GTTS Systems for the SWS Task at MediaEval 2013

Luis J. Rodriguez-Fuentes, Amparo Varona, Mikel Penagarikano,
Germán Bordel, Mireia Diez

Software Technologies Working Group (http://gtts.ehu.es), DEE, ZTF/FCT
University of the Basque Country UPV/EHU, Barrio Sarriena, 48940 Leioa, Spain
{luisjavier.rodriguez, amparo.varona, mikel.penagarikano, german.bordel, mireia.diez}@ehu.es

ABSTRACT

This paper briefly describes the systems presented by the Software Technologies Working Group (http://gtts.ehu.es, GTTS) of the University of the Basque Country (UPV/EHU) to the Spoken Web Search (SWS) task at MediaEval 2013. GTTS systems consist of four main modules: (1) feature extraction; (2) speech activity detection; (3) DTW-based query matching; and (4) score calibration and fusion. The most remarkable contributions are the use of phone log-likelihood ratio features, the normalization of the DTW distance matrix and the calibration/fusion approach (which is imported from language/speaker verification).

1. INTRODUCTION

The MediaEval 2013 Spoken Web Search (SWS) task consists of searching for a spoken query within a set of audio documents [4]. The locations and durations of all the occurrences of spoken queries in the audio documents must be obtained. System performance is primarily measured in terms of the Average Term-Weighted Value (ATWV) [5], but also in terms of a normalized cross-entropy metric and the processing resources (real-time factor and peak memory usage) required by the submitted systems [6]. For more details on the SWS task at MediaEval 2013, see [2].

2. SYSTEM OVERVIEW

2.1 Feature extraction

The Brno University of Technology (BUT) phone decoders for Czech, Hungarian and Russian [7] are applied to decode both the spoken queries and the audio documents. BUT decoders are trained on 8 kHz SpeechDat(E) databases recorded over fixed telephone networks, containing 12, 10 and 18 hours of speech and featuring 45, 61 and 52 units for Czech, Hungarian and Russian, respectively (three of them being non-phonetic units that stand for short pauses and noises).

Given an input signal of length T, the decoder outputs the posterior probability of each state s ($1 \leq s \leq S$) of each unit i ($1 \leq i \leq M$) at each frame t ($1 \leq t \leq T$), $p_{i,s}(t)$, where M is the number of units and S the number of states per unit. The posterior probability of each unit i at each frame t are computed by adding the posteriors of its states:

$$p_i(t) = \sum_{\forall s} p_{i,s}(t) \tag{1}$$

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

Finally, the posteriors of the three non-phonetic units are added and stored as a single *non-speech* posterior. Thus, the size of the frame-level feature vectors is 43, 59 and 50 for the Czech, Hungarian and Russian BUT decoders, respectively.

2.2 Speech Activity Detection

Given an audio signal, Speech Activity Detection (SAD) is performed by discarding those phone posterior feature vectors for which the non-speech posterior is the highest. The remaining vectors, along with their corresponding time offsets, are stored for further use, but the component corresponding to the non-speech unit is deleted. If the number of speech vectors is too low (in this evaluation, that threshold was arbitrarily set to 10, that is, 0.1 seconds), the whole signal is discarded, to save time and to avoid false alarms.

2.3 DTW-based query matching

Given two SAD-filtered sequences of feature vectors corresponding to a spoken query q and a spoken document x, the cosine distance is computed between each pair of vectors, $q[i]$ and $x[j]$ as follows:

$$d(q[i], x[j]) = -\log \frac{q[i] \cdot x[j]}{|q[i]| \cdot |x[j]|} \tag{2}$$

Note that $d(v, w) \geq 0$, with $d(v, w) = 0$ if and only if v and w are perfectly aligned and $d(v, w) = +\infty$ if and only if v and w are orthogonal. The distance matrix computed according to Eq. 2 is further normalized with regard to the spoken document x, as follows:

$$d_{norm}(q[i], x[j]) = \frac{d(q[i], x[j]) - d_{min}(j)}{d_{max}(j) - d_{min}(j)} \tag{3}$$

with $d_{min}(j) = \min_i d(q[i], x[j])$ and $d_{max}(j) = \max_i d(q[i], x[j])$. In this way, matrix values are all comprised between 0 and 1, so that a perfect match would produce a quasi-diagonal sequence of zeroes.

The best match of a query q of length m in a spoken document x of length n is defined as that minimizing the average distance in a *crossing path* of the matrix d_{norm}. A crossing path starts at any given frame of x, $k_1 \in [1, n]$, then traverses a region of x which is optimally aligned to q (involving L vector alignments), and ends at frame $k_2 \in [k_1, n]$. The average distance in this crossing path is:

$$d_{avg}(q, x) = \frac{1}{L} \sum_{l=1}^{L} d_{norm}(q[i_l], x[j_l]) \tag{4}$$

where i_l and j_l are the indices of the vectors of q and x in the alignment l, for $l = 1, 2, \ldots, L$. Note that $i_1 = 1$, $i_L = m$, $j_1 = k_1$ and $j_L = k_2$. The minimization operation

Table 1: Results of GTTS on-time and late systems submitted to the required (single-example) condition.

	development queries				evaluation queries					
	MTWV/ATWV	C_{nxe} (act/min)	SSF	PMU_s	MTWV/ATWV	C_{nxe} (act/min)	SSF	PMU_s	ISF	PMU_i
p	0.4174 / 0.4078	0.7962 / 0.6605	0.2509	0.325	0.3992 / 0.3806	0.8159 / 0.6570	0.2350	0.226	0.8015	0.027
c1	0.3601 / 0.3586	0.9976 / 0.6877	0.1219	0.325	0.3457 / 0.3430	1.0229 / 0.6838	0.1187	0.226	0.8015	0.027
c2	0.2726 / 0.2687	1.4365 / 0.7559	0.0399	0.298	0.2586 / 0.2538	1.5588 / 0.7543	0.0311	0.200	0.2473	0.023
c3	0.2704 / 0.2466	1.0274 / 0.7710	0.0457	0.302	0.2408 / 0.2221	0.9514 / 0.7665	0.0449	0.204	0.2862	0.027
c4	0.2491 / 0.2437	1.2606 / 0.7716	0.0434	0.300	0.2418 / 0.2372	1.2125 / 0.7534	0.0403	0.202	0.2680	0.024
p-late	0.4186 / 0.4164	0.6659 / 0.6603	0.2509	0.325	0.3994 / 0.3989	0.6582 / 0.6567	0.2350	0.226	0.8015	0.027
c1-late	0.3601 / 0.3590	0.6878 / 0.6877	0.1219	0.325	0.3457 / 0.3438	0.6848 / 0.6838	0.1187	0.226	0.8015	0.027
c2-late	0.2726 / 0.2722	0.7561 / 0.7559	0.0399	0.298	0.2586 / 0.2570	0.7645 / 0.7543	0.0311	0.200	0.2473	0.023

Table 2: Results of the GTTS late system submitted to the extended (multiple-example) condition.

	development queries				evaluation queries					
	MTWV/ATWV	C_{nxe} (act/min)	SSF	PMU_s	MTWV/ATWV	C_{nxe} (act/min)	SSF	PMU_s	ISF	PMU_i
c2-late	0.3038 / 0.3004	0.6845 / 0.6844	0.0192	0.298	0.2970 / 0.2939	0.6943 / 0.6942	0.0173	0.200	0.2473	0.023

is accomplished by means of a dynamic programming procedure, which is $\Theta(n \cdot m \cdot d)$ in time (d: size of feature vectors) and $\Theta(n \cdot m)$ in space. The detection score is computed as $1 - d_{avg}(q, x)$. The starting time and the duration of each detection are obtained by retrieving the time offsets corresponding to frames k_1 and k_2 in the SAD-filtered spoken document.

This procedure is iteratively applied to find not only the best match but also less likely matches in the same document. To that end, a queue of search intervals is defined and initialized with $(1, n)$. Let us consider an interval (a, b), and assume that the best match is found at (a', b'), then the intervals (a, a') and (b', b) are added to the queue (for further processing) if: (1) the score of the current match is greater than a given threshold (in this evaluation, 0.85); (2) the interval is long enough (in this evaluation, half the query length); and (3) the number of matches (already computed + pendant) is less than a given maximum (in this evaluation, 7). Finally, the list of matches for each query is truncated to the N with the highest scores (in this evaluation, $N = 1000$).

Under the extended (multiple examples) condition, only the examples passing SAD filtering (i.e. with enough speech samples) are considered for each query. The longest example is taken as reference and DTW-aligned to the other available examples. Finally, the vectors aligned at each frame are averaged and a *single average example* is obtained and processed as in the required (single example) condition.

2.4 Score calibration and fusion

System scores are transformed according to [1], which is an adaptation of the discriminative calibration/fusion approach commonly applied in speaker and language recognition.

First, the so-called *q-norm* (query normalization) is applied, so that zero-mean and unit-variance scores are obtained per query. Then, if n different systems are fused, detections are aligned so that only those supported by $n/2$ or more systems are retained for further processing (this is known as *majority voting* validation). Let us consider one of such validated detections, corresponding to a query q; if a system A does not provide a score for it, we use instead the minimum score that A has output for q. The same value is assigned to missed detections and non-target trials. In this way, a complete set of scores is prepared, which besides the ground truth (target/non-target labels) can be used to discriminatively estimate a linear transformation that produces well-calibrated scores that can be linearly combined to get fused scores. Under this approach, the Bayes optimal threshold —given by the effective prior (0.0148 for this evaluation)— is applied. The BOSARIS toolkit [3] is used to estimate and apply the calibration/fusion models.

3. RESULTS

Tables 1 and 2 show the results (performance and processing resources) for GTTS systems in the required and extended conditions, respectively. All the experiments have been carried out on a 2× Xeon E5-2450 (×8 core, ×2 HT) @2.10GHz, 64GB, under Linux Fedora 3.3.4-5.fc17.x86_64. The indexing phase involves just applying BUT decoders to extract phone posterior features. ISF, SSF and PMU values have been computed as if all the computation had been performed sequentially in a single processor (see [6]). Calibration and fusion costs have been neglected.

The contrastive systems 2, 3 and 4 (c2, c3 and c4) use the BUT decoders for Czech, Hungarian and Russian, respectively. The contrastive system 1 (c1) uses the concatenation of phone posteriors from the three decoders as features (and the average of non-speech posteriors for SAD). The primary system (p), which is the fusion of the four contrastive systems, increases MTWV in 5 absolute points (15% relative) with regard to the best contrastive (c1). In all cases, calibration and fusion parameters have been estimated on the development set. Late submissions fixed a bug in the fusion script (which did not count missed detections), thus leading to better calibrated systems. Note also that a 15% relative MTWV increase (nearly 4 absolute points) is obtained by using multiple examples under the approach described above (system c2-late).

4. REFERENCES

[1] A. Abad, L. J. Rodriguez Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel. On the calibration and fusion of heterogeneous spoken term detection systems. In *Interspeech 2013*, Lyon, France, August 25-29 2013.

[2] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L.-J. Rodriguez-Fuentes. The Spoken Web Search Task. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[3] N. Brümmer and E. de Villiers. The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing. Technical report, 2011. https://sites.google.com/site/bosaristoolkit/.

[4] MediaEval Benchmarking Initiative for Multimedia Evaluation. *The 2013 Spoken Web Search Task*, June 2013. http://www.multimediaeval.org/mediaeval2013/sws2013/.

[5] NIST. *The Spoken Term Detection (STD) 2006 Evaluation Plan*, September 2006. http://www.itl.nist.gov/iad/mig/tests/std/2006/.

[6] L.-J. Rodriguez-Fuentes and M. Penagarikano. MediaEval 2013 Spoken Web Search Task: System Performance Measures. Technical report, GTTS, UPV/EHU, May 2013. http://gtts.ehu.es/gtts/NT/fulltext/rodriguezmediaeval13.pdf.

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