GTTS System for the Spoken Web Search Task at MediaEval 2012

Amparo Varona
DEE, UPV/EHU, B. Sarriena
48940 Leioa, Spain
amparo.varona@ehu.es

Mikel Penagarikano
DEE, UPV/EHU, B. Sarriena
48940 Leioa, Spain
mikel.penagarikano@ehu.es

L.J. Rodriguez-Fuentes
DEE, UPV/EHU, B. Sarriena
48940 Leioa, Spain
luisjavier.rodriguez@ehu.es

German Bordel
DEE, UPV/EHU, B. Sarriena
48940 Leioa, Spain
german.bordel@ehu.es

Mireia Diez
DEE, UPV/EHU, B. Sarriena
48940 Leioa, Spain
mireia.diez@ehu.es

ABSTRACT

This paper briefly describes the system presented by the Working Group on Software Technologies (GTTS)[1] of the University of the Basque Country (UPV/EHU) to the Spoken Web Search task at MediaEval 2012. The GTTS system apply state-of-the-art phone decoders to search for approximate matchings of the N-best decodings of a spoken query in the phone lattice of the target audio document.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Spoken Term Detection, Phone Lattice, String Matching

1. INTRODUCTION

The spoken web search task included in MediaEval 2012 consists of searching for a spoken query within a set of audio documents [3]. Since both the queries and the audio documents may contain any language, a language-independent audio search system must be developed. The locations and durations of all the occurrences of spoken queries in the audio documents must be obtained [2].

2. SYSTEM ARCHITECTURE

The system developed by GTTS looks for approximate string matchings at the phone level (see Figure 1). Phone lattices are computed for both the spoken queries and the audio documents. In the case of spoken queries, phone lattices are further processed to extract N-best phone decodings. Then, approximate matchings of the N-best hypotheses corresponding to each query are searched on the phone lattices of audio documents, each match being assigned a detection score. Scores are converted to log-likelihood ratios and length-normalized. Finally, for each query, only the K detections with the highest scores are output.

[1]http://gtts.ehu.es

MediaEval 2012 Workshop, October 4-5, 2011, Pisa, Italy

2.1 Computing phone lattices

As a first step, the open software Brno University of Technology (BUT) phone decoders for Czech, Hungarian and Russian [4] are applied to decode both the spoken queries and the audio documents. Remind that BUT decoders have been trained on 8kHz SpeechDat(E) databases recorded over fixed telephone networks, containing 12, 10 and 18 hours of speech and featuring 45, 61 and 52 phonetic units for Czech, Hungarian and Russian, respectively. For each unit, a three-state model is used, so three posterior probabilities per frame are calculated.

Since exact (or almost exact) matchings are required to detect queries, the number of phonetic units may be too high for this application. Note that the same sound may be decoded in different ways, producing similar (but different) units. To compensate for this effect, the set of units is reduced by defining groups of similar (i.e. highly confusable) units, according to their characterization in the International Phonetic Alphabet (IPA). Besides, three non-phonetic units used by BUT decoders are fused into a single non-phonetic unit model. Eventually, we use 25 units for Czech, 23 for Hungarian and 21 for Russian.

Let us consider one of the BUT decoders, featuring M phone units, each of them typically represented by means of a left-right model of S states. The posterior probability of each state s ($1 \leq s \leq S$) of each phone model i ($1 \leq i \leq M$) at each frame t, $p_{i,s}(t)$, is directly provided by the phone decoder. When considering a reduced set of units, each unit j clusters a number of similar units, and its posterior probability at each state s and each frame t cand be computed by adding the posterior probabilities of all of them:

$$p_{j,s}(t) = \sum_{\forall i \in S_j} p_{i,s}(t) \quad (1)$$

with $1 \leq j \leq R$, R being the number of clusters in the reduced set and S_j the subset of phone units in cluster j. Finally, posterior probabilities are used to produce phone lattices —which encode multiple hypotheses with acoustic likelihoods—, by means of the HTK tool *HVite* [9].

2.2 Searching phone lattices

For each spoken query, the N phone decodings with the highest likelihoods are extracted from the phone lattice by

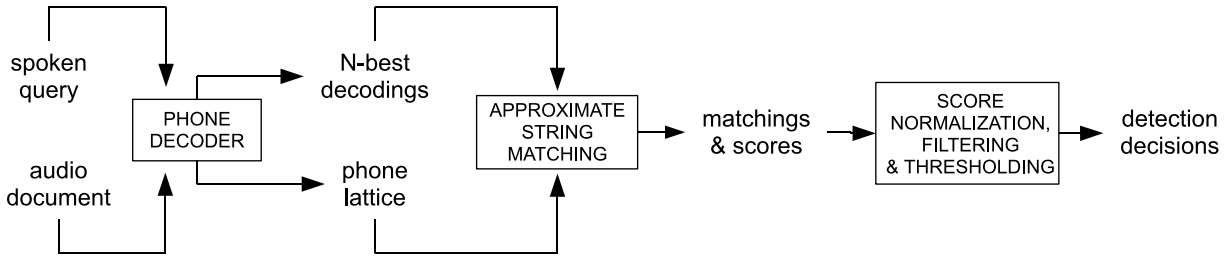

Figure 1: Processing steps of the GTTS system for the MediaEval 2012 Spoken Web Search task.

means of the *lattice-tool* of *SRILM* [5]. Then, the *Lattice2Multigram* (L2M) tool by Dong Wang [6, 8, 7][2] is applied. L2M takes two inputs: a list of phone strings (the queries) and a list of phone lattices (the documents), and outputs detections in MLF format [9]. The behaviour of L2M is controlled by several parameters, which have been tuned on the development dataset. In particular, *LogLikeliBaumWelsh* has been used as lattice score computation method and *SubstInsDel* as confusion method, being n the maximum number of edition operations allowed for approximate matching (see Table 1).

2.3 Handling the scores

Three filters are sequentially applied to MLF detection files:

- *mlf2mlf*: log-likelihood ratios are computed based on the scores of the MLF file, and normalized by the length (in frames) of each detected segment i, in the following way:

$$new\_score_i = \log \frac{e^{\frac{score_i}{length_i}}}{\sum_{\forall j \neq i} e^{\frac{score_j}{length_j}}} \quad (2)$$

  Given a set of spoken queries and a set of audio documents, three MLF files are produced, based on the BUT decoders for Czech, Hungarian and Russian, respectively. Detection files can be either mixed and processed jointly, or processed independently. In any case, for each audio document, overlapping detections are processed such that only the most likely detection is taken into account, the remaining ones being discarded.

- *mlf2std*: detection information is converted to the final STD format.

- *std2std*: for each query, only the K most likely detections are retained, scores are z-normalized and a threshold is applied.

3. PRELIMINARY EXPERIMENTS

Table 1 summarizes the results obtained in preliminary experiments on the development set, using different configurations (MLF files mixed and jointly processed, $K = 50$ and the final threshold not effectively applied): N-best from $N = 1$ to $N = 3$, and approximate matching from $n = 1$ to $n = 3$. The Actual Term Weighted Value (ATWV) is used as primary evaluation measure [1]. False alarm and miss probabilities are shown too. Best performance was attained with $N = 3$ and $n = 2$, so these were the tunings chosen for the *primary system*. For the contrastive system, we chose the second best configuration: $N = 1$ and $n = 2$.

[2] http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html

Table 1: Preliminary experiments on the development set using different configurations.

N-best	n	ATWV	P(FA)	P(Miss)
1-best	1	0.070	0.00140	0.911
	2	**0.102**	**0.00217**	**0.865**
	3	0.096	0.00276	0.863
2-best	1	0.088	0.00156	0.891
	2	0.102	0.00226	0.864
3-best	1	0.093	0.00161	0.883
	2	**0.110**	**0.00229**	**0.856**
5-best	1	0.096	0.00164	0.880
	2	0.111	0.00239	0.854

Table 2: ATWV performance for the primary (3-best, n=2) and contrastive (1-best, n=2) systems.

	devC-devQ	devC-evalQ	evalC-devQ	evalC-evalQ
Pri	0.105	0.108	0.078	0.081
Con	0.098	0.083	0.069	0.070

4. RESULTS

Table 2 presents the results attained in all the required conditions. Clearly, the primary system outperforms the contrastive system. Though results are not competitive, we expect further improvements under the proposed phone-lattice + approximate string matching approach. We are currently exploring a change in the focus, by searching for the best detection of each spoken query in each audio document.

5. REFERENCES

[1] J. Fiscus, J. Ajot, J. Garafolo, and G. Doddington. Results of the 2006 spoken term detection evaluation. In *ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, Amsterdam, 2007.

[2] F. Metze, E. Barnard, M. Davel, C. van Heerden, X. Anguera, G. Gravier, and N. Rajput. The spoken web search task. In *MediaEval 2012 Workshop*, Pisa, Italy, 2012.

[3] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. van Heerden, G. V. Mantena, A. Muscariello, K. Prahallad, I. Szoke, and J. Tejedor. The Spoken Web Search Task at MediaEval 2011. In *ICASSP*, pages 3487–3491, Kyoto, Japan, 2012.

[4] P. Schwarz. *Phoneme recognition based on long temporal context*. PhD thesis, FIT, BUT, Brno, Czech Republic, 2008.

[5] A. Stolcke. SRILM - an extensible language modeling toolkit. In *ICSLP*, pages 257–286, 2002.

[6] D. Wang, S. King, and J. Frankel. Stochastic pronunciation modelling for out-of-vocabular spoken term detection. *IEEE Trans. on Audio, Speech, and Language Processing*, 2011.

[7] D. Wang, S. King, and J. Frankel. Direct posterior confidence estimation for spoken term detection. *ACM Transactions on Information Systems*, 2012.

[8] D. Wang, J. Tejedor, S. King, and J. Frankel. Term-dpendent confidence normalization for out-of-vocabulary spoken term detection. *Journal of Computer Science and Technology*, 2012.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Lui, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (Version 3.4)*. Cambridge, 2006.