COPING WITH CHANNEL MISMATCH IN QUERY-BY-EXAMPLE - BUT QUESST 2014

Igor Szöke, Miroslav Skácel, Lukáš Burget, and Jan "Honza" Černocký*

BUT Speech@FIT, Brno University of Technology, Czech Republic
{szoke}@fit.vutbr.cz

ABSTRACT

The paper investigates into Query by Example (QbE) – a spoken term detection technique with queries entered by voice. It describes BUT QbE system that achieved the best accuracy in MediaEval QUESST2014 evaluations. This evaluation was challenging because of severe mismatch between queries and utterances, and introduction of new types of queries. The paper provides an analysis of DTW sub-system's in mismatched conditions (especially targeting DTW metrics) and discusses approaches investigated for QUESST2014: generation of calibration side-information by a language identification system, and handling T2 and T3 queries relaxing the constraints of an exact match. All results are provided on QUESST2014 development and evaluation data.

Index Terms— query-by-example spoken term detection, spoken document retrieval, acoustic keyword spotting, dynamic time warping, fusion, m-norm, logistic regression

1. MOTIVATION

As shown in a recent summary paper [1], the QbE approaches can be roughly divided into two categories: the pattern-matching ones look for similarities at the feature level and are mostly represented by a Dynamic Time Warping (DTW)-style comparison of query and utterance segments. The second category is represented by symbolic systems (based on Weighted Finite State Transducers) where the queries and utterances are represented by a sequence or graph of discrete symbols (phones for example). Our Acoustic Keyword Spotting (AKWS) builds a model of query and processes the utterances by looking at log-likelihood ratio of the keyword model and a background model. The AKWS can be considered in between the DTW and symbolic systems.

The QUESST2014 dataset was challenging because of mixed language, acoustic conditions, and 3 types of queries (exact match, variation, and reordering). We wanted to thoroughly compare and combine the DTW and the AKWS approaches. In our previous research aiming at SWS2012 evaluations [2], we found bottleneck features superior to standard posteriors, so our goal was to compare posterior and bottleneck features in mismatched-channel scenario and also to compare different distance metrics for DTW.

In comparison to our last year system [3], we used lower number of systems in parallel and used bottleneck features. Our goal was to further investigate the sensitivity of particular approaches to the language / channel mismatch in the query and utterance data. Also, coping with different types of queries was challenging this year [4]. Similarly to SWS2013, we used systems already available at BUT (socalled Atomic Systems).

* Igor Szöke was supported by Grant Agency of Czech Republic postdoctoral project No. GP202/12/P567.

	min./seg.	dev / eval	type
Albanian	127/968	50(20/13/16) / 50(18/13/17)	read
Basque	192/1841	70(16/33/21) / 70(30/19/21)	broadcast
Czech	237/2652	100(77/24/27) / 100(73/27/32)	conversational
NNEnglish	273/2438	138(46/46/46) / 138(46/46/46)	TEDx
Romanian	244/2272	100(46/21/31) / 100(43/27/30)	read
Slovak	312/2320	102(102/53/14) / 97(97/47/10)	parliamentary
SUM dev	1385/12491	560(307/190/155)	mixed
SUM eval	1385/12491	555(307/179/156)	mixed

Table 1. *Set of 6 European languages. The first column: amounts of data per language. The second column: the numbers of development (dev) and evaluation (eval) queries (all(T1/T2/T3)). The last column is type of speech.*

2. DATA AND SCORING

The QUESST2014 organizers brought the evaluations closer to a real scenario – voice search over a set of audios. The database consists of only one set of utterances – used both for development and evaluation – and two sets of queries: one for development and the other for evaluation. The overall length of utterance data is 23 hours, in 6 languages (table 1).

Utterances in the search repository were shuffled and no side information was provided to participants regarding the spoken language, acoustic conditions, or query type. Therefore, any adaptation needs to rely on unsupervised algorithms. QUESST2014 brings two interesting research areas:

- Cross-channel problem in query – utterance. All of the queries were dictated and recorded by mobile telephone. The utterances were from different sources and speaking style (read, conversational, lecture, broadcast).
- 3 different types of queries: Type 1 (T1) queries seek the exact match (the same way as in SWS2013, SWS2012). If the query is `white horse`, an utterances such as `...my white horse is nice...` should be found. Type 2 (T2) queries are queries with variations (for example inflections). Utterances such as `...my whiten horses are nice...` are expected. Type 3 (T3) queries are queries with reordering, for example `...my horses are nice and white...`. Each word in query has at least 4 phonemes, no other information was provided. As the queries were randomized, each participant should find a way how to detect T1/T2/T3 and search them. One query could appear in all forms (T1/T2/T3) in the data.

The evaluation goal was changed from keyword spotting task (Where is the occurrence of Query005?) to detection task (Is Query005 in Utterance0156?). Normalized cross-entropy cost (C_{nxe}) was chosen as primary metric by the organizers [4]. Well known Term Weighted Value (TWV) defined by NIST [5] was the secondary metric. Actual TWV (ATWV) score is calculated ac-

cording to a hard YES/NO decision for each detection given by a system. Maximum TWV (MTWV) is then calculated by searching for a global threshold (to set YES/NO decision) with respect to maximization of ATWV. $ATWV = 1$ means 100% accurate system (no false alarms (FA) and no misses), and lower $ATWV$ represents worse system. System with no output has $ATWV = 0$. ATWV can be negative (no hits, lots of FAs).

On the other hand, C_{nxe} is based on system scores. It measures the fraction of information, with regard to the ground truth, that is not provided by system scores, assuming that they can be interpreted as log-likelihood ratios (LLR). A perfect system would get $C_{nxe} = 0$ and a non-informative system would get $C_{nxe} = 1$, whereas $C_{nxe} > 1$ would indicate severe miscalibration of the log-likelihood ratio scores. More details on both evaluation metrics used for QUESST2014 can be found in [1].

As C_{nxe} was the primary metric, we did not optimize on TWV. We just found the best global threshold (maximizing the TWV by hard YES/NO decision). The QUESST2014 dataset is available here[1].

3. SYSTEM OVERVIEW

We followed our system architecture from SWS2013 [3]. Only brief system description is provided with emphasis on differences and new things in this paper. The reader is kindly asked to read also our previous paper [6].

Our **Query-by-Example** (QbE) system (figure 1) is based on phoneme-state posterior extractors and bottleneck features extractors. Each extractor (denoted as an **Atomic system**) is an artificial neural network taking raw audio file as the input (either query example or test utterance) producing phoneme state posteriors (POST) or bottleneck features (BN) as the output. We used 7 Atomic Systems – $3\times$ phoneme state posteriors, $4\times$ bottleneck features. See section 3.1 for details.

Phoneme state posteriors were then processed by a **Query-by-Example Subsystems**. We have two types of subsystems, one based on the AKWS (section 3.2) and the other based on the DTW (section 3.3). The input of each subsystem is the matrix of phoneme state posteriors or bottleneck features for query example and utterance. The output is a set of detections of given query example in the utterance.

The next step is a **score normalization and calibration**. It takes the set of detections and normalizes the scores. Next, only one score per query-utterance pair is generated. Finally, the scores are calibrated with respect to the normalized cross entropy – C_{nxe} (section 4).

Fusion is the final stage of the QbE system. It takes calibrated outputs of all subsystems and fuses them into one output. Again, we optimize the fusion parameters with respect to the normalized cross entropy – C_{nxe} (section 5).

3.1. Atomic Systems

All our Atomic systems use Artificial Neural Network classifiers (ANN) to estimate per-frame phoneme state posterior probabilities (posteriorgrams) or bottle-neck features (outputs of hidden layer). Our motivation was to re-use already trained ANNs available at Brno University of Technology (BUT).

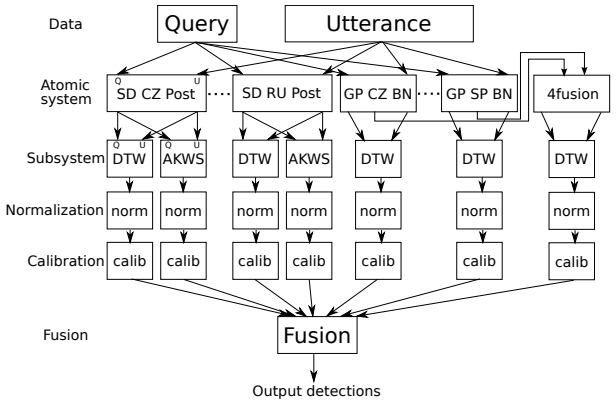

Fig. 1. *Query-by-Example system schema: Q means Queries as the input, U means Utterances as the input, SD means SpeechDat atomic systems where the output are phoneme posteriors. GP means GlobalPhone atomic systems where the output are bottleneck features.*

The ANNs were trained as acoustic models for phoneme recognizers in several past BUT projects. Altogether, we ended-up with 7 Atomic systems with the following architectures and trained on the following datasets:

- $3\times$ **SpeechDat**[2] (Czech, Hungarian and Russian; monolingual LCRC systems [7], trained on 20 hours of read speech per language) for estimations of phoneme state posteriors. Denoted as *SD Post*. The Hungarian system was also used as Speech Activity Detector (SAD).
- $4\times$ **GlobalPhone** (Czech, Portuguese, Russian, Spanish; monolingual stacked-bottleneck systems [8, 9] for BN features, trained on 20 hours of read speech per language). Denoted as *GP BN*.

3.2. Acoustic Keyword Spotting based QbE

In AKWS [6], we build an HMM for each query and then calculate log likelihood ratio between the query model and a background model (free phone loop). In QbE task, however, we need to generate the phoneme sequence for each of the acoustic examples – a **query-to-text step**. This is achieved by decoding each example using free phoneme loop. We remove all silence labels (if present) in decoded queries. The AKWS works on top of phoneme posteriors.

3.3. Dynamic Time Warping based QbE

In our implementation, we follow the standard QbE recipe – subsequence DTW [10]. A single DTW is run for each combination of query and utterance and the query is allowed to start at any frame of the utterance. When selecting the locally optimal path in the standard DTW algorithm, transition from the smallest accumulated distance is chosen. In our implementation, we compare the accumulated distances (including the current local distance) normalized by the corresponding path lengths on-the-fly. Note that in the standard subsequence DTW, no on-the-fly path length normalization in performed, which results in the inappropriate preference for shorter (recently started) paths. We applied SAD and removed all silence

[1] http://speech.fit.vutbr.cz/software/quesst-2014-multilingual-database-query-by-example-keyword-spotting

[2] http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

Features	corr	cos	euc	logcos	logdot
SD CZ POST	0.687(0.534)	0.768(0.633)	0.852(0.806)	**0.649**(0.453)	0.658(0.460)
SD HU POST	**0.646**(0.505)	0.712(0.584)	0.805(0.731)	0.679(0.510)	0.691(0.523)
SD RU POST	0.653(0.509)	0.706(0.557)	0.789(0.712)	**0.652**(0.495)	0.662(0.510)
GP CZ BN	0.593(0.435)	**0.585**(0.425)	0.777(0.672)	0.722(0.601)	-
GP PO BN	0.659(0.536)	**0.650**(0.522)	0.882(0.830)	0.819(0.750)	-
GP RU BN	0.668(0.533)	**0.658**(0.516)	0.862(0.800)	0.814(0.726)	-
GP SP BN	0.673(0.558)	**0.663**(0.540)	0.849(0.773)	0.822(0.741)	-
GP CZ+PO+RU+SP BN 4fusion	0.586(0.432)	**0.579**(0.418)	0.761(0.671)	0.713(0.601)	-

Table 2. *Comparison of distance metrics scored by $minC_{nxe}$. Number for T1 queries only are shown in parenthesis. SD POST denotes 3-state posterior features from some of SpeechDat-East languages, GP BN denotes bottle-neck features from some of GlobalPhone languages. The last line is a fusion.*

System	sideinfo	eval	dev
LID	-	0.926(1.3)	0.929(2.0)
best_single	-	0.556(42.4)	0.586(42.7)
best_single	QU	0.535(41.7)	0.539(40.5)
best_single	LID	0.551(**43.4**)	0.581(**43.6**)
best_single	QU+LID	**0.528**(41.9)	**0.533**(41.2)
bigfusion	-	**0.464**(**49.7**)	**0.486**(**49.8**)
bigfusion	QU	0.465(47.3)	0.461(46.4)
bigfusion	LID	0.473(49.0)	0.496(49.5)
bigfusion	QU+LID	0.470(46.7)	0.466(46.2)

Table 3. *Comparison of different calibration side-info. The scores are $minC_{nxe}$ with MTWV in parenthesis. The best_single is single QbE subsystem based on DTW on GP Czech bottlenecks. The bigfusion system is fusion of 11 subsystems. LID is a subsystem based only on LID sideinfo used as scores. Notice the differences between dev and eval accuracy for single subsystem and system fusion.*

frames from queries. We experimented with different distance metrics and input features in DTW.

We used different metrics for measuring distances between query-utterance vectors: *Pearson product moment correlation* distance (corr), *cosine* distance (cos), *Euclidean* distance (euc), *logarithm of the cosine* distance (logcos) and *logarithm of the dot product* (logdot) – see table 2. The Euclidean distance gave us the worst results for both posteriors and BNs. For posteriors, the logarithm of the cosine provided us the best results and we also obtained acceptable accuracy using the logarithm of the dot product. For BNs, the best accuracy was clearly achieved using the cosine distance. However, we used the Pearson correlation as the main metric for QUESST2014, as it seems to be the most universal distance metric regardless the features. Note, that accuracy on T1 (in parenthesis) shows the same trends as the overall score.

3.4. Score normalization

For both DTW and AKWS subsystems, the local maxima of frame-by-frame accumulated detection scores are selected as candidate detections. For overlapping detections, only the best scoring ones are preserved. For the AKWS, the accumulated detection scores are normalized by the length of the detection, for the DTW, by the length of warping path (done on-the-fly). After the length normalizations, we apply an *m-norm* which was found the best last year [3]. Finally, we take the highest score of particular query in particular utterance and attach this score to the query–utterance pair.

4. SCORE CALIBRATION

Next, we calibrate the scores using binary logistic regression, where the input to the logistic regression is a vector of normalized scores augmented with different per-term, per-query side-information scores, denoted as *sideinfo*: number of phonemes, log of number of phonemes, number of speech frames, log of number of speech frames, average log-posterior of speech frames taken from SAD and optionally the language indentification (LID) i-vector score. The side-info is generated for queries and utterances so the final "feature vector" for calibration consists of: 1 detection score (query–utterance pair), 5 query side-info, 5 utterance side-info. The logistic regression parameters are trained on development set. We denote this query+utterance side-info parameters as *QU*.

Another set of experiments was performed with a language identification (LID) system. The motivation to use LID was simple: the languages of query and utterance should be matching, for example, a Czech query should not be searched in Basque utterances, etc. We used a state-of-the-art system based on i-vectors [11]. As acoustic features, we used Shifted Delta Cepstra. Gaussian mixture model with 2048 Gaussians serves as Universal Background Model for 600 dimensional, gender-independent, i-vector extractor. LID produces a distance (Pearson product moment correlation) between query and utterance i-vectors. This distance was used as a side-info (denoted as *LID*), and we also scored LID as separate subsystem as the LID provides unique score for each query-utterance pair.

We summarized our results in table 3. The conclusion is that side-info is helpful for a single subsystem. The QU bring 5% absolute improvement on dev and 2% absolute improvement on eval data. This shows over-fitting of QU side-info. The LID brings only tiny improvement of 0.5% on dev and eval data without over-fitting. The LID can be considered also as another subsystem because it calculates distance of query and utterance i-vectors. The accuracy of LID subsystem 0.926 shows low performance which is expected, because the LID is not suitable for keyword search. On the other hand, per language scoring (see table 5) shows that LID extracts some query-utterance information as the per language scores are lower than 1 for $minC_{nxe}$.

When fusion of 11 calibrated subsystems was done, the side-info started to degrade system accuracy on evaluation data. Probably due to the over-fitting.

5. FUSION

We used two fusions: the first one on the level of feature vectors going to DTW (inspired by Fuentes et al. [12]). Here, we concatenated feature vectors of 4 GP BN atomic systems and then processed them by our DTW subsystem (denoted as 4fusion). Fuentes reported significant gain from this fusion (10% on MTWV) on the SWS2013 data. However, we got only 2% on MTWV and 0.7% on $minC_{nxe}$ improvement compared to single best subsystem (DTW-GP-Czech BN), see lines 1 and 4 in table 5. Anyway, we let the 4fusion subsystem in for "bigfusion".

The second type of fusion was on level on subsystem outputs (system combination). All calibrated scores from the individual subsystems were fused using binary logistic regression linear classifier. 11 subsystems took part in the bigfusion system – 3 AKWS based

Approach	sideinfo	eval $minC_{nxe}$	MTWV	dev $minC_{nxe}$	MTWV
bigfusion - primary	QU	$\mathbf{0.465}(0.310/0.461/0.673)$	47.3	$0.461(0.309/0.513/0.624)$	46.4
bigfusionnoside		$0.464(0.323/0.470/0.660)$	49.7	$0.486(0.333/0.554/0.624)$	49.8
best_single	QU LID	$\mathbf{0.528}(0.374/0.546/0.714)$	41.9	$0.533(0.376/0.600/0.675)$	41.2
LID		$0.926(0.897/0.946/0.920)$	1.3	$0.929(0.896/0.961/0.901)$	2.0
AKWS-cz	QU	$\mathbf{0.648}(0.519/0.645/\mathbf{0.848})$	25.6	$\mathbf{0.641}(0.500/0.680/\mathbf{0.824})$	25.0
AKWS-T3-cz	QU	$\mathbf{0.674}(0.597/0.694/\mathbf{0.756})$	19.4	$\mathbf{0.673}(0.581/0.742/\mathbf{0.718})$	19.2
bigfusionnoside - bestmetric		$0.455(0.310/0.462/0.653)$	50.0	$0.479(0.321/0.549/0.611)$	50.6

Table 4. *Results for the approaches in minimum C_{nxe} and minimum TWV with per query type (T1/T2/T3). The bigfusion system was our primary system submitted to the evals. AKWS-cz and AKWS-T3-cz is experiment on T3 queries. Notice accuracy improvement on T3 but loss on overall score.*

Approach	sideinfo	ALL	Albanian	Basque	Czech	NNEnglish	Romanian	Slovak
DTW-GP-Czech BN	-	$\mathbf{0.586}(42.7)$	$\mathbf{0.519}(43.8)$	$\mathbf{0.773}(32.3)$	$\mathbf{0.591}(42.8)$	$\mathbf{0.853}(11.2)$	$\mathbf{0.368}(59.4)$	$0.444(57.5)$
DTW-SP-Czech Post	-	$0.677(33.8)$	$0.551(40.7)$	$0.843(24.0)$	$0.693(29.0)$	$0.881(8.9)$	$0.424(47.2)$	$0.564(49.3)$
AKWS-SP-Czech Post	-	$0.665(25.7)$	$0.799(12.9)$	$0.865(13.6)$	$0.628(35.2)$	$0.927(2.9)$	$0.597(29.9)$	$\mathbf{0.394}(59.4)$
DTW-GP-4fusion BN	-	$0.579(44.7)$	$0.522(44.7)$	$0.773(36.1)$	$0.590(41.4)$	$0.831(13.6)$	$0.329(65.3)$	$0.453(55.8)$
bigfusion	-	$0.486(49.8)$	$0.497(46.4)$	$0.715(38.7)$	$0.505(50.1)$	$0.811(16.2)$	$0.279(65.2)$	$0.320(67.4)$
bigfusion	QU	$0.461(46.4)$	$0.525(41.8)$	$0.688(36.5)$	$0.474(49.1)$	$0.796(16.3)$	$0.295(56.6)$	$0.310(65.5)$
LID		$0.929(2.0)$	$0.903(2.1)$	$0.984(0.4)$	$0.918(4.9)$	$0.973(0.2)$	$0.930(0.7)$	$0.930(1.3)$

Table 5. *Dev data results (and per language results) in $minC_{nxe}$ and minimum TWV in parenthesis. Notice: 1) Difference between best single subsystem (line 1) and 4fusion on feature level (line 4). 2) Difference between DTW (line 2) and AKWS (line 3) approaches on SP-Czech Post 3) Difference between best single subsystem (line 1) and bigfusion (line 5 and 6).*

on SD Post, 3 DTW based on SD Post, 4 DTW based on GP BN, 1 DTW based on 4fusion. We summarized results of interesting subsystems and the bigfusion in table 5. The bigfusion improved the score of best single subsystem by 10% $minC_{nxe}$.

6. RESULTS, T2 AND T3 QUERIES

Table 4 summarizes accuracies of systems we submitted to the QUESST2014 evaluations. In this scetion we aim at query type analysis (the impact of score calibration was discussed in section 4 and fusion was discussed in section 5). As you can notice, the T1 queries – exact match – achieved the best accuracy (0.310 $minC_{nxe}$ for bigfusion system). The T2 queries – variations – achieved significantly lower accuracy (0.461 $minC_{nxe}$) and the T3 – reordering – achieved the worst one (0.673 $minC_{nxe}$). This is given by 1) no optimization of our systems on T2 and T3 and 2) difficulty to identify T2/T3 in randomized data and search them correctly.

Although no T2/T3 handling was included in our bigfusion submission system, we experimented with improving accuracy on these queries. The AKWS approach was modified to allow the last phoneme in the query to be any phoneme. We found a tiny improvement of 0.4% on T2, but overall 1% $minC_{nxe}$ deterioration. The other experiments (for example, use "wild-card" for the first, second, and the last two phonemes) led to loss of accuracy even on T2.

The best approach to cope with T3 we found was to split queries longer than 7 phonemes in the middle. Then, we searched for these two particular sub-queries independently. Finally, we merged the sub-query results by forbidding sub-queries overlap longer than 10 frames. Results of this experiment are in table 4 lines 5 and 6. System AKWS-cz is reference system where we search for T3 in the same way as for T1 (exact match). We implemented the above mentioned split to sub-queries in system AKWS-T3-cz. We got improvement 9% on T3 but the overall deterioration is 2.4% of C_{nxe} on eval queries.

In all T2/T3 experiments, we obtained a small improvement on T2/T3 followed by large loss of accuracy on T1. It can be explained by about twice more T1 occurrences than T2/T3 (see table 1). While implementing "softness" to cover T2/T3, we get hit on T1 where any softness in unwanted. Our conclusion here is, that it does not make sense to cover T2 queries by a special approach (search algorithm), as these queries are covered enough by "softness" of the standard DTW algorithm.

In table 5, we analyzed also per language accuracy. We conclude that good accuracy is achieved in scenarios, where the query and utterance channel and type of speech are matching. What is surprising is the very low accuracy on non-native English. This is probably caused by the channel rather than accent mismatch (as the accent of TEDx is not so strong). Next, we observed superiority of AKWS over DTW (based on SP Czech Post) on Slovak data. This can be explained by language match (Slovak is very close to Czech) and also type of speech match (SD Czech achieved higher accuracy than GP Czech).

7. CONCLUSIONS

We presented our QUESST2014 bigfusion system, which achieved the best accuracy on C_{nxe} metric. Moreover, our single best system based on Czech GlobalPhone bottleneck features, DTW and sideinfo calibration scored the second with difference (0.7%). We conclude the superiority of bottleneck features in DTW to posteriors in cross-channel and multilingual environments. We found Pearson product moment correlation distance a good universal metric, while one can achieve a small improvement by using log-cosine distance for posteriors and cosine distance for bottle-necks. Our conclusion on different query types is, that there is no advantage of coping with query variations (T2) as these can lead to significant accuracy loss on exact match (T1). On the other hand, there is interesting room to improve search algorithms for reordering queries (T3) while avoiding accuracy loss on exact match (T1).

8. REFERENCES

[1] Xavier Anguera, J. Luis Rodriguez-Fuentes, Igor Szöke, Andi Buzo, and Florian Metze, "Query-by-example spoken term detection evaluation on low-resource languages," in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under- resourced Languages SLTU-2014. St. Petersburg, Russia*, 2014, pp. 24–31.

[2] Igor Szöke, Michal Fapšo, and Karel Veselý, "BUT2012 approaches for spoken web search - Mediaeval 2012," in *MediaEval 2012 Workshop*, Pisa, Italy, October 4-5 2012, CEUR Workshop Proceedings, Vol. 2012, No. 927, DE.

[3] Igor Szöke, Lukáš Burget, František Grézl, and Lucas Ondel, "Calibration and fusion of query-by-example systems - BUT SWS 2013," in *Proceedings of ICASSP 2014*, 2014, pp. 7899–7903.

[4] X. Anguera, L.J. Rodriguez-Fuentes, Igor Szöke, Andy Buzo, and Florian 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.

[5] J. Fiscus, J. Ajot, and G. Doddington, "The spoken term detection (STD) 2006 evaluation plan," Tech. Rep., National Institute of Standards and Technology (NIST) USA, September 2006.

[6] Igor Szöke, Petr Schwarz, Lukáš Burget, Martin Karafiát, Pavel Matějka, and Jan Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," *Lecture Notes in Computer Science*, vol. 2005, no. 3658, pp. 8, 2005.

[7] Petr Schwarz, Pavel Matějka, and Jan Černocký, "Towards lower error rates in phoneme recognition," in *Proceedings of 7th International Conference Text,Speech and Dialoque 2004*. 2004, p. 8, Springer Verlag.

[8] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proceedings of IEEE 2012 Workshop on Spoken Language Technology*. 2012, pp. 336–341, IEEE Signal Processing Society.

[9] František Grézl and Martin Karafiát, "Hierarchical neural net architectures for feature extraction in ASR," in *Proceedings of INTERSPEECH 2010*, 2010, vol. 2010, pp. 1201–1204.

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

[11] Niko Brummer, Sandro Cumani, Ondřej Glembek, Martin Karafiát, Pavel Matějka, Jan Pešán, Oldřich Plchot, Mohammad Mehdi Soufifar, Edward Villiers de, and Jan Černocký, "Description and analysis of the Brno276 system for LRE2011," in *Proceedings of Odyssey 2012: The Speaker and Language Recognition Workshop*. 2012, pp. 216–223, International Speech Communication Association.

[12] Luis Javier Rodríguez-Fuentes, Amparo Varona, Mikel Peñagarikano, Germán Bordel, and Mireia Díez, "High-performance query-by-example spoken term detection on the SWS 2013 evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, 2014, pp. 7819–7823.