IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

GTM-IRLab Systems for Albayzin 2018 Search on Speech Evaluation

Paula Lopez-Otero[1], *Laura Docio-Fernandez*[2]

[1]Universidade da Coruña - CITIC, Information Retrieval Lab
[2]Universidade de Vigo - atlanTTic Research Center, Multimedia Technology Group
paula.lopez.otero@udc.gal, ldocio@gts.uvigo.es

Abstract

This paper describes the systems developed by the GTM-IRLab team for the Albayzin 2018 Search on Speech evaluation. The system for the spoken term detection task consists in the fusion of two subsystems: a large vocabulary continuous speech recognition strategy that uses the proxy words approach for out-of-vocabulary terms, and a phonetic search system based on the probabilistic retrieval model for information retrieval. The query-by-example spoken term detection system is the result of fusing four subsystems: three of them are based on dynamic time warping search using different representations of the waveforms, namely Gaussian posteriorgrams, phoneme posteriorgrams and a large set of low-level descriptors; and the other one is the phonetic search system used for spoken term detection with some modifications to manage spoken queries.

Index Terms: Spoken term detection, query-by-example spoken term detection, large vocabulary continuous speech recognition, out-of-vocabulary terms, phoneme posteriorgrams, Gaussian posteriorgrams, probabilistic information retrieval, phonetic search

1. Introduction

In this paper, the systems developed by the GTM-IRLab team for the Albayzin 2018 Search on Speech evaluation are described. Systems were submitted to the spoken term detection (STD) and the query-by-example spoken term detection (QbESTD) tasks.

In the STD task, a fusion of two subsystems was proposed. The first system consists in a strategy based on large vocabulary continuous speech recognition (LVCSR). This LVCSR system was built using the Kaldi toolkit [1] to train a set of acoustic models, to generate the output lattices and to perform lattice indexing and term search [2]. The proxy words strategy described in [3] was used to deal with out-of-vocabulary (OOV) terms. The second system performs phonetic search (PS) using an approach that adapts the probabilistic retrieval model [4] for information retrieval to the search on speech task similarly as described in [5, 6].

For the QbESTD task, the proposed system consists in a fusion of four different systems. Three of them rely on dynamic time warping (DTW) search with different representations of the speech data, namely phoneme posteriorgrams [7], low-level descriptors [8] and Gaussian posteriorgrams [9]. The fourth system is an adaptation of the PS approach used for STD that copes with spoken queries.

The rest of this paper is organized as follows: Section 2 and 3 describe the systems for the STD and QbESTD tasks, respectively; Section 4 presents the preliminary results obtained for the different tasks on the development data; and Section 5 presents some conclusions extracted from the experimental validation of the different systems.

2. Spoken term detection system

The proposed STD system consists in the fusion of two subsystems: one is based in a LVCSR system while the other is a PS system that adapts the probabilistic retrieval model for information retrieval to the STD task.

2.1. LVCSR system

An LVCSR system was built using the Kaldi open-source toolkit [1]. Deep neural network (DNN) based acoustic models were used; specifically, a DNN-based context-dependent speech recognizer was trained following Karel Veselý's DNN training approach [10]. The input acoustic features to the neural network are 40 dimensional Mel-frequency cepstral coefficients (MFCCs) augmented with three pitch and voicing related features [11], and appended with their delta and acceleration coefficients. The DNN has 6 hidden layers with 2048 neurons each. Each speech frame is spliced across ± 5 frames to produce 1419 dimensional vectors which are the input to the first layer, and the output layer is a soft-max layer representing the log-posteriors of the context-dependent HMM states.

The Kaldi LVCSR decoder generates word lattices [12] using the above DNN-based acoustic models. These lattices are processed using the lattice indexing technique described in [2] so that the lattices of all the utterances in the search collection are converted from individual weighted finite state transducers (WFST) to a single generalized factor transducer structure in which the start-time, end-time and lattice posterior probability of each word token is stored as a 3-dimensional cost. This factor transducer is actually an inverted index of all word sequences seen in the lattices. Thus, given a list of keywords or phrases, a simple finite state machine is created such that it accepts the keywords/phrases and composes them with the factor transducer to obtain all the occurrences of the keywords/phrases in the search collection.

The proxy words strategy included in Kaldi [3] was used for OOV term detection. This approach uses phone confusion to find the in-vocabulary (INV) term that is the most similar, in terms of its phonetic content, to the corresponding OOV term, and search is performed using the INV term.

The data used to train the acoustic models of this LVCSR system were extracted from the Spanish material used in the 2006 TC-STAR automatic speech recognition evaluation campaign[1] and from the Galician broadcast news database Transcrigal [13]. It must be noted that all the non-speech parts as well as the speech parts corresponding to transcriptions with pronunciation errors, incomplete sentences and short speech utterances were discarded, so in the end the acoustic training material consisted of approximately 104 hours and 30 minutes.

The language model (LM) was constructed using a text database of 150 MWords composed of material from several

[1]http://www.tc-star.org

10.21437/IberSPEECH.2018-50

sources (transcriptions of European and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, online courses and transcriptions of the Mavir sessions included in the development set[2] [14]). Specifically, two fourgram-based language models were trained following the Kneser-Ney discounting strategy using the SRILM toolkit [15], and the final LM was obtained by mixing both LMs using the SRILM static n-gram interpolation functionality. One of the LMs was trained using the RTVE2018 subtitles data provided for the Albayzin 2018 Text-to-Speech challenge and the other LM was built using the other text corpora. The LM vocabulary size was limited to the most frequent 300K words and, for each search task, the set of OOV keywords were removed from the language model.

2.2. PS system

A system based on phonetic search following the probabilistic retrieval model for information retrieval was developed for the STD task:

- Indexing. First, the phone transcription of each document is obtained, and then the documents are indexed in terms of phone n-grams of different size [16, 5]. According to the probabilistic retrieval model, each document is represented by means of a language model [4]. In this case, given that the phone transcriptions have errors, several hypotheses for the best transcription are used to improve the quality of the language model [6]. The start time and duration of each phone are also stored in the index.
- Search. First, a phonetic transcription of the query is obtained using the grapheme-to-phoneme model for Spanish included in Cotovia [17]. Then, the query is searched within the different indices, and a score for each document is computed following the query likelihood retrieval model [18]. It must be noted that this model sorts the documents according to how likely they contain the query, but the start and end times of the match are required in this task. To obtain these times, the phone transcription of the query is aligned to that of the document by computing their minimum edit distance, and this allows the recovery of the start and end times since they are stored in the index. In addition, the minimum edit distance is used to penalize the score returned by the query likelihood retrieval model as described in [6].

The minimum and maximum size of the n-grams were set to 1 and 5, respectively, according to [5]. The different hypotheses for the phone transcriptions of the documents were extracted from the phone lattice obtained employing the LVCSR system described above, and the number of hypotheses to be used for indexing was empirically set to 40. Indexing and search were performed using Lucene[3].

2.3. Fusion

Discriminative calibration and fusion were applied in order to combine the outputs of the different STD systems [19]. The global minimum score produced by the system for all queries was used to hypothesize the missing scores. After normalization, calibration and fusion parameters were estimated by logistic regression on a development dataset in order to obtain improved discriminative and well-calibrated scores [20]. Calibration and fusion training was performed using the Bosaris toolkit [21].

3. Query-by-example spoken term detection system

The primary system submitted for the QbESTD evaluation consists in the fusion of four systems. Three of those systems follow the same scheme: first, feature extraction is performed in order to represent the queries and documents by means of feature vectors; then, the queries are searched within the documents using a search approach based on DTW; finally, a score normalization step is performed. The other system is an adaptation of the PS system described above to the QbESTD task.

3.1. DTW-based systems

3.1.1. Speech representation

Three different approaches for speech representation were used; given a query Q with n frames (and equivalently, a document D with m frames), these representations result in a set $\mathrm{Q} = \{\mathrm{q}_1, \ldots, \mathrm{q}_n\}$ of n vectors of dimension U (and equivalently, a set $\mathrm{D} = \{\mathrm{d}_1, \ldots, \mathrm{d}_m\}$ of m vectors of dimension U):

- Phoneme posteriorgram (PhnPost). One subsystem relies on phoneme posteriorgrams [7] for speech representation: given a speech document and a phoneme recogniser with U phonetic units, the a posteriori probability of each phonetic unit is computed for each time frame, leading to a set of vectors of dimension U that represent the probability of each phonetic unit at every time instant. The English (EN) phone decoder developed by the Brno University of Technology was used to obtain phoneme posteriorgrams; in this decoder, each phonetic unit has three different states and a posterior probability is output for each of them, so they were combined in order to obtain one posterior probability for each unit [22]. After obtaining the posteriors, a Gaussian softening was applied in order to have Gaussian distributed probabilities [23].
- Low-level descriptors (LLD). A large set of features, summarised in Table 1, was used to represent the queries and documents; these features, obtained using the OpenSMILE feature extraction toolkit [24], were extracted every 10 ms using a 25 ms window, except for F0, probability of voicing, jitter, shimmer and HNR, for which a 60 ms window was used.
- Gaussian posteriorgram (GP). Gaussian posteriorgrams [9] were used to represent the audio documents and queries. Given a Gaussian mixture model (GMM) with U Gaussians, the a posteriori probability of each Gaussian is computed for each time frame, leading to a set of vectors of dimension U that represent the probability of each Gaussian at every time instant. In this system, 19 MFCCs were extracted from the waveforms, accompanied with their energy, delta and acceleration coefficients. Feature extraction and Gaussian posteriorgram computation were performed using the Kaldi toolkit [1]. The GMM was trained using MAVIR training and development data, as well as RTVE development recordings.

[2] http://cartago.lllf.uam.es/mavir/index.pl?m=descargas

[3] http://lucene.apache.org

Table 1: *Acoustic features used in the proposed search on speech system.*

Description	# features
Sum of auditory spectra	1
Zero-crossing rate	1
Sum of RASTA style filtering auditory spectra	1
Frame intensity	1
Frame loudness	1
Root mean square energy and log-energy	2
Energy in frequency bands 250-650 Hz (energy 250-650) and 1000-4000 Hz	2
Spectral Rolloff points at 25%, 50%, 75%, 90%	4
Spectral flux	1
Spectral entropy	1
Spectral variance	1
Spectral skewness	1
Spectral kurtosis	1
Psychoacoustical sharpness	1
Spectral harmonicity	1
Spectral flatness	1
Mel-frequency cepstral coefficients	16
MFCC filterbank	26
Line spectral pairs	8
Cepstral perceptual linear predictive coefficients	9
RASTA PLP coefficients	9
Fundamental frequency (F0)	1
Probability of voicing	1
Jitter	2
Shimmer	1
log harmonics-to-noise ratio (logHNR)	1
LCP formant frequencies and bandwidths	6
Formant frame intensity	1
Deltas	102
Total	204

3.1.2. Search algorithm

The search stage was carried out using the subsequence DTW (S-DTW) [25] variant of the classical DTW approach. To perform S-DTW, first a cost matrix $\mathrm{M} \in \Re^{\mathrm{n}\times\mathrm{m}}$ must be defined, in which the rows and columns correspond to the query and document frames, respectively:

$$\mathrm{M}_{i,j} = \begin{cases} c(\mathrm{q}_i, \mathrm{d}_j) & \text{if} \quad i = 0 \\ c(\mathrm{q}_i, \mathrm{d}_j) + \mathrm{M}_{i-1,0} & \text{if} \quad i > 0, j = 0 \\ c(\mathrm{q}_i, \mathrm{d}_j) + \mathrm{M}^*(i,j) & \text{else} \end{cases} \quad (1)$$

where $c(\mathrm{q}_i, \mathrm{d}_j)$ is a function that defines the cost between the query vector q_i and the document vector d_j, and

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

Pearson's correlation coefficient r [26] was the metric used to define the cost function by mapping it into the interval [0,1] applying the following transformation:

$$c(\mathrm{q}_i, \mathrm{d}_j) = \frac{1 - r(\mathrm{q}_i, \mathrm{d}_j)}{2} \quad (3)$$

Once matrix M is computed, the end of the best warping path between Q and D is obtained as

$$\mathrm{b}^* = \underset{b\in 1,\ldots,\mathrm{m}}{\arg\min}\, \mathrm{M}(\mathrm{n}, b) \quad (4)$$

The starting point of the path ending at b^*, namely a^*, is computed by backtracking, hence obtaining the best warping path $\mathrm{P}(\mathrm{Q},\mathrm{D}) = \{\mathrm{p}_1, \ldots, \mathrm{p}_k, \ldots, \mathrm{p}_\mathrm{K}\}$, where $\mathrm{p}_k = (i_k, j_k)$, i.e. the k-th element of the path is formed by q_{i_k} and d_{j_k}, and K is the length of the warping path.

It is possible that a query Q appears several times in a document D, especially if D is a long recording. Hence, not only the best warping path must be detected but also others that are less likely. One approach to overcome this issue consists in detecting a given number of candidate matches $\mathrm{n_c}$: every time a warping path, that ends at frame b^*, is detected, $\mathrm{M}(\mathrm{n}, \mathrm{b}^*)$ is set to ∞ in order to ignore this element in the future.

A score must be assigned to every detection of a query Q in a document D. First, the cumulative cost of the warping path M_{n,b^*} is length-normalized [27] and, after that, z-norm is applied so that all the scores of all the queries have the same distribution [28].

3.2. PS system

The system described in Section 2.2 was also used for QbESTD. Since, in this experimental setup, the queries are spoken, the LVCSR system described in Section 2.1 was used to obtain phone transcriptions of the queries. In this system, the number of transcription hypotheses of the documents was empirically set to 50.

3.3. Fusion

The fusion strategy described in Section 2.3 was used to combine the QbESTD systems described in this section.

4. Preliminary Results

The systems described in the previous sections were evaluated in terms of the average term weighted value (ATWV) and maximum term weighted value (MTWV), which are the evaluation metrics defined for Albayzin 2018 evaluation. The results included in this section were achieved using the development data provided by the organizers. Since two different datasets (MAVIR and RTVE) were used for development, and in order to avoid overfitting when choosing the decision threshold, the groundtruth labels of MAVIR and RTVE were joined into a single set (namely MAVIR+RTVE) to compute the decision

Table 2: *STD results on development data*

	MAVIR		RTVE		MAVIR+RTVE	
System	MTWV	ATWV	MTWV	ATWV	MTWV	ATWV
LVCSR (con1)	0.5314	0.5179	0.5976	0.5798	0.5992	0.5991
PS (con2)	0.4828	0.4739	0.6286	0.5993	0.6173	0.6167
LVCSR-NP (con3)	0.5068	0.4079	0.5801	0.5794	0.5704	0.5700
Fusion (pri)	**0.5470**	**0.5290**	**0.6550**	**0.6183**	**0.6826**	**0.6791**

Table 3: *QbESTD results on development data*

	MAVIR		RTVE		MAVIR+RTVE	
System	MTWV	ATWV	MTWV	ATWV	MTWV	ATWV
PhnPost (con2)	0.1971	0.1742	0.7145	0.7081	0.5180	0.5160
LLD	0.2017	0.1774	0.7136	0.7114	0.5156	0.5136
GP	0.1877	0.1628	0.6731	0.6718	0.4856	0.4841
PS (con3)	0.2383	0.2029	0.3540	0.3528	0.3519	0.3507
Fusion DTW (con1)	0.2699	**0.2649**	0.7211	**0.7076**	0.5471	0.5451
Fusion (pri)	**0.2896**	0.2470	**0.7273**	0.6964	**0.6195**	**0.6174**

threshold, which was subsequently applied to each dataset individually.

4.1. STD experiments

Table 2 shows the results achieved using the systems described in Section 2. The Table also includes an additional system, namely LVCSR-NP, which consists in the aforementioned LVCSR without using the proxy words strategy for OOV terms; this means that the LVCSR-NP system does not detect any OOV terms. Comparing LVCSR-NP and LVCSR systems, it can be seen that using the proxy words strategy is beneficial specially when dealing with MAVIR data. The table also shows that the PS system outperforms the LVCSR system on RTVE dataset, and it also leads to a better overall result. The combination of both systems achieves a significant improvement in all the experimental conditions, which suggests that both strategies are strongly complementary.

4.2. QbESTD experiments

Table 3 shows the results achieved by the QbESTD systems described in Section 3. The best performance in MAVIR data was achieved with the PS system, which also exhibited the lowest performance in RTVE data. PhnPost and LLD systems achieved almost the same results for RTVE and MAVIR+RTVE data.

The table also displays the results obtained when fusing the three DTW approaches (Fusion DTW) and when fusing the four systems (Fusion). The MTWV is always higher when fusing the four systems but, for the individual datasets, the ATWV is higher when fusing only the DTW systems. Nevertheless, the overall result is better when combining the four systems, so this system was selected as the primary (pri) for this evaluation, while the fusion of the three DTW systems was presented as contrastive (con1).

5. Conclusions and future work

This paper presented the systems developed for the STD and QbESTD tasks of Albayzin 2018 Search on Speech evaluation. The STD system consists in a fusion of a LVCSR system with a phonetic search approach based on the probabilistic retrieval model for information retrieval. The LVCSR system relied on the proxy words approach for OOV words, which were also managed by the phonetic search system. The QbESTD system is a fusion of three DTW-based systems with the phonetic search system used in the STD task.

The performance obtained in STD and QbESTD tasks are not straightforwardly comparable because the queries used to compute the evaluation metrics are not the same for both tasks, but the results suggest that spoken queries lead to better results in RTVE dataset. This might be caused by a greater amount of OOV words, so this will be investigated by further analysis of the results.

In future work, a system that combines word-level and phone-level representations with the probabilistic retrieval model for information retrieval will be assessed. This idea is motivated by the fact that, according to the results exhibited in the STD task, the LVCSR and phonetic search systems are strongly complementary, and designing smart combination strategies might improve the performance of logistic regression fusion.

The DTW-based systems for QbESTD used in this paper are language-independent, i.e. the system can be used regardless the language spoken in the recordings. Given that a LVCSR system for Spanish was trained for the STD system, the use of the activations of the LVCSR network will be investigated in future work in order to assess QbESTD performance in a language-dependent setting.

6. Acknowledgements

This work has received financial support from i) "Ministerio de Economía y Competitividad" of the Government of Spain and the European Regional Development Fund (ERDF) under the research projects TIN2015-64282-R and TEC2015-65345-P, ii) Xunta de Galicia (projects GPC ED431B 2016/035 and GRC 2014/024), and iii) Xunta de Galicia - "Consellería de Cultura, Educación e Ordenación Universitaria" and the ERDF through the 2016-2019 accreditations ED431G/01 ("Centro singular de investigación de Galicia") and ED431G/04 ("Agrupación estratéxica consolidada").

7. References

[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[2] D. Can and M. Saraclar, "Lattice indexing for spoken term detection." *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[3] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," in *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU*, 2013, pp. 416–421.

[4] J. Ponte and W. Croft, "A language modeling approach to information retrieval," in *Proceedings of ACM SIGIR*, 1998, pp. 275–281.

[5] P. Lopez-Otero, J. Parapar, and A. Barreiro, "Efficient query-by-example spoken document retrieval combining phone multigram

representation and dynamic time warping," *Information Processing and Management*, vol. 56, pp. 43–60, 2019.

[6] ——, "Probabilistic information retrieval models for query-by-example spoken document retrieval," *Speech Communication (submitted)*, 2018.

[7] T. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU*, 2009, pp. 421–426.

[8] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Finding relevant features for zero-resource query-by-example search on speech," *Speech Communication*, vol. 84, pp. 24–35, 2016.

[9] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU*, 2009, pp. 398–403.

[10] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013).*, no. 8, 2013, pp. 2345–2349.

[11] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of ICASSP*, 2014, pp. 2494–2498.

[12] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiát, S. Kombrink, P. Motlícek, Y. Qian, K. Riedhammer, K. Veselý, and N. T. Vu, "Generating exact lattices in the WFST framework," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2012, pp. 4213–4216.

[13] C. Garcia-Mateo, J. Dieguez-Tirado, L. Docio-Fernandez, and A. Cardenal-Lopez, "Transcrigal: A bilingual system for automatic indexing of broadcast news," in *in Proc. Int. Conf. on Language Resources and Evaluation*, 2004.

[14] A. M. Sandoval and L. C. Llanos, "MAVIR: a corpus of spontaneous formal speech in Spanish and English," in *Iberspeech 2012: VII Jornadas en Tecnología del Habla and III SLTech Workshop*, 2012.

[15] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at Sixteen: Update and outlook," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, December 2011.

[16] J. Parapar, A. Freire, and A. Barreiro, "Revisiting n-gram based models for retrieval in degraded large collections," in *Proceedings of the 31st European Conference on Information Retrieval Research: Advances in Information Retrieval*, ser. Lecture Notes in Computer Science, vol. 5478. Springer International Publishing, 2009, pp. 680–684.

[17] E. Rodríguez-Banga, C. Garcia-Mateo, F. Méndez-Pazó, M. González-González, and C. Magariños, "Cotovía: an open source TTS for Galician and Spanish," in *Proceedings of Iberspeech 2012*, 2012, pp. 308–315.

[18] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[19] A. Abad, L. J. Rodríguez-Fuentes, M. Peñagarikano, A. Varona, and G. Bordel, "On the calibration and fusion of heterogeneous spoken term detection systems." in *Proceedings of Interspeech*, 2013, pp. 20–24.

[20] N. Brümmer and D. van Leeuwen, "On calibration of language recognition scores," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 1–8.

[21] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," Tech. Rep., 2011. [Online]. Available: https://sites.google.com/site/nikobrummer

[22] 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.

[23] A. Varona, M. Penagarikano, L. Rodriguez-Fuentes, and G. Bordel, "On the use of lattices of time-synchronous cross-decoder phone co-occurrences in a SVM-phonotactic language recognition system." in *12th Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 2901–2904.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM Multimedia (MM)*, 2010, pp. 1459–1462.

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

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

[27] A. Abad, R. Astudillo, and I. Trancoso, "The L2F spoken web search system for Mediaeval 2013," in *Proceedings of the MediaEval 2013 Workshop*, 2013.

[28] 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 the 37th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7899–7903.