A COMPARISON OF PHONE AND GRAPHEME-BASED SPOKEN TERM DETECTION

Dong Wang[1], *Joe Frankel*[1], *Javier Tejedor*[1,2] *and Simon King*[1]

1. The Centre for Speech Technology Research,
University of Edinburgh
2. Human Computer Technology Laboratory,
Escuela Politecnica Superior UAM

ABSTRACT

We propose grapheme-based sub-word units for spoken term detection (STD). Compared to phones, graphemes have a number of potential advantages. For out-of-vocabulary search terms, phone-based approaches must generate a pronunciation using letter-to-sound rules. Using graphemes obviates this potentially error-prone hard decision, shifting pronunciation modelling into the statistical models describing the observation space. In addition, long-span grapheme language models can be trained directly from large text corpora.

We present experiments on Spanish and English data, comparing phone and grapheme-based STD. For Spanish, where phone and grapheme-based systems give similar transcription word error rates (WERs), grapheme-based STD significantly outperforms a phone-based approach. The converse is found for English, where the phone-based system outperforms a grapheme approach. However, we present additional analysis which suggests that phone-based STD performance levels may be achieved by a grapheme-based approach despite lower transcription accuracy, and that the two approaches may usefully be combined. We propose a number of directions for future development of these ideas, and suggest that if grapheme-based STD can match phone-based performance, the inherent flexibility in dealing with out-of-vocabulary terms makes this a desirable approach.

Index Terms— Spoken term detection, graphemes

1. INTRODUCTION

Information retrieval from spoken audio has attracted the attention of a number of research groups, in part driven by the recent NIST Spoken Term Detection (STD) evaluation. A standard approach is to split the task into two stages. In the first, a large vocabulary continuous speech recognition (LVCSR) system is used to generate a word or phone lattice corresponding to the audio, and in the second, lattice search is used to determine likely occurrences of the search terms.

Searching a word-based lattice works well for terms which occur in the LVCSR system's vocabulary. However, a relatively high proportion of search terms will be out-of-vocabulary (OOV) in many applications, for example names, places, acronyms, some neologisms etc. A standard method for dealing with OOV terms is to generate a phone sequence corresponding to the terms, which may then be searched for in a phone lattice.

In this work, we propose using context-dependent graphemes (CDGs) as sub-word units for STD, in particular for out-of-vocabulary search terms. In essence, this approach moves pronunciation modelling away from the letter-to-sound rules which are used to generate phone strings, and into the Gaussian mixture models which describe the observation space. Generating a pronunciation is usually a hard decision (i.e., not probabilistic), and errors introduced at this step are hard to recover from. In addition, words which have multiple pronunciations have a single grapheme representation, which simplifies the subsequent search. Large text corpora can be used to train long-span grapheme-based language models for use in lattice generation. These language models have words implicit within them, though given suitable smoothing should have the capacity to support previously unseen words. Finally, without the necessity of lexicon construction, a grapheme system can be built quickly, and can be applied to minority languages in which linguistic resources are limited.

We present experimental results on Spanish and English corpora. For Spanish, the correspondence between grapheme and phoneme is very regular, and grapheme-based LVCSR systems can achieve a similar Word Error Rate (WER) to phoneme-based systems [1]. By comparison, speech sounds in English are hard to predict accurately from the graphemes, so grapheme-based units typically perform worse than phoneme-based units for acoustic modelling [1].

We propose that given the very different cost functions of ASR, in which accurate inference of all words is considered uniformly important, and spoken term detection, in which only reliable recovery of specific words is considered, grapheme-based units cannot be dismissed on the grounds of higher WER.

Prior to search, terms must first be converted into a sequence of sub-word units. A phone-based approach requires a letter-to-sound module, for example a classification and regression tree (CART) as commonly used in text-to-speech synthesis. The performance of this module degrades for out-of-vocabulary words, giving a system built on grapheme units, in which the letter-to-sound conversion is trivial, a natural advantage.

We present an investigation into grapheme- and phone-based spoken term detection. We do not suggest that grapheme-based approaches will lead to reduced word error rates on standard transcription tasks. However, with the development of techniques capable of modelling the irregular letter-to-sound relationships that exist in languages such as English, grapheme-based methods have considerable benefits to offer open-vocabulary applications such as STD.

2. SPOKEN TERM DETECTION

We follow a standard two stage approach to spoken term detection in which prior knowledge of the search terms is not required. In the first step, the audio is indexed by decoding and producing a lattice. The second step is then to search through the lattice for the terms of interest.

It is common in spoken term detection to index the audio with both word and sub-word. In-vocabulary terms can then be searched for in a word lattice, and for out-of-vocabulary terms, the system 'backs off' to generating a sequence of sub-word units corresponding to the term, and searching for these in the sub-word lattice. It is

the sub-word scenario which we focus on in this work.

2.1. HMM-based acoustic modelling

The Spanish and English systems differ with regard to the phone and grapheme sets, though the same core architecture is used. In both cases, 12 Mel-frequency cepstral coefficients (MFCCs) are derived at 10ms intervals within 25ms windows on the acoustic signal. Energy plus first and second order derivatives are then appended, giving a 39-dimensional feature vector. Experiments using both context independent (CI) and cross-word context-dependent (CD) hidden Markov models (HMMs) are reported.

For the phone systems, all allophone and silence models have a conventional 3-state, left-to-right topology; a short pause model which has a single emitting state and a skip transition is also included in the inventory. Decision trees with phonetically motivated questions are used to cluster the triphone states.

The grapheme systems were built in an identical fashion to the phoneme-based systems, the only difference being the inventories of sub-word units and the questions used for state clustering.

Building a set of state-tied context-dependent grapheme models (trigraphemes) requires a set of questions with which to construct the decision tree. In previous work by Schultz and colleagues [1], it has been reported that singleton questions give the best performance, and these were used for state tying in our experiments. HTK [2] was used for feature extraction, acoustic modelling and lattice generation.

2.2. The lattice decoder

The HTK tool HVite is run in N-best mode to produce lattices. A depth of $N = 5$ was found to be suitable in preliminary experiments.

Once generated, the Viterbi algorithm is used to find all path fragments in the lattice that exactly match the sub-word string representing the keyword. We use an implementation of this in a tool provided by collaborators at the Brno University of Technology (BUT) [3]. In previous work, Szoke et al found that for dense lattices, exact matches are sufficient for term detection [3]. Since our main purpose is to compare phone and grapheme-based systems, we only consider exact matches in this work.

The confidence score C_K for each potential keyword K in the lattice is calculated as:

$$C_K = L_a(K) + L_b(K) + L(K) - L_{best} \quad (1)$$

where $L(K)$ is the log likelihood of the keyword K, and $L_a(K)$ and $L_b(K)$ are the log likelihoods of the best paths from the beginning of the lattice to the first node of K, and from the last node of K to the end of the lattice respectively. L_{best} is the log likelihood of the 1-best path in the whole lattice, and provides normalization. A threshold on the confidence score is estimated on the validation data in order to balance the miss and false alarm rates.

2.3. Evaluation metrics

We present results according to a number of metrics. The figure of merit (FOM) was originally defined for the task of keyword spotting [4], and gives the average detection rate over the range $[1, 10]$ false alarms per hour per keyword. The NIST STD 2006 evaluation plan [5] defines the metrics *occurrence-weighted value* (OCC) and *actual term-weighted value* (ATWV), both of which are specifically tailored to the task of spoken term detection.

Results are presented for the three metrics FOM, OCC and ATWV, and in each case the systems are tuned on validation data to the metric with which they are evaluated.

3. SPANISH EXPERIMENTS - ALBAYZIN DATABASE

The Spanish experiments use the geographical-domain ALBAYZIN corpus [6] which consists of utterances incorporating the names of mountains, rivers, cities, etc. We chose 80 in-vocabulary keywords based on their high frequency of occurrence and suitability as search terms for information retrieval in this domain.

ALBAYZIN contains two separate sub-corpora. The first has orthographic and phonetic labels, consists of 4800 phonetically-balanced sentences from 164 speakers, and is used for model training. The second of these is the Geographic corpus, which is only labelled at the orthographic level. From this, 4400 sentences from 88 speakers are used as a validation set for parameter tuning, and the remaining 2400 sentences from 48 speakers provide test material. The train, validation and test sets are disjoint with no overlap of speakers.

3.1. Phone and grapheme inventories for Spanish

The phone models were based on an inventory of 47 allophones of Spanish [7], along with beginning and end of utterance silence models. This set was selected as it achieved higher phone accuracy than a 26-phone inventory in preliminary experiments.

We use the term "grapheme" to refer to a unit consisting of a sequence of one or more letters, to be used for acoustic modelling. This may not be precisely the same as the alphabet used for writing because we can expect better performance if we account for a small number of language-specific special cases. We use a total of 27 grapheme units, (a, b, c, ch, d, e, f, g, i, j, k, l, ll, m, n , ñ, and o – z), as described in [8].

3.2. Spoken term detection results

Table 1 presents STD results for Spanish. Two trends are apparent, and consistent across metrics: firstly, CD models give better performance than CI models, and secondly, the grapheme-based systems outperform their phone-based counterparts. In fact this difference is so pronounced that the context-independent grapheme-based system outperforms the context-dependent phone-based system. Paired t-tests showed that these increases for each metric were significant with $p < 0.01$.

	phone		grapheme	
	CI	CD	CI	CD
FOM	44.0	47.1	58.1	64.0
OCC	0.40	0.42	0.53	0.61
ATWV	0.15	0.19	0.28	0.31

Table 1. Spanish spoken term detection results for CI and CD phone- and grapheme-based systems.

The detection error trade-off (DET) curve in Figure 1 shows miss against false alarm probability. This gives an indication of the system performance at a number of operating points. It is clear that both grapheme-based systems outperform the phone-based systems regardless of the desired balance of precision and recall.

4. ENGLISH EXPERIMENTS - MEETINGS DOMAIN

Experiments on English use data from the meetings domain. The training data is publicly available and was recorded in instrumented meeting rooms at multiple locations. These include the International Computer Science Institute (ICSI), National Institute for Standards and Technology (NIST), Carnegie Mellon University Interactive Systems Laboratory (ISL), plus partners of the Augmented Multiparty Interaction (AMI) project. In total, the training data amounted

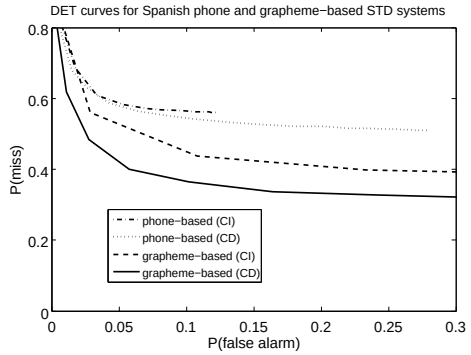

Fig. 1. DET curves for Spanish STD with context dependent and independent phone- and grapheme-based systems.

to a little over 100 hours. For both training and testing we use data from the independent headset microphone (IHM) condition which offers much higher signal-to-noise ratios than the distant microphone conditions.

Results are presented on test data from the NIST Spring 2004 Rich Transcription (RT04s) evaluation. The development set from the same year was used for intermediate parameter tuning. There were 90 words, including most frequently used words, named entities and compound words, selected as search terms from the reference transcription.

Whilst all search terms in the Spanish experiments were in-vocabulary (INV), for English the terms were divided equally between INV and out-of-vocabulary (OOV). Pronunciation generation is handled by Festival's letter-to-sound module [9], which is based on the Carnegie Mellon University (CMU) dictionary. Pronunciations occurring in the CMU dictionary are INV and are used directly. OOV words are generated using a CART module which is considered to be state-of-the-art, yet still has an error rate in the region of 30%.

4.1. Phone and grapheme inventories for English

To ensure consistency between HMM states and phone pronunciations, the set of 42 English phones from the CMU dictionary were used for the phone-based system. The grapheme set simply uses the 26 letters found in the English alphabet, with the addition of short pause and silence models.

4.2. Automatic speech recognition results

It is interesting to compare phone and grapheme-based performance on an ASR task. Table 2 summarizes these results for triphone and trigrapheme systems on the RT04s IHM test data.

	phone (CD)	grapheme (CD)
WER	44.5%	54.5%
PER / GER	48.2%	46.3%

Table 2. WER on English meetings data fro triphone- and trigrapheme-based ASR. Phone and grapheme error rates (PER and GER respectively) are also given.

As expected, we find that the WER for the phone-based system is much lower, around 10% absolute, than that of the grapheme-based system. The phone error rate (PER) at 48.2% is slightly higher than the grapheme error rate (GER) of 46.3%, though these results are not directly comparable as there are fewer graphemes than phones.

4.3. Spoken term detection results

As with the Spanish language experiments, simple bigram language models were used during phone and grapheme decoding. In order to ensure consistency between the phone language model and pronunciations generated to accompany search terms, the phone bigram training material was generated using the letter-to-sound module by converting the transcript of the acoustic model training data.

	phone (CD)			grapheme (CD)		
	ALL	INV	OOV	ALL	INV	OOV
FOM	20.5	19.0	31.4	18.0	17.1	24.3
OCC	0.44	0.47	0.22	0.34	0.38	0.06
ATWV	0.25	0.28	0.22	0.16	0.23	0.09

Table 3. English STD performance for CI and CD phone- and grapheme-based systems. Results are given for the complete set of evaluation terms (ALL), as well as divided into in-vocabulary (INV) and out-of-vocabulary (OOV) terms.

Results for phone and grapheme-based systems are given in Table 3 for each of the 3 evaluation metrics. The phone-based system outperforms the grapheme-based approach, a result which is further illustrated in the DET curve of Figure 2. However, we observed substantial variation in the detection rates across the set of terms, and paired t-tests show the difference between phone and grapheme-based performance is statistically significant ($p < 0.01$) based on evaluation with the occurrence-weighted measures FOM and OCC, though not with term-weighted ATWV.

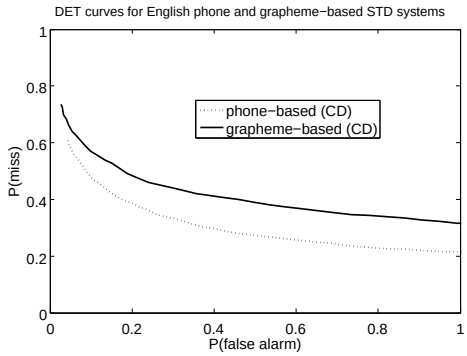

Fig. 2. DET curves for English STD with context dependent phone- and grapheme-based systems.

In addition to results on the complete set of evaluation terms (ALL), Table 3 presents the results divided into in-vocabulary (INV) and OOV search terms. Again, the conclusions which can be drawn from the data differ according to the evaluation metric. Based on FOM, we find a significant difference between phone- and grapheme-based performance for INV terms, though no significant difference in the performance on OOV terms. Conversely, for OCC and ATWV, our tests show systematic differences in the STD performance on OOV terms, though not for INV terms. The high STD performance variation across terms coupled with the low significance of many of the paired tests suggests that there are different error patterns for phone and grapheme-based approaches, and hence complementary information which may be exploited.

This is confirmed by the preliminary system combination results presented in Table 4. A simple voting scheme was used in which

the terms hypothesised by both systems were considered as potential matches, and rescored using a combination of the phone and grapheme-based scores. We observe increases in the FOM value across all sets of the data, which for the ALL and INV sets are statistically significant with $p < 0.025$.

	ALL	INV	OOV
phone (CD) FOM	20.5	18.9	31.4
phone (CD) + grapheme (CD) FOM	23.7	22.2	34.3

Table 4. English STD performance as FOM for CD phone and system combination of CD phone and CD grapheme.

An STD performance reduction using graphemes rather than phones is expected given the significant increase in transcription WER. In order to gain insight into how ASR and STD performance relate, we built a series of phone and grapheme-based HMMs with varying numbers of Gaussian mixture components associated with each state. These were intended to simulate various levels of ASR performance, and were evaluated in terms of WER and STD FOM. The results are plotted in Figure 3, and show that for equal word error rate, grapheme-based STD gives higher FOM. Extrapolation of this graph is clearly risky, though it supports that suggestion that grapheme-based STD may achieve comparable performance to phone-based despite higher WER on transcription tasks.

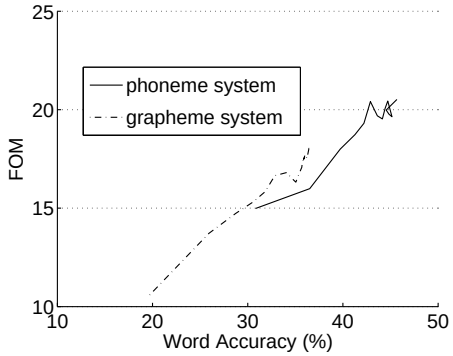

Fig. 3. Relationship between word accuracy and FOM for the phone-based system and grapheme-based system.

5. CONCLUSIONS AND FUTURE WORK

In this paper we proposed grapheme-based units as a basis for spoken term detection, and evaluated this idea on both Spanish and English tasks. Our finding was that for Spanish, in which ASR performance is similar using phones and grapheme units, the grapheme-based system significantly outperformed the phone-based system for STD.

For English, we find that the grapheme-based system is outperformed by the phone-based approach. However, inspection of the relationship between the ASR accuracy and STD performance suggests that the grapheme-based approach may be able to match phone-based performance despite higher WER. If so, the inherent flexibility in dealing with out of vocabulary terms makes graphemes a desirable approach. We also find that complementary information exists between phone and grapheme units which may usefully be exploited.

In order to provide phone/grapheme experiments which were directly comparable, simple bigram language models (LMs) were used. However, preliminary experiments show that the longer span N-gram LMs (e.g. $N = 8$) contribute more to grapheme- than phone-based STD. Future work will include investigating letter LMs, in particular the smoothing which is required in order to ensure that the LMs generalize to OOV words.

Another research focus will be to increase the core ASR performance of the grapheme-based system. Possibilities include long-span letter-based LMs as discussed above, wider context for tri-grapheme state clustering, multi-letter models, and tandem features. In addition, we plan to develop methods which replace the decision trees that map from context-independent to context-dependent grapheme units with probabilistic mappings. This will have the effect of removing a further hard decision from the system, and improve the system's ability to model unusual pronunciations.

6. ACKNOWLEDGEMENTS

Some of the Spanish material has been submitted to the Speech Communication special issue on Iberian languages [8]. DW is a Fellow on the EdSST interdisciplinary Marie Curie training programme. JF is funded by the Edinburgh Stanford Link. JT is a visiting researcher at the CSTR, University of Edinburgh. SK is an EPSRC Advanced Research Fellow. Many thanks to Igor Szoke and colleagues in the Speech Processing Group of FIT, Brno University of Technology for providing the lattice search tools. Part of this work is also funded by the Spanish Ministry of Science and Education (TIN 2005-06885)

7. REFERENCES

[1] M. Killer, S. Stker, and T. Schultz, "Grapheme based speech recognition," in *Proc. Eurospeech*, December 2003.

[2] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. amnd Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Microsoft Corp. and Cambridge University Engineering Department, 2006.

[3] Igor Szoke, PetrSchwarz, Pavel Matejka, Lukas Burget, Martin Karafiat, Michal Fapso, and Jan Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Proc. Interspeech*, Portugal, September 2005, pp. 633–636.

[4] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting.," in *Proc. ICASSP*, UK, 1989, vol. 1, pp. 627–630.

[5] NIST, *The spoken term detection (STD) 2006 evaluation plan*, National Institute of Standards and Technology, Gaithersburg, MD, USA, v10 edition, September 2006.

[6] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mario, and C. Nadeu, "Albayzin speech database: Design of the phonetic corpus," in *Proc. Eurospeech*, September 1993, vol. 1, pp. 653–656.

[7] A. Quilis, *El comentario fonologico y fonetico de textos*, ARCO/LIBROS, S.A., 1998.

[8] J. Tejedor, D. Wang, J. Frankel, S. King, and J. Colas, "A comparison of grapheme and phoneme-based units for Spanish spoken term detection," *Speech Communication*, 2007 (Submitted).

[9] R. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, April 2007.