A Posterior Probability-Based System Hybridisation and Combination for Spoken Term Detection

Javier Tejedor[1], Dong Wang[2], Simon King[2], Joe Frankel[2], José Colás[1]

[1]Human Computer Technology Laboratory,
Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain
[2]The Centre for Speech Technology Research,
University of Edinburgh, UK

`javier.tejedor@uam.es, dwang2@inf.ed.ac.uk`

Abstract

Spoken term detection (STD) is a fundamental task for multi-media information retrieval. To improve the detection performance, we have presented a direct posterior-based confidence measure generated from a neural network. In this paper, we propose a detection-independent confidence estimation based on the direct posterior confidence measure, in which the decision making is totally separated from the term detection. Based on this idea, we first present a hybrid system which conducts the term detection and confidence estimation based on different sub-word units and then propose a combination method which merges detections from heterogeneous term detectors based on the direct posterior-based confidence. Experimental results demonstrated that the proposed methods improved system performance considerably for both English and Spanish.

Index Terms: speech recognition, spoken term detection, confidence estimation, grapheme

1. Introduction

Information retrieval from speech, in the way of spoken term detection (STD), has been receiving much interest of late, in part due to the evaluation organised by NIST [1]. The standard architecture of a STD system consists of a speech recogniser to transcribe the input speech to a word or sub-word lattice (offline indexing) and a term detector to search putative occurrences of the enquiry terms from the lattice (online indexing).

An essential component of a STD system is the *decision maker*, which examines each putative detection and determines if it is a reliable hit or a false alarm (FA). This hit/FA decision is based on certain confidence measures that commonly derive from a posterior probability of the event that a search term K appears in a specific segment of the input speech O, formally written as

$$c(d) = p(K_{t_1}^{t_2}|O) \qquad (1)$$

where $c(d)$ is the confidence of a detection d of term K, and $K_{t_1}^{t_2}$ denotes the event that K appears in the speech segment from t_1 to t_2.

A commonly used posterior probability-derived confidence measure is computed from the lattice [2, 3], which we call *lattice-based confidence*. A drawback of this confidence measure is that the detection and the confidence estimation are 'glued' together, which makes integrating multiple confidence difficult.

In previous work [4] we presented a new direct posterior-based confidence that derives the posterior probability from a multi-layer perceptron (MLP). The new confidence has exhibited better performance than the conventional lattice-based confidence, partly because of the discrimination power provided by the MLP. Another distinct advantage accompanying the direct posterior confidence estimation is that the confidence estimation can be separated from the term detection. With this separation, decision making can be addressed with more reliable confidence and multiple confidence measures can be integrated to improve the decision quality.

This paper presents our work on the detection-independent confidence estimation approach. We first propose a hybrid approach which uses different sub-word units for term detection and confidence estimation, and then we propose a combined method which merges detections from heterogeneous systems based on different sub-word units. We chose phonemes and graphemes as the two sub-word units to test the hybridisation and combination: phonemes are the most widely used units in sub-word-based STD systems and graphemes have been demonstrated working well and complementary to phonemes by our previous work [5].

In the following section, we first review the direct posterior-based confidence and then describe the direct posterior-based system hybridisation and combination in Section 3 and Section 4 respectively. Section 5 presents our experiments on English and Spanish and Section 6 concludes the whole paper.

2. Direct posterior-based confidence

In previous work [4], we presented a direct posterior-based confidence measure. In this approach, an MLP is used to estimate the posterior probability that phone Q is spoken at time t given the input speech O, denoted as $p(Q_t|O)$. Then the posterior probability of a search term K appears in the speech segment from t_1 to t_2 is given by Equation 2-3 under some assumptions [4].

$$p(K_{t_1}^{t_2}|O) \approx p(K_{t_1}^{t_2}|O)\frac{p(K^l|C'_{K^l})}{p(K)} \qquad (2)$$

$$= \prod_{t=t_1}^{t_2} p(Q_t|O)\,\frac{p(K^l|C'_{K^l})}{p(K)} \qquad (3)$$

where K^l denotes the spelling form of K, C'_{K^l} denotes the best context of K^l and Q_t is the phone at time t in the phone path of K.

Notice that the information the direct posterior confidence requires from the term detector is only the term identity K and
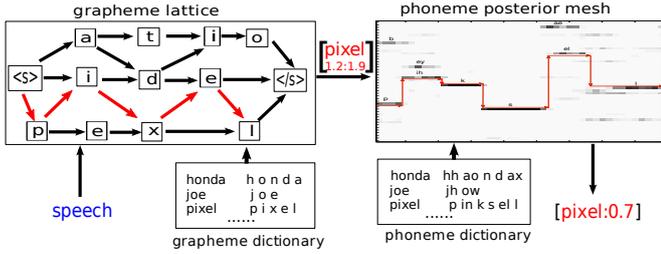
Figure 1: The hybrid system with a grapheme-based term detector and a phoneme-based confidence estimation. The red path in the lattice represents a detection of the term "pixel". The posterior probabilities generated by the MLP construct a posterior mesh which has a posterior vector at each frame. The red path in the posterior mesh represents a possible path corresponding to the detection [pixel 1.2:1.9], whose confidence is computed according to Equation 2-3, resulting in the confidence-bearing detection [pixel:0.7].

the starting and ending time t_1 and t_2. With these three quantities provided, the confidence estimation is totally independent of the term detection.

3. System hybridisation with posterior-based confidence

Since the confidence estimation can be separated from the term detection with the posterior-based confidence, we can employ different sub-word units to perform the detection and estimate the confidence, which is the idea of system hybridisation. An example of the hybrid system is shown in Figure 1, where we trained a grapheme-based ASR system to decode the speech, used a grapheme-based dictionary to detect search terms and trained a phoneme-based MLP to estimate the confidence of the putative detections.

A problem inherent to the hybrid approach is that the time alignment of the sub-word path of a detection is unavailable when the confidence estimation is based on sub-word units different from those used by the term detection. We solve this problem by a Baum-Welch approach which accumulates the confidence of all possible alignments, as expressed by Equation 4-5.

$$p(K_{t_1}^{t_2}|O) \quad \approx \quad p(K_{t_1}^{t_2}|O)\frac{p(K^l|C'_{K^l})}{p(K)} \qquad (4)$$

$$= \quad \sum_{\xi} \prod_{Q_t \in \xi, t=t_1}^{t_2} p(Q_t|O) \; \frac{p(K^l|C'_{K^l})}{p(K)} \quad (5)$$

where ξ is a possible phone path of K in the speech segment from t_1 to t_2 and Q_t is the phone at time t in the path ξ. In fact, Equation 5 was used in general in our study, even for the system which uses the same sub-word units for detection and confidence.

4. System combination with posterior-based confidence

The second approach we employ the term-independent confidence estimation is a heterogeneous system combination. In this approach, we either accumulate multiple confidence for detections from a single detector, or merge detections from different detectors based on the same confidence measure, or even combine different systems in case all of them use the direct posterior-based confidence. We hypothesize that all these combinations could utilise information coming from multiple resources, resulting in a better system performance.

The combination, regardless of each type, can be implemented as a simple fusion in the spirit of ROVER [6]. Figure 2 illustrates the combination of a phoneme-based system and a grapheme-based system. Detections from each system are aligned and then each hypothesised term detection is examined. If a hypothesis does not overlap with a hypothesis from the other system, it is simply copied to the final result along with its confidence score. If the same term is hypothesised by the two systems, an output hypothesis is generated which has the earliest and latest hypothesised start and end times and the confidence of the first system *plus* the confidence of the second system weighted by a fusion factor α, tuned on the development set.

Remind that with the detection-independent confidence estimation approach, detections from any heterogeneous systems can be combined. For example, we can combine detections from word-based systems, phoneme-based systems with different order of language models (LMs) and even keyword spotting systems.

5. Experiments

We tested the proposed hybrid and combined approaches on both English and Spanish. A phoneme-based system and a grapheme-based system are built for each language. The HTK toolkit from Cambridge was used to train the acoustic models and generate lattices for input speech for both languages and to train the LMs for Spanish, the SRI LM toolkit was used to train the LMs for English and QuickNet from ICSI was used to train the MLPs and predict posterior probabilities. The term detector was implemented with *Lattice2Multigram* provided by the Speech Processing Group, FIT, Brno University of Technology.

The actual term weighted value (ATWV) defined by NIST for STD, which considers both hit ratio and false alarm rate in a single metric, averaged over all search terms was used as a single point evaluation metric [1]. To examine the behaviour of a STD system at different hit/FA ratios, we present detection (DET) curves.
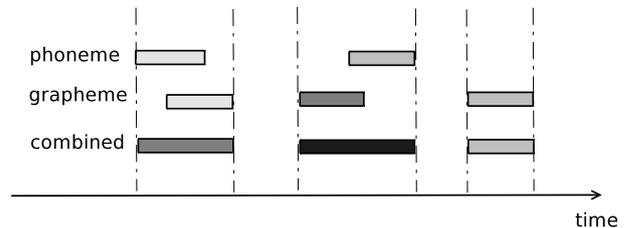

Figure 2: An example of a phoneme- and a grapheme-based system combination. Hypotheses from the two systems that overlap in time are merged as a single detection, whilst detections without overlap are duplicated in the final result directly. Confidence scores are accumulated for overlapped detections and unchanged for non-overlapped detections. Shading represents confidence, with darker being greater. Reminding that overlapped detections of each individual system have been merged to a single detection before combination, by accumulating the duration and confidence scores.

5.1. MLP training

3-layer MLPs were trained to predict the posterior-based confidence. For both English and Spanish, the input layer of the MLP consists of 9 frames, amounting 351 nodes. The output layer corresponds to the sub-word units, which are 44 phonemes and 26 graphemes plus a short and a long silence for English, and 47 phones and 28 graphemes plus a short silence and a long initial and a long final silence for Spanish. The size of the hidden layer was optimised by cross-validation.

5.2. English experiments

Conventional MFCC features, which include 12 MFCCs with the zero-order coefficient and their 1st and 2nd order derivatives, were improved using the tandem ANN/HMM approach [7]. The tandem-based acoustic models were discriminatively trained using Minimum Phone Error (MPE) [8]. Conventional 12 PLP features plus the zero-order coefficient, and their 1st and 2nd order derivatives, were used as MLP input features.

English experiments are performed in the domain of multi-party meetings. Acoustic models, based on triphone and trigrapheme HMMs, were trained on over 100 hours of speech collected in various instrumented meeting rooms using headset-mounted microphones. Results are presented on test data from the NIST Spring 2004 Rich Transcription (RT04s) evaluation; the development set of the same year is used for parameter tuning. Long-span sub-word LMs, trained from approximately 51 million words of text, including transcriptions of meetings and a large news archive, were used in sub-word decoding and lattice search. The LM order is determined through tuning on the development set: a 7-gram for the phoneme-based system and an 8-gram for the grapheme-based system. A total of 90 search terms were selected which include company and city names, some compound words, and commonly used terms.

5.2.1. System hybridisation

For English, we treat the phoneme-based system (phoneme-based detection and phoneme-based confidence) as the baseline, since it worked better than the grapheme-based system in our previous work [5]. There are various hybrid variants, as shown in the first row of the results in Table 1. The interesting observation is that the hybrid system with phoneme-based detection and grapheme-based confidence gave the best performance; a paired t-test shows that the improvement the phoneme-grapheme hybrid system achieved over the baseline system is weakly significant ($p < 0.05$). The DET curves in Figure 3 show that the hybrid system outperforms the baseline in most operating regions.

5.2.2. System combination

Table 1 shows the results of the combined systems. We find that nearly all the combined systems worked better than individual systems, and the best results came from the combination of the two hybrid systems. The paired t-test shows that the best combined system outperforms the baseline system significantly in statistics ($p < 0.01$). The DET curves are shown in Figure 3 and they indicate that the combined systems outperformed individual systems consistently.

5.3. Spanish experiments

In Spanish experiments, standard 12 MFCCs + zero-order coefficient plus their 1st and 2nd derivatives were used as acoustic features for HMMs and standard 12 PLPs + zero-order coefficient plus their 1st and 2nd derivatives were used as MLP input

	ATWV			
	ph⟨ph⟩	ph⟨gr⟩	gr⟨ph⟩	gr⟨gr⟩
Hybrid	0.379	0.393	0.301	0.266
+ph⟨ph⟩	–	0.396	0.416	0.409
+ph⟨gr⟩	–	–	0.440	0.426
+gr⟨ph⟩	–	–	–	0.305
+gr⟨gr⟩	–	–	–	–

Table 1: STD results in terms of ATWV for the system hybridisation and system combination for English. *ph* denotes phoneme and *gr* denotes grapheme. A hybrid system is represented as a *ph/gr* pair, i.e., ph⟨gr⟩ represent a hybridisation of phoneme-based detection and grapheme-based confidence. The row *Hybrid* reports performance of the hybrid systems, and the following rows present combined systems.
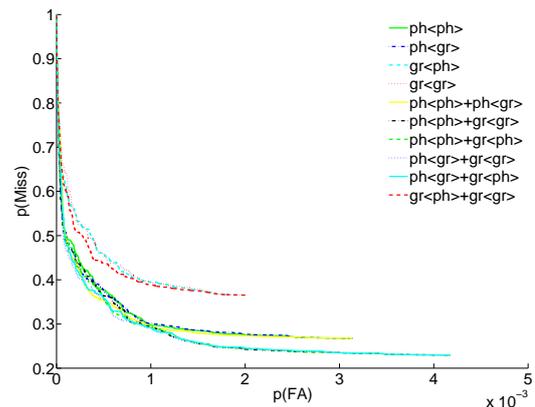

Figure 3: DET curves of the phoneme- and grapheme-based systems for English, as well as their hybridisation and combination. *ph* denotes phoneme and *gr* denotes grapheme. A hybrid system is represented as a *ph/gr* pair and a combined system is represented by an 'addition' of two individual systems. For example, ph⟨gr⟩+gr⟨ph⟩ denotes a combination of two hybrid systems, one has phoneme-based detection and grapheme-based confidence, and the other has grapheme-based detection and phoneme-based confidence.

features.

The geographical domain Albayzin read speech database [9] was used for Spanish experiments. Triphone and trigrapheme HMMs were used as acoustic models for the phoneme- and grapheme-based system respectively, and 2-gram phoneme and grapheme LMs were used for decoding. We selected 80 search terms based on their occurrences in the development and test sets, including city, river and mountain names.

5.3.1. System hybridisation

In these experiments, we select the grapheme-based system as the baseline due to its better performance than that of the phoneme-based system [10]. Results are shown in Table 2. Different from the English experiments, we did not find obvious performance improvement with the hybridisation, which might be because of the predominant priority of the grapheme-based system in Spanish. DET curves are shown in Figure 4, confirming the same observation.

5.3.2. System combination

The experimental results of the combined systems are shown in Table 2. We find that the best performance comes from the com-

	ATWV			
	ph⟨ph⟩	ph⟨gr⟩	gr⟨ph⟩	gr⟨gr⟩
Hybrid	0.203	0.189	0.250	0.252
+ph⟨ph⟩	–	0.185	0.280	0.251
+ph⟨gr⟩	–	–	0.262	0.262
+gr⟨ph⟩	–	–	–	0.246
+gr⟨gr⟩	–	–	–	–

Table 2: STD results in terms of ATWV for the system hybridisation and system combination for Spanish. The same notations are used as in the English experiments.

bined system which merges detections from the grapheme- and phoneme-based systems, and estimates their confidence based on the same phoneme posterior probabilities. These results support our hypothesis that detections from heterogeneous detection systems can be merged with the detection-independent confidence estimation. DET curves in Figure 4 demonstrate that all combinations outperform the baseline consistently in most of the operating range, although the performance improvement achieved by the best combined system over the baseline is not significant in a paired t-test.
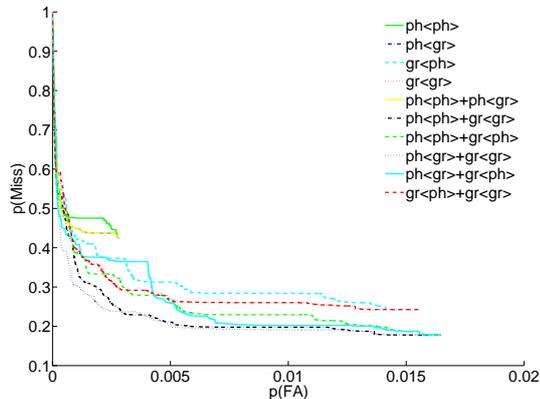

Figure 4: DET curves of the phoneme- and grapheme-based systems for Spanish, as well as their hybridisation and combination. The same notations are used as in the English experiments.

5.4. Discussion

Inspecting the results we obtained in the experiments of English and Spanish, we observe different pattern performance. The performance of the English systems were more significantly improved by the hybridisation and combination than that of the Spanish systems. This can be attributed to the different grapheme-phoneme relationships within these two languages. In English, the pronunciation rules are rather complex, leading to an irregular relationship between graphemes and phonemes, which makes the grapheme- and phoneme-based systems capture different information of the language and exhibit complementary performance on STD. This complementarity is the underlying assumption that a hybrid or combined system works. Contrary, for Spanish, the relationship between graphemes and phonemes is more regular, hence the hybridisation and combination do not provide much extra information. This is why we observed such remarkable performance improvement in the English experiments, while the improvement is relatively insignificant in the Spanish experiments.

6. Conclusions

In this paper we presented a detection-independent confidence estimation approach for spoken term detection. Based on this approach, we provided a hybridisation method to employ different sub-word units for term detection and confidence estimation. In addition, we proposed a combination method to merge detections from heterogeneous detection systems based on the direct posterior-based confidence measure. Both methods provide substantial performance improvement in experiments conducted on English and Spanish.

7. Acknowledgements

8. References

[1] NIST, *The spoken term detection (STD) 2006 evaluation plan*, v10 ed., National Institute of Standards and Technology, Gaithersburg, MD, USA, September 2006. [Online]. Available: http://www.nist.gov/speech/tests/std

[2] F. Wessel, K. Macherey, and R. Schluter, "Using word probabilities as confidence measures," in *Proc. ICASSP'98*, 1998.

[3] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiat, M. Fapso, and J. Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Proc. Interspeech'05*, Lisbon, Portugal, 2005, pp. 633–636.

[4] D. Wang, J. Tejedor, J. Frankel, S. King, and J. Colás, "Direct posterior based confidence for spoken term detection," in *To appear in Proc. of ICASSP*, Taipei, Taiwan, April 2009.

[5] D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in *Proc. of ICASSP*, Las Vegas, NV, US, March-April 2008, pp. 4969–4972.

[6] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Proc. of IEEE ASRU Workshop*, Santa Barbara, CA, US, December 1997, pp. 347–354.

[7] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Proc. ICASSP'00*, Istanbul, June 2000.

[8] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proc. ICASSP'02*, vol. 1, Orlando, FL, USA, May 2002, pp. 105–108.

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

[10] J. Tejedor, D. Wang, J. Frankel, S. King, and J. Colás, "A comparison of grapheme and phoneme-based units for spanish spoken term detection," *Speech Communication*, vol. 50, no. 11-12, pp. 980–991, November 2008.