonetic features, such as manner, place, height. According to linguistic theory, each phoneme can be decomposed into some independent and distinctive features; the combination of these features serves to uniquely identify each phoneme [10]. In our studies, we use the phonetic feature values similar to the one used in [10, Chapter 7]. Table 2 presents the different broad-phonetic-classes that we have used and their corresponding values. It could be seen from the table that the number of values for manner, place and height broad-phonetic-classes are 10, 12, and 7, respectively. So, by collapsing the phonemes into a broad-phonetic-class (many-to-one mapping) we could train a grapheme-broad-phonetic-class system which models the relation between the grapheme and the values of the broad-phonetic-class. 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. Different broad-phonetic-classes and their values.

Broad-phonetic-class	Values
Manner	vowel, approximant, nasal, stop, voiced stop, fricative, voiced fricative, closure, silence
Place	front, mid, back, retroflex, lateral, labial, dental, alveolar, dorsal, closure, unknown, silence
Height	maximum, very low height, low height, high height, very high height, closure, silence

We studied three different grapheme-broad-phonetic-class systems corresponding to the different broad-phonetic-classes, 1. manner (System 1), 2. place (System 2) and 3. height (System 3). We train acoustic models for both grapheme units and values of the broad-phonetic-class by training a single MLP via embedded Viterbi training. During training, at each iteration, we marginalize out the broad-phonetic-class as per (9) and perform Viterbi decoding according to (2) to get the segmentation in-terms of graphemes.

We performed recognition studies just using graphemes as the subword units i.e. orthographic transcription of the words like the grapheme baseline system. In order to do so, we marginalize out the broad-phonetic-class as per (9) to estimate the scaled-likelihoods of the grapheme units (i.e. the broad-phonetic-class acts like an auxiliary information which is used during the training; but hidden during recognition.) and then perform decoding like any standard ASR. Table 3 presents the experimental results of this study.

The experimental results show that performance of the grapheme-based system which uses just the orthographic transcription of the word can be significantly improved by

Table 3. Performance of grapheme-based ASR system using broad-phonetic-class as auxiliary information. The performance is expressed in terms of Word Error Rate (WER).

System	Broad-phonetic-class	# of o/p units	WER
Baseline	-	28	43.0%
System 1	Manner	280	29.2%
System 2	Place	336	27.2%
System 3	Height	196	27.9%

modelling the phonetic related information and the grapheme information together.

Next, with the improved grapheme-based system we study whether the grapheme information could help us to improve the performance of ASR if used as an auxiliary information. We investigate this in the lines of (10), where we assume independence between the phoneme units and grapheme units. We model them by separate MLPs, and, during recognition multiply the scaled-likelihood estimates obtained from the two systems in order to estimate $p(x_n|q_n, l_n)$. We conducted recognition experiments by combining the scaled-likelihood estimates of the phoneme units and the scaled-likelihood estimates of the grapheme units estimated from different MLPs, corresponding to the grapheme baseline system and the different grapheme-broad-phonetic-class systems. This yielded results slightly poorer compared to the phoneme baseline system.

It could be observed from (10) that the scaled-likelihood estimates of phoneme units and grapheme units are two different kinds of probability streams that are combined with equal weights. Hence, we performed experimental studies by weighting the $\log$ probability streams differently. The weights could be estimated automatically during recognition or could be a fixed weight [11, 12].

In order to see how crucial the weights are in determining the performance of the system, we conducted an experiment where we fixed the weights and performed recognition experiments on the test set. We then varied the weights in steps of 0.05 and performed recognition experiments at each step. The result of this study is shown in Figure 1. The best performance obtained was 4.1% for the case where the grapheme probabilities were estimated from the grapheme-broad-phonetic-class system using the place broad-phonetic-class as auxiliary information. The resulting model is significantly[3] better than the baseline system with 95% confidence. It could be seen from the figure that the operating points of the different systems are different. It is also closely related to how the grapheme-based systems perform individually.

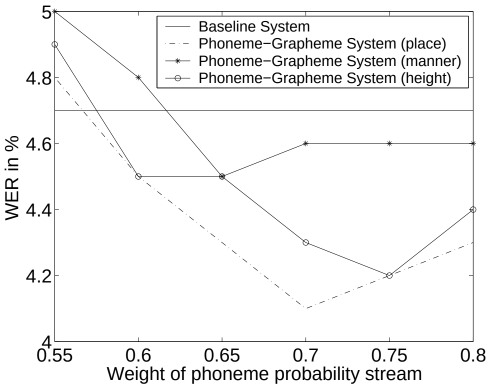

Fig. 1. Plot illustrating the relationship between the weight and the word error rate of the phoneme-grapheme system.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented an approach to model an auxiliary information which could be hidden during training as well as recognition similar to the states of HMM. In this framework, we studied the application of graphemes as subword units in standard ASR.

An ASR system was trained using graphemes as the subword units. This system yielded poor results. However, this system performs above the chance level suggesting that it might be still useful if modelled well. So, we trained a grapheme-broad-phonetic-class system in the proposed framework, where the broad-phonetic-class acts as an auxiliary information. Recognition experiments were conducted just using the grapheme subword units (orthographic transcription) by marginalizing out the broad-phonetic-class. We obtained a significant improvement in the performance of grapheme-based ASR but still is not comparable to the phoneme-based system. This suggests that it is possible to obtain a grapheme-based recognizer with considerable performance, if we could train a system with phonemes as auxiliary information.

Finally, we investigated a phoneme-grapheme system assuming independence between the two subword units. This system yielded significant improvement over the phoneme-baseline system for speaker-independent task-independent isolated word recognition task in English language. Our studies suggest that the graphemes do contain useful information for speech recognition application which, if properly modelled and utilized instead of ignoring it, could improve the performance of the ASR.

[3] The significant tests are done with standard proportion test, assuming a binomial distribution for the targets, and using a normal approximation.

In future, we would like to investigate other techniques to dynamically estimate the weights for each probability stream. We would also like to study a phoneme-grapheme system where we could train models without making the independence assumption. One such direction would be to investigate the possibility of a system where we could model the phonemes and graphemes through a single MLP. Furthermore, it would be interesting to extend the phoneme-grapheme system for a short vocabulary connected word recognition task such as OGI Numbers.

5. ACKNOWLEDGEMENT

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. We would also like to thank Prof. Hynek Hermansky for his valuable comments and suggestions.

6. REFERENCES

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

[2] Mathew Magimai.-Doss, Todd A. Stephenson, and Hervé Bourlard, "Using pitch frequency information in speech recognition," in *Eurospeech*, Geneva, September 2003, pp. 2525–2528.

[3] Todd A. Stephenson, Mathew Magimai.-Doss, and Hervé Bourlard, "Speech recognition with auxiliary information," *To appear in IEEE Trans. Speech and Audio Processing*, 2003.

[4] Zoubin Ghahramani and Michael I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[5] Beth Logan and Pedro J. Moreno, "Factorial hidden Markov models for speech recognition: Preliminary experiments," Technical Report Series CRL 97/7, Cambridge Research Laboratory, Massachusetts, USA, September 1997.

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

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

[8] J. F. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. Leung, "PhoneBook: A phonetically-rich isolated-word telephone-speech database," in *ICASSP*, 1995, pp. 1767–1770.

[9] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM/ANN systems for training independent tasks: Experiments on 'PhoneBook' and related improvements," in *ICASSP*, 1767-1770, 1997, pp. 524–528.

[10] John-Paul Hosom, *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*, PhD dissertation, CSLU, Oregon Graduate Institute of Science and Technology (OGI), USA, 2000.

[11] Astrid Hagen, *Robust speech recognition based on multi-stream processing*, PhD dissertation, EPFL, Lausanne, Switzerland, December 2001.

[12] Hemant Misra, Hervé Bourlard, and Vivek Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *ICASSP*, HongKong, April 2003, pp. II–741–II–744.