GTTS-EHU Systems for QUESST at MediaEval 2014

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 Query-by-Example Search on Speech Task (QUESST) at MediaEval 2014. The GTTS-EHU systems consist of four modules: (1) feature extraction; (2) speech activity detection; (3) DTW-based query matching; and (4) score calibration and fusion. The submitted systems follow the same approach used in our SWS 2013 submissions, with two minor changes (needed to address the new task): the search stops at the most likely query detection (no further detections are looked for) and a score is produced for each (query, document) pair. The two approximate matching types introduced in QUESST have not received special treatment. This year, we have just explored the use of reduced feature sets, obtaining worse results but at lower computational costs.

1. INTRODUCTION

The MediaEval 2014 Query-by-Example Search on Speech Task (QUESST) consists of searching for a spoken query within a set of spoken documents. For each pair (query, document), a score in the range $(-\infty, +\infty)$ must be produced, the higher (the more positive) the score, the more likely that the query appears in the document. System performance is primarily measured in terms of a normalized cross-entropy cost C_{nxe}. Term-Weighted Value metrics (ATWV/MTWV) are used as secondary metrics, along with the processing resources (real-time factor and peak memory usage) required by the submitted systems. For more details on QUESST, see [2].

2. SYSTEM OVERVIEW

2.1 Feature extraction

The Brno University of Technology (BUT) phone decoders for Czech, Hungarian and Russian [6] 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, 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:

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

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.

2.1.1 Reduced feature sets

In [4], several dimensionality reduction techniques were successfully applied on phone posterior features to reduce the computational cost while keeping performance on spoken language recognition tasks. Following one of the approaches proposed in [4], here we define a reduced set of features by adding the posteriors of phones with the same manner and place of articulation. This leads to feature sets of size 25, 23 and 21, 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, 10, meaning 0.1 seconds), the whole signal is discarded (thus saving time and possibly avoiding many false alarms) and a *floor* score is output (in this evaluation, 10^{-5}).

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:

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

with $d_{min}(i) = \min_j d(q[i], x[j])$ and $d_{max}(i) = \max_j d(q[i], x[j])$.

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

Table 1: **Performance and processing costs of GTTS-EHU systems on QUESST 2014. Full sets: 2×Xeon E5-2450 (×8core×2HT) @2.10GHz, 64GB, 22892 MFlops. Reduced sets: 2×Xeon E5-649 (×6core×2HT) @2.53GHz, 24GB, 14300 MFlops.**

	development queries				evaluation queries					
	C_{nxe} (act/min)	ATWV/MTWV	SSF	PMU$_s$	C_{nxe} (act/min)	ATWV/MTWV	SSF	PMU$_s$	ISF	PMU$_i$
p	0.6540 / 0.6353	0.3567 / 0.3663	0.064	0.208	0.6207 / 0.5994	0.3621 / 0.3671	0.064	0.179	0.427	0.028
c1	0.7180 / 0.6849	0.3164 / 0.3293	0.008	0.208	0.6956 / 0.6631	0.3078 / 0.3230	0.008	0.179	0.427	0.028
c2	0.7498 / 0.7115	0.2765 / 0.2947	0.009	0.123	0.7316 / 0.6968	0.2658 / 0.2937	0.009	0.098	0.427	0.012
c3	0.6599 / 0.6408	0.3558 / 0.3588	0.027	0.208	0.6266 / 0.6053	0.3593 / 0.3624	0.027	0.179	0.427	0.028
c4	0.6987 / 0.6747	0.3235 / 0.3300	0.037	0.123	0.6707 / 0.6450	0.3146 / 0.3311	0.037	0.098	0.427	0.012

Table 2: C_{nxe} **and TWV performance of the GTTS-EHU primary system, disaggregated per matching types and per language.**

	development queries		evaluation queries	
	C_{nxe} (act/min)	ATWV/MTWV	C_{nxe} (act/min)	ATWV/MTWV
T1	0.4832 / 0.4514	0.5567 / 0.5594	0.4773 / 0.4396	0.5353 / 0.5375
T2	0.7585 / 0.7323	0.2960 / 0.3153	0.6573 / 0.6407	0.3196 / 0.3276
T3	0.7627 / 0.7390	0.1361 / 0.1493	0.8118 / 0.7724	0.1548 / 0.1620
Albanian	0.6256 / 0.5824	0.3005 / 0.3244	0.6778 / 0.6313	0.3961 / 0.4102
Basque	0.8647 / 0.8279	0.2354 / 0.2476	0.7920 / 0.7616	0.2982 / 0.3052
Czech	0.6455 / 0.6263	0.4026 / 0.4048	0.5863 / 0.5699	0.3603 / 0.3720
NNEnglish	0.9384 / 0.8741	0.0600 / 0.0707	0.9305 / 0.8478	0.1008 / 0.1061
Romanian	0.4365 / 0.4058	0.4991 / 0.5398	0.5748 / 0.5520	0.4081 / 0.4452
Slovak	0.5495 / 0.5105	0.5533 / 0.5579	0.4917 / 0.4465	0.5841 / 0.6096

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:

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

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 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 $1 - d_{avg}(q, x)$. Once the best match is obtained, the search procedure stops. As noted above, if either q or x have not enough speech samples, no alignment is performed and a *floor* score (10^{-5}) is output. Note that a detection score must be mandatorily produced for each pair (q, x).

2.4 Score calibration and fusion

First, the so-called *q-norm* (query normalization) is applied, so that zero-mean and unit-variance scores are obtained per query [1]. Then, if n different systems are fused, since all of them contain a complete set of scores, for each trial the set of n scores is considered, 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: 0.0741 for this evaluation) is applied. The BOSARIS toolkit [3] is used to estimate and apply the calibration/fusion models.

3. RESULTS

Table 1 shows the performance and processing costs of GTTS-EHU systems on QUESST 2014. To speed up computations, experiments with the full and reduced sets of features were carried out on different machines (see Table 1), which makes it nonsense to compare the reported times. Indexing involves just applying BUT decoders to extract phone posterior features. The *Indexing Speed Factor* (ISF), the *Searching Speed Factor* (SSF) and the *Peak Memory Usage* (PMU) values have been computed as if all the computation was performed sequentially in a single processor (see [5]). Calibration and fusion costs have been neglected.

The contrastive systems 1 and 2 (c1 and c2) use the concatenation of phone posteriors from the three decoders as features, for the full and reduced feature sets, respectively. The system c3 is the fusion of four subsystems, using the full set of phone posteriors for Czech, Hungarian, Russian and the concatenation of them, respectively. The system c4 is equivalent to c3 but using the reduced sets of features. Finally, the primary system is the fusion of the eight available subsystems. In all cases, calibration and fusion parameters have been estimated on the development set. Note that the primary system yields only slightly better performance than system c3, meaning that reduced sets of features provide little additional information to full sets of features. In fact, the full sets outperform the reduced sets in all cases.

As shown in Table 2, performance strongly degrades from T1 to T2 and (not so much) from T2 to T3; on the other hand, the non-native English and (to a lesser extent) the Basque subsets seem problematic. Future work may involve some kind of language detection and adaptation, plus specific strategies for matching types T2 and T3.

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, L.-J. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze. Query by Example Search on Speech at Mediaeval 2014. In *Working Notes Proceedings of the Mediaeval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[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] M. Diez, L. J. Rodriguez Fuentes, M. Penagarikano, A. Varona, and G. Bordel. Dimensionality reduction of phone log-likelihood ratio features for spoken language recognition. In *Interspeech 2013*, Lyon, France, August 25-29 2013.

[5] 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.

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