GTM-UVigo Systems for the Query-by-Example Search on Speech Task at MediaEval 2015

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo
AtlantTIC Research Center
E.E. Telecomunicación, Campus Universitario S/N, 36310 Vigo
{plopez,ldocio,carmen}@gts.uvigo.es

ABSTRACT

In this paper, we present the systems developed by GTM-UVigo team for the query by example search on speech task (QUESST) at MediaEval 2015. The systems consist in a fusion of 11 dynamic time warping based systems that use phoneme posteriorgrams for speech representation; the primary system introduces a technique to select the most relevant phonetic units on each phoneme decoder, leading to an improvement of the search results.

1. INTRODUCTION

The query by example search on speech task (QUESST) aims at searching for audio content within audio content using an audio content query [13], having special focus on low-resource languages. This paper describes the systems developed by GTM-UVigo team to address this task[1].

2. GTM-UVIGO SYSTEM DESCRIPTION

GTM-UVigo systems consist in the fusion of 11 individual systems that represent the documents and queries by means of phoneme posteriorgrams, and then subsequence dynamic time warping (S-DTW) is used to perform the search. The primary system features a phonetic unit selection strategy, which is briefly described in this Section.

2.1 Phoneme posteriorgrams

Three architectures were used to obtain phoneme posteriorgrams:

- *lstm*: a context-independent phone recognizer based on a long short-term memory (LSTM) neural network was trained using the KALDI toolkit [5]. A 2-layer LSTM was used; the input of the first layer consists of 40 log filter-bank energies augmented with 3 pitch related features [4] and the output layer dimension was the number of context independent phone units.

- *dnn*: a deep neural network (DNN)-based context-dependent phone recognizer was trained using the KALDI toolkit following Karel Veselý's DNN training implementation [15]. The network has 6 hidden layers, each with 2048 units, and it was trained on LDA-STC-fMLLR features obtained from auxiliary Gaussian mixture models (GMM) [15]. The dimension of the input layer was 440 and the output layer dimension was the number of context-dependent states.

- *traps*: the phone decoder based on long temporal context developed at the Brno University of Technology (BUT) was used [11].

11 models, summarized in Table 1, were trained using data in 6 languages: Galician (GA), Spanish (ES), English (EN), Czech (CZ), Hungarian (HU) and Russian (RU).

Table 1: Databases used to train the acoustic models. BUT models were used in the traps systems.

System	Database	Duration (h)
GAdnn, GAlstm	Transcrigal [3]	35
ESdnn, ESlstm	TC-STAR [2]	78
ENdnn, ESlstm	LibriSpeech [8]	100
CZdnn, ENlstm	Vystadial 2013 [6]	15
CZ/HU/RUtraps	Speech-Dat	n/a

2.2 Dynamic Time Warping Strategy

The search of the spoken queries within the audio documents is performed by means of S-DTW [7]. First, a cost matrix $M \in \Re^{n \times m}$ is defined, where the rows and the columns correspond to the frames of the query Q and the document D, respectively:

$$M_{i,j} = \begin{cases} c(q_i, d_j) & \text{if} \quad i = 0 \\ c(q_i, d_j) + M_{i-1,0} & \text{if} \quad i > 0,\ j = 0 \\ c(q_i, d_j) + M^*(i,j) & \text{otherwise} \end{cases} \quad (1)$$

where $c(q_i, d_j)$ represents the cost between query vector q_i and document vector d_j, both of dimension U, and

$$M^*(i,j) = min\left(M_{i-1,j}, M_{i-1,j-1}, M_{i,j-1}\right) \quad (2)$$

Pearson's correlation coefficient r is used as distance metric [14]:

$$r(q_i, d_j) = \frac{U(q_i \cdot d_j) - \|q_i\|\|d_j\|}{\sqrt{(U\|q_i^2\| - \|q_i\|^2)(U\|d_j^2\| - \|d_j\|^2)}} \quad (3)$$

In order to use r as a cost function, it is linearly mapped to the range [0,1], where 0 corresponds to correlations equal to 1 and 1 corresponds to correlations equal to -1.

[1]The code of GTM-UVigo systems will be released at `https://github.com/gtm-uvigo/MediaEval_QUESST2015`

MediaEval 2015 Workshop, Sept. 14–15, 2015, Wurzen, Germany.

Table 2: Performance of systems submitted by GTM-UVigo team.

		Dev				Eval				Dev-late				Eval-late			
System	Metric	All	T1	T2	T3	All	T1	T2	T3	All	T1	T2	T3	All	T1	T2	T3
Primary	actCnxe	0.917	0.881	0.943	0.918	0.919	0.864	0.959	0.913	0.875	0.841	0.890	0.882	0.871	0.815	0.916	0.866
	minCnxe	0.905	0.861	0.928	0.904	0.905	0.844	0.946	0.882	0.847	0.788	0.865	0.860	0.838	0.758	0.895	0.824
	lowerbound	0.627	0.562	0.672	0.631	0.629	0.532	0.702	0.627	0.593	0.526	0.633	0.606	0.592	0.490	0.657	0.601
Contrastive	actCnxe	0.998	0.998	0.997	1.000	0.999	0.999	0.997	1.000	0.907	0.897	0.916	0.904	0.898	0.852	0.933	0.896
	minCnxe	0.918	0.874	0.942	0.898	0.923	0.865	0.953	0.907	0.864	0.811	0.877	0.880	0.852	0.785	0.900	0.843
	lowerbound	0.635	0.588	0.681	0.627	0.633	0.555	0.693	0.624	0.618	0.559	0.655	0.633	0.613	0.521	0.669	0.622

In order to detect n_c candidate matches of a query in a spoken document, every time a candidate match is detected, which ends at frame b^*, $M(n, b^*)$ is set to ∞ in order to ignore this match.

2.3 Phoneme Unit Selection

A technique to select the most relevant phonemes among the phonetic units of the different decoders was used in the primary system. Given the best alignment path P(Q,D) of length K between a query and a matching document, the correlation and the cost at each step of the path can be decomposed so there is a different term for each phonetic unit u:

$$r(q_i, d_j, u) = \frac{U q_{i,u} d_{j,u} - \frac{1}{U}\|q_i\|\|d_j\|}{\sqrt{(U\|q_i^2\| - \|q_i\|^2)(U\|d_j^2\| - \|d_j\|^2)}} \quad (4)$$

In this way, the cost accumulated by each phonetic unit through the best alignment path can be computed:

$$R(P(Q,D), u) = \frac{1}{K}\sum_{k=1}^{K} c(q_{i_k}, d_{j_k}, u) \quad (5)$$

This value $R(P(Q,D), u)$ can be considered as the relevance of the phonetic unit u (the lower the contribution to the cost, the more relevant the phonetic unit). Hence, the phonetic units can be sorted from more relevant to less relevant in order to keep the most relevant ones and to discard those who increased the cost of the best alignment path.

Using only one alignment path may not provide a good estimate of the relevance of the phonetic units; hence, the relevance of the different pairs query-matching document in the development set were accumulated in order to robustly estimate the relevance. The number of relevant phonetic units was empirically selected for each system.

2.4 Normalization and fusion

Score normalization and fusion were performed following [12]. First, the scores were normalized by the length of the warping path. A binary logistic regression was used for fusion, as described in [1].

3. RESULTS AND DISCUSSION

Table 2 shows the results obtained on QUESST 2015 data using the submitted systems. The Table shows that the primary system, that features phoneme unit selection, clearly outperforms the contrastive system, suggesting that the proposed technique obtains the expected improvement. Another fact that can be observed is that Dev and Eval results are very similar, showing the generalization capability of the systems. Late systems feature z-norm normalization of the query scores, obtaining an improvement with respect to the original submissions, where only path-length normalization was applied. In Table 3, actCnxe obtained with and without applying the phoneme unit selection approach in some individual systems are compared.

Table 3: actCnxe of some individual systems with and without applying phoneme unit selection.

	With				Without			
System	Global	T1	T2	T3	Global	T1	T2	T3
CZdnn	0.889	0.829	0.902	0.906	0.915	0.867	0.927	0.922
CZlstm	0.902	0.864	0.922	0.901	0.907	0.864	0.932	0.904
CZtraps	0.902	0.840	0.924	0.910	0.931	0.883	0.945	0.938
HUtraps	0.903	0.856	0.926	0.899	0.934	0.894	0.950	0.936
RUtraps	0.895	0.844	0.918	0.894	0.925	0.886	0.944	0.922

Table 4 shows the indexing speed factor (ISF), searching speed factor (SSF), peak memory usage for indexing (PMU_I) and searching (PMU_S) and processing load (PL)[2], computed as described in [9]. ISF and PMU_I are rather high because, in the dnn systems, first an automatic speech recognition system (ASR) is applied in order to obtain the input features to the DNN; hence, the peak memory usage is so large due to the memory requirements of the language model, and the large computation time is caused by the two recognition steps that are performed to estimate the transformation matrix used to obtain the fMLLR features that are the input to the DNN. In future work, the ASR step of dnn systems will be replaced with a phonetic network in order to avoid these time and memory consuming steps.

Table 4: Required amount of processing resources.

ISF	SSF	PMU_I	PMU_S	PL
12.1	0.09	6	0.014	7.3

4. ACKNOWLEDGEMENTS

This research was funded by the Spanish Government ('SpeechTech4All Project' TEC2012-38939-C03-01), the Galician Government through the research contract GRC2014/024 (Modalidade: Grupos de Referencia Competitiva 2014) and 'AtlantTIC Project' CN2012/160, and also by the Spanish Government and the European Regional Development Fund (ERDF) under project TACTICA.

[2]These values were computed using 2xIntel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz, 12cores/24threads, 128GB RAM.

5. REFERENCES

[1] M. Akbacak, L. Burget, W. Weng, and J. Houtvan. Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams. In *Proceedings of ICASSP*, pages 8267–8271, 2013.

[2] L. Docio-Fernandez, A. Cardenal-Lopez, and C. Garcia-Mateo. TC-STAR 2006 automatic speech recognition evaluation: The UVIGO system. In *TC-STAR Workshop on Speech-to-Speech Translation*, 2006.

[3] C. Garcia-Mateo, J. Dieguez-Tirado, L. Docio-Fernandez, and A. Cardenal-Lopez. Transcrigal: A bilingual system for automatic indexing of broadcast news. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, 2004.

[4] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pages 2494–2498, 2014.

[5] A. Graves, A. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 6645–6649, 2013.

[6] M. Korvas, O. Plátek, O. Dušek, L. Žilka, and F. Jurčíček. Vystadial 2013 - Czech data, 2014. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

[7] M. Müller. *Information Retrieval for Music and Motion*. Springer-Verlag, 2007.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. LibriSpeech: an ASR corpus based on public domain audio books. In *Proceedings of ICASSP*, pages 99–105, 2015.

[9] L. Rodriguez-Fuentes and M. Penagarikano. MediaEval 2013 spoken web search task: system performance measures. Technical report, Dept. Electricity and Electronics, University of the Basque Country, 2013.

[10] L. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez. GTTS systems for the SWS task at MediaEval 2013. In *Proceedings of the MediaEval 2013 Workshop*, 2013.

[11] P. Schwarz. *Phoneme Recognition based on Long Temporal Context*. PhD thesis, Brno University of Technology, 2009.

[12] I. Szöke, L. Burget, F. Grézl, J. Černocký, and L. Ondel. Calibration and fusion of query-by-example systems - BUT SWS 2013. In *Proceedings of ICASSP 2014*, pages 7899–7903. IEEE Signal Processing Society, 2014.

[13] I. Szöke, L. Rodriguez-Fuentes, A. Buzo, X. Anguera, F. Metze, J. Proenca, M. Lojka, and X. Xiong. Query by example search on speech at Mediaeval 2015. In *Proceedings of the MediaEval 2015 Workshop*, 2015.

[14] I. Szöke, M. Skácel, and L. Burget. BUT QUESST2014 system description. In *Proceedings of the MediaEval 2014 Workshop*, 2014.

[15] K. Veselý, A. Ghoshal, L. Burget, and D. Povey. Sequence-discriminative training of deep neural networks. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pages 2345–2349, 2013.