DCU Search Runs
at MediaEval 2014 Search and Hyperlinking

David N. Racca, Maria Eskevich, Gareth J.F. Jones
CNGL Centre for Global Intelligent Content
School of Computing
Dublin City University, Ireland
{dracca, meskevich, gjones}@computing.dcu.ie

ABSTRACT

We described Dublin City University (DCU)'s participation in the Search sub-task of the Search and Hyperlinking Task at MediaEval 2014. Exploratory experiments were carried out to investigate the utility of prosodic prominence features in the task of retrieving relevant video segments from a collection of BBC videos. Normalised acoustic correlates of loudness, pitch, and duration were incorporated in a standard TF-IDF weighting scheme to increase weights for terms that were prominent in speech. Prosodic models outperformed a text-based TF-IDF baseline on the training set but failed to surpass the baseline on the test set.

1. INTRODUCTION

Increasing amounts of multimedia content are being produced and stored on a daily basis. In order to make this data useful, computer applications are required that facilitate search, browsing, and navigation through these large data collections. The MediaEval *Search and Hyperlinking task* seeks to contribute to addressing this problem.

In contrast with previous years where a known-item task was examined, this year an ad-hoc search task was introduced. The retrieval collection consists of an extension of last year's collection, comprising 4021 hours of BBC TV Broadcast content split into training and test sets of 1335 and 2686 hours respectively. For every video file in the collection, the organizers provided human-generated subtitles, three different automatic speech recognition (ASR) transcripts (LIMSI/Vocapia, LIUM, and NST-Sheffield), prosodic features, shot boundaries, visual concept detection output, and additional metadata associated with each TV-show. The training set includes 50 text queries while the test set comprises 30 queries. More details about the data collection and task evaluation metrics can be found in [4].

Previous research has demonstrated that prosodic information is useful for a wide range of speech processing tasks [6], including speech search tasks. In [2], Crestani suggests that there might be a direct relationship between acoustic stress of terms and their TF-IDF score in the *OGI Stories Corpus*, while Chen reports improvements on a spoken document retrieval task by using energy and durational features [1]. In [5], Guinaudeau and Hirschberg improve a topic tracking system by incorporating intensity and pitch values into the retrieval weighting scheme.

This paper describes an implementation of an approach that incorporates loudness, duration, and pitch into TF-IDF weights in order to examine their potential to improve retrieval effectiveness of video segments.

2. FEATURE PROCESSING

Following Guinaudeau's method [5], loudness and pitch correlates were extracted from the speech signal and normalised and aligned to each word occurrence in the transcripts. To perform this alignment, word timestamps were used in the case of LIMSI/Vocapia and NST-Sheffield transcripts. For subtitles, word timestamps had to be approximated from each segment's starting and ending timestamps. This was done by dividing the number of words included in a segment by its length to obtain the average word duration for that segment. Starting times and duration of words were then approximated by considering the starting time of a segment plus multiples of its average word duration. In the case of the LIUM transcript, duration of words was approximated for the test set by the average word duration of all words in the training set.

After the alignment was performed, minimum, maximum, mean, and standard deviation of loudness and pitch were computed for each word. These four statistics were normalised in order to be compared against other words spoken in different acoustic conditions. The final objective was to calculate an acoustic score for each spoken word that represents how salient a word is relative to its surounding context. With this in mind, two different definitions of surounding context for a word were then considered:

- Context given by the words that belong to the same speech segment predicted by the ASR (seg).
- Full length of document, this is, all the words spoken in the video (doc).

Finally, two normalisation functions were explored for normalising a feature f_i over a context C:

1. Range: $(f_i - \min_C)/(\max_C - \min_C)$.
2. Z-score: $(f_i - \mu_C)/\sigma_C$.

3. RETRIEVAL FRAMEWORK

Text transcripts were segmented into fixed-time adjacent (non-overlapping) segments of 90 seconds duration. Before indexing, stop words from the standard Terrier list [7] were removed and Porter stemming applied. Segments were then indexed using a modified version of Terrier-3.5 that associated acoustic features with term occurrences in the inverted

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

Transcript Type	Normalisation Type		Run Parameters			Results					
	Function	Context	Weighting Scheme	θ_{ir}	θ_{ac}	MAP	MAP-bin	MAP-tol	P@5	P@10	P@20
Subtitles	-	-	TF-IDF	1	0	0.639	0.394	0.293	0.727	0.627	0.478
	range	seg	G-pr	2	3	0.599	0.371	0.278	0.727	0.583	0.452
	z-score	doc	G-lp	3	1	0.533	0.336	0.265	0.633	0.570	0.433
	-	-	G-dur	1	3	0.345	0.206	0.154	0.367	0.327	0.280
NST-Sheffield	-	-	TF-IDF	1	0	0.440	0.295	0.215	0.600	0.513	0.393
	range	seg	**G-pr**	2	3	0.434	0.294	0.214	0.613	0.510	0.393
	z-score	doc	G-lp	3	1	0.435	0.294	0.218	0.587	0.503	0.393
	-	-	G-dur	1	3	0.404	0.272	0.199	0.567	0.457	0.363
LIMSI	-	-	TF-IDF	1	0	0.525	0.339	0.242	0.620	0.543	0.430
	range	seg	G-pr	2	3	0.508	0.331	0.239	0.607	0.543	0.423
	z-score	doc	G-lp	3	1	0.428	0.283	0.201	0.460	0.457	0.370
	-	-	**G-dur**	1	3	0.505	0.330	0.237	0.607	0.537	0.413
LIUM	-	-	TF-IDF	1	0	0.451	0.300	0.233	0.693	0.573	0.430
	range	seg	G-pr	2	3	0.444	0.293	0.222	0.653	0.527	0.418
	z-score	doc	**G-lp**	3	1	0.436	0.291	0.215	0.633	0.547	0.412
	-	-	G-dur	1	3	0.358	0.240	0.186	0.540	0.453	0.357

Table 1: Evaluation results over the test set. Overlap MAP [MAP], Binned MAP [MAP-bin], and Tolerance to Irrelevance MAP [MAP-tol] are shown in the Results columns. Precision at different cut-offs are based on a "Tolerance to Irrelevance" definition of relevance.

index. Note that due to stemming, multiple words can be mapped to the same stem. In these cases, acoustic feature vectors associated with each non-stemmed word ccurrence in a segment were treated as belonging to the same stem and thus were linked with this term in the inverted index.

Retrieval was performed using Terrier's standard implementation of the vector space model (VSM) with a modified TF-IDF weighting function that takes into account the acoustic features from the inverted index when computing term weights. The weight of a term t in a segment was computed using Guinaudeau's harmonic mean [5]:

$$w(t) = \frac{\theta_{ir} * \mathrm{idf}_t * \mathrm{tf}_t + \theta_{ac} * \mathrm{ac}_t}{\theta_{ir} + \theta_{ac}}$$

Different definitions were explored for the acoustic score (ac_t). In all cases, ac_t was intended to represent the level of salience of t from its surrounding context. In particular, simple multiplications of the maximum loudness and maximum pitch (G-lp), pitch range considering the maximum and minimum pitch (G-pr), and the maximum duration (G-dur) with which t was pronounced in the segment were used as definitions for ac_t. Values for the free parameters θ_{ir} and θ_{ac} were selected to optimise mean reciprocal rank (MRR), mean generalised average presicion (mGAP), and mean average segment precision (MASP) [3] on the training set for individual ASR transcripts and normalisation type. Specifically, the runs G-lp, G-pr, and G-dur were optimised for the LIUM, NST-Sheffield, and LIMSI transcripts respectively.

4. RESULTS AND CONCLUSIONS

Table 1 shows details of submitted runs and presents a summary of evaluation results over the test set for every type of transcript. Runs that made use of prosodic information for computing term weights, namely G-pr, G-lp, and G-dur, clearly underperformed the baseline TF-IDF system in general. Given the fact that the baseline system can be beaten on the training set, results over the test set suggest that models optimised on MRR, mGAP, and MASP for a known-item task evaluation scheme, as it is the case of the training set, fail to generalise to an ad-hoc retrieval task performed over the test set.

It is important to note here, that text queries for training and test sets were produced with different objectives in mind. This could be another reason why the models presented in this work seem to have overfitted the training set.

In future work, an error analysis will be carried out in order to identify queries for which prosodic-based models could have outperformed the baseline.

5. ACKNOWLEDGMENTS

This work was supported by Science Foundation Ireland (Grant 12/CE/I2267) as part of the Centre for Global Intelligent Content CNGL II project at DCU.

6. REFERENCES

[1] B. Chen, H.-M. Wang, and L.-S. Lee. Improved spoken document retrieval by exploring extra acoustic and linguistic cues. In *Proceedings Interspeech'01*, pages 299–302, Aalborg, Denmark, 2001.

[2] F. Crestani. Towards the use of prosodic information for spoken document retrieval. In *Proceedings ACM SIGIR'01*, pages 420–421, New Orleans, LA, USA, 2001.

[3] M. Eskevich, R. Aly, R. Ordelman, S. Chen, and G. J. F. Jones. The search and hyperlinking task at MediaEval 2013. In *Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[4] M. Eskevich, R. Aly, D. N. Racca, R. Ordelman, S. Chen, and G. J. F. Jones. The search and hyperlinking task at MediaEval 2014. In *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

[5] C. Guinaudeau and J. Hirschberg. Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news. In *Proceedings Interspeech'11*, pages 1401–1404, Florence, Italy, 2011.

[6] J. Hirschberg. Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1):31–43, 2002.

[7] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in Terrier: a search engine for advanced retrieval on the web. *Novatica/UPGRADE Special Issue on Next Generation Web Search*, pages 49–56, 2007.