PHONETIC UNIT SELECTION FOR CROSS-LINGUAL QUERY-BY-EXAMPLE SPOKEN TERM DETECTION

Paula Lopez-Otero, Laura Docio-Fernandez, Carmen Garcia-Mateo

AtlantTIC Research Center, E.E. Telecomunicación, Campus Universitario S/N, 36310, Vigo

ABSTRACT

Cross-lingual query-by-example spoken term detection (QbE STD) has caught the attention of speech researchers, as it makes it possible to develop systems for low-resource languages, in which the available amount of labelled data makes the training of automatic speech recognition approaches prohibitive. The use of phonetic posteriorgrams for speech representation combined with dynamic time warping search is a widely used approach for this task, but little attention has been focused in the suitability of a set of phonetic units to represent speech information spoken in a different language. This paper proposes a technique for estimating the relevance of phonetic units aiming at selecting the most suitable ones for a given target language. Experiments in a Spanish database using phoneme posteriorgrams in Czech, English, Hungarian and Russian proved the validity of the proposed method, as QbE STD performance was enhanced by reducing the set of phonetic units.

Index Terms— Query-by-example spoken term detection, phoneme posteriorgram, subsequence dynamic time warping

1. INTRODUCTION

The amount of available spoken documents is steadily increasing, which leads to the need for tools to perform automatic search within large audio databases. Generating transcriptions as a previous step to this search in audio contents is not an option nowadays, as the cost of manual transcriptions is prohibitive and automatic speech recognition systems are not yet robust enough [1]. This led to a new perspective for tackling this issue, which consists in searching for audio content, within audio content, using an audio content query [2]. This definition encompasses two different approaches: searching for the time instant in which a given query was spoken, namely query-by-example spoken term detection (QbE STD), or searching for the documents in which a given query was pronounced, namely query-by-example spoken document retrieval (QbE SDR). The growing interest around these tasks led to the organization of different international competitions in order to encourage research in this field [2][3][4][5][6][7].

A common approach for QbE STD and QbE SDR relies on generating a word-level or phoneme-level transcription of the spoken queries and the documents by means of large vocabulary continuous speech recognition (LVCSR) approaches, making it possible to use tools such as lattice search [8][9]. The main limitation of these techniques is that, in low-resource languages, the amount of transcribed audio that is required for training a good ASR system is not available. Hence, the research community is focusing on zero- or low-resource approaches to overcome the search on speech problem; these techniques are mostly based in template matching techniques [6], but they differ in the type of features they use to represent the documents and the queries. Zero-resource approaches usually represent the speech information by means of acoustic features extracted from the waveform, such as Mel-frequency cepstral coefficients [10][11][12], and low-resource approaches usually rely on a phonetic posteriorgram representation. The latter consists in a time vs. class matrix representing the posterior probability of each phonetic class for each specific time [13], which are obtained using phone decoders that are not necessarily developed in the target language. These cross-lingual strategies, when combined with dynamic time warping (DTW) search, have led to encouraging results [14][15][16][17].

One issue that is not usually considered when performing cross-lingual QbE STD is that not all the languages share the same phonetic units, so it is possible that some phonemes are suitable for this task in a given language, but they might act as nuisance for other languages. The most logical manner to approach this concern would be using feature selection techniques but, unfortunately, the most common feature selection techniques used in pattern recognition tasks cannot be straightforwardly applied in this situation, as this task does not involve a division of the data to be processed into classes in order to be able to compute the relevance of the different features. This paper presents a technique for phonetic unit selection in the context of QbE STD, which consists in decomposing the contribution of each phonetic unit to the cost of the best alignment path in DTW matching in order to measure the relevance of these units. This cross-lingual approach was validated in the framework of Albayzin 2014 search on speech evaluation [7], in which QbE STD was performed on a database in Spanish using phone decoders for English, Czech,

978-1-4799-7291-3/15/$31.00 ©2015 IEEE

Hungarian and Russian [18].

The rest of this paper is organized as follows: Section 2 describes the QbE STD approach used in this work; Section 3 presents the proposed technique for estimating the phoneme relevance of the different phonetic units; Section 4 defines the experimental framework used to validate the proposed strategy; Section 5 specifies the experimental settings; Section 6 shows the experimental results; and Section 7 summarizes the conclusions and defines some future work.

2. QBE STD APPROACH

The approach for QbE STD used in this work has two main steps: the first one encompasses the extraction of relevant features from the waveform, specifically phoneme posteriorgrams were used; and the second one encompasses the search algorithm, which in this case consists in a variant of the classic DTW approach.

2.1. Phonetic posteriorgram representation

The representation of spoken queries and documents can be done by means of phonetic posteriorgrams [13]; given a spoken document and a phone decoder with U phonetic units, the posterior probability of each phonetic unit is computed for each time frame, leading to a set of vectors of dimension U that represents the phonetic probability of each phonetic unit at each time frame.

2.2. Search algorithm: subsequence DTW

The search of the spoken queries within the audio documents is commonly performed using the DTW algorithm [19]: given a document $D = \{d_1, \ldots, d_m\}$ and a query $Q = \{q_1, \ldots, q_n\}$ of m and n frames respectively, with $n \ll m$, DTW finds the best alignment path between these two sequences. In search on speech applications, it is common to use modifications of this algorithm such as subsequence DTW (S-DTW) [20], non-segmental DTW [11] and memory efficient approaches [21] [22]. In this work, the S-DTW approach was chosen.

To perform S-DTW, 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 and the document, respectively:

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

where $c(q_i, d_j)$ is a function that defines the cost between query vector q_i and document vector d_j, and

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

Different metrics can be used to compute the cost function, such as the Euclidean distance or the cosine distance; in this system, Pearson's correlation coefficient r is used [16], as it showed a superior performance when compared with other metrics:

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

where $q_i \cdot d_j$ denotes the dot product of q_i and d_j, and U is the number of phonetic units as defined above.

High values of correlation correspond to a low cost and vice versa; hence, in order to use this correlation as a cost function defined in [0,1], the following transformation is performed:

$$c(q_i, d_j) = \frac{1 - r(q_i, d_j)}{2} \tag{4}$$

Once the matrix M is computed, the end of the best match between Q and D is the one whose last frame is

$$b^* = \underset{b \in 1, \ldots, m}{\arg\min}\, M(n, b) \tag{5}$$

The starting point of the matching path, a^*, is computed by backtracking, which serves to obtain the best path $P(Q, D) = \{p_1, \ldots, p_k, \ldots, p_K\}$, where $p_k = (i_k, j_k)$, i.e. the k^{th} element of the path is formed by q_{i_k} and d_{j_k}. Once the start and end frames of the match are obtained, the score for this match can be computed as $\frac{M(n,b^*)}{b^* - a^* + n}$ as in [15].

It is possible that a query Q appears several times in a document D, specially if D is a long recording. Thus, not only the best match must be detected but also less likely matches. One approach to overcome this issue consists in detecting a given number of candidates n_c: every time a candidate match is detected, which ends at frame b^*, $M(n, b^*)$ is set to ∞ in order to ignore this match; in this way, a maximum of n_c candidates per query and document with the lowest cost are output by the system (this value is usually much lower than n_c because overlapping matches or very short matches are discarded).

3. PHONEME RELEVANCE ESTIMATION

The main contribution of this paper consists in a technique to select the most relevant phonemes in a QbE STD context, which is based on the best alignment path obtained when performing S-DTW. We hypothesize that the phonetic units which most contribute to increasing the cost in the best path are assumed to be nuisance, while the phonetic units which obtain small costs are considered the most relevant. Hence, given a pair query-document, we aim at obtaining a decomposition of $c(q_i, d_j)$ such that each phonetic unit u has an associated $c(q_i, d_j, u)$ that fulfils

$$\sum_{u=1}^{U} c(q_i, d_j, u) = c(q_i, d_j) \tag{6}$$

To this effect, Eq. (8) can be decomposed as:

$$r(q_i, d_j) = \frac{U q_{i,1} d_{j,1} + \ldots + U q_{i,U} d_{j,U} - \|q_i\| \|d_j\|}{\sqrt{(U\|q_i^2\| - \|q_i\|^2)(U\|d_j^2\| - \|d_j\|^2)}} \tag{7}$$

where $q_{i,u}$ represents the u^{th} element in q_i. Using Eq. (7), $r(q_i, d_j, u)$ can be defined as:

$$r(q_i, d_j, u) = \frac{U q_{i,u} d_{j,u} - \frac{1}{U} \|q_i\| \|d_j\|}{\sqrt{(U\|q_i^2\| - \|q_i\|^2)(U\|d_j^2\| - \|d_j\|^2)}} \tag{8}$$

which fulfils $\sum_{u=1}^{U} r(q_i, d_j, u) = r(q_i, d_j)$. Finally, we define the contribution of a phonetic unit u to the cost $c(q_i, d_j)$ as

$$c(q_i, d_j, u) = \frac{\frac{1}{U} - r(q_i, d_j, u)}{2} \tag{9}$$

which fulfils Eq. (6).

Given a pair query-document, its best path $P(Q, D)$ of length K is computed. Assuming that Q is present in D, we decompose the cost of each p_k in $P(Q, D)$ in order to obtain the contribution of each phonetic unit to the cost of this path as defined in Eq. (9). Finally, we define the relevance $R(P(Q, D), u)$ of a phonetic unit u in $P(Q, D)$ as:

$$R(P(Q, D), u) = \frac{1}{K} \sum_{k=1}^{K} c(q_{i_k}, d_{j_k}, u) \tag{10}$$

We assume that those phonetic units that less contribute to the cost in the best path $P(Q, D)$ are more relevant, as ideally the cost of the best path would be 0. Hence, the set of phonetic units can be sorted by its relevance $R(P(Q, D), u)$, and the least relevant ones can be removed from the phoneme posteriorgrams. In order to find a good estimate of the relevance, given a set of queries and their location in a given set of documents, the relevances of the different phonetic units obtained from the different pairs query-document are summed.

4. EVALUATION FRAMEWORK

The framework of Albayzin 2014 Search on Speech (SOS) evaluation [7] is used in this work to assess the proposed phonetic unit selection approach. The task to be performed in this evaluation consisted in searching for a set of acoustic examples within large recordings in Spanish language.

4.1. Database

MAVIR database [23] was used in Albayzin 2014 Search on Speech evaluation. It consists of a set of lectures in spontaneous formal speech taken place during the MAVIR workshops held in 2006, 2007 and 2008, which dealt with language technologies topics. Some of the recordings included more than one speaker, both male and female. The lectures were performed in large conference rooms with diverse types of microphones, leading to noise and mismatched acoustic conditions [6], and they were recorded using different audio formats; nevertheless, for the evaluation, all the recordings were converted to PCM, 16 kHz, single channel and 16 bits per sample.

Of all the recordings included in MAVIR database, 10 were selected for Albayzin 2014 SOS evaluation. These 10 recordings were split into a training and a test set in order to avoid biasing the results. A large set of spoken queries were manually extracted from these recordings, some of them being used for training and the rest of them for testing. Table 1 summarizes some specific information about the database used in these experiments.

Table 1. Summary of MAVIR database

Partition	# recordings	Total duration	# queries	# occurrences
Train	7	5 h 20 min	94	1415
Test	3	2 h 1 min	99	1162

4.2. Evaluation metric

The evaluation metric used in this paper is the average term weighted value (ATWV) in accordance with Albayzin 2014 SOS evaluation. Given a detection threshold Θ, ATWV is defined as [24]:

$$\mathrm{ATWV}(\Theta) = 1 - \underset{\mathrm{term}}{\mathrm{average}}\{\mathrm{P_{Miss}}(\mathrm{term}, \Theta) + \beta \cdot \mathrm{P_{FA}}(\mathrm{term}, \Theta)\} \tag{11}$$

where $\mathrm{P_{Miss}}(\mathrm{term}, \Theta)$ is the probability of missing hits of term given the decision threshold Θ, $\mathrm{P_{FA}}(\mathrm{term}, \Theta)$ is the probability of inserting false hits of term given Θ, and

$$\beta = \frac{C}{V}\left(PR_{term}^{-1} - 1\right) \tag{12}$$

where PR_{term} is the probability of term, which was fixed to 10^{-4} and $\frac{C}{V} = 0.1$ is a weighting factor that gives more or less relevance to either false alarms or miss detections.

A secondary metric was used in Albayzin 2014 SOS evaluation, namely the maximum term weighted value (MTWV), which is defined as the maximum possible ATWV, i.e. the ATWV obtained when selecting the optimal threshold.

5. EXPERIMENTAL SETTINGS

5.1. Feature extraction

In this paper, the phone decoders based on long temporal context [18] developed at the Brno University of Technology were used; specifically, the Czech (CZ), English (EN), Hungarian (HU) and Russian (RU) systems were used. In these decoders, each unit has three different states and a posterior probability is output for each of them, so they are combined in order to obtain one posterior probability for each unit [14]. After obtaining the posteriors, a Gaussian softening was applied in order to have Gaussian distributed probabilities [25].

It must be noted that the pre-trained models corresponding to Czech, Hungarian and Russian were trained using 8 kHz data; in order to overcome this sample frequency mismatch, MAVIR recordings were downsampled to 8 kHz in order to be able to use these phone decoders.

Table 2. Phone decoders used in the experiments.

Language	# phonemes	# fillers	# Phonetic units
CZ	42	3	45
EN	38	1	39
HU	58	3	61
RU	49	3	52

5.2. System tuning

The search strategy used in this paper for QbE STD has a tuning parameter, namely the decision threshold, which decides which scores correspond to hits (and therefore, must be output by the system) and which must be discarded by the system. The training set of Albayzin 2014 SOS evaluation was used to tune this threshold: given the output of a QbE STD system in the training set, the threshold that achieves the MTWV is chosen and subsequently applied in the test experiment. The number of candidate matches n_c is also a tuning parameter, which was empirically set to 100.

The proposed phoneme relevance approach sorts the phonetic units by relevance, but there is no criterion to decide where to stop considering that a phonetic unit is relevant. Thus, the most suitable number of phonetic units was also adjusted using the training documents and queries.

5.3. Contrastive system

The performance of the cross-lingual approach proposed in this paper is compared to a language-dependent approach, in order to find out whether using different languages to model the queries and the documents is having a big impact in system performance. Hence, a large vocabulary continuous speech recognition (LVCSR) system built for Albayzin 2014 SOS evaluation was used as a contrastive system [9]. This system, which was built using the Kaldi open-source toolkit [26], uses standard perceptual linear prediction (PLP) analysis to extract 13 dimensional acoustic features, and follows a state-of-the-art maximum likelihood (ML) acoustic training recipe, which begins with a flat-start initialization of context-independent phonetic Hidden Markov Models (HMMs), and ends with a speaker adaptive training (SAT) of state-clustered triphone HMMs with Gaussian mixture model (GMM) output densities. The ML stage is followed by the training of a Universal background model (UBM) from speaker-transformed training data, which is then used to train a subspace GMM (SGMM) that is used in the decoding stage.

The Kaldi LVCSR decoder generates word lattices [27] for the queries and the documents, using the above SGMM models. These lattices are converted into weighted finite state transducers (WFST) as described in [28], which are used to perform the search of the queries in the documents.

The data used to train the acoustic models of this Kaldi-based LVCSR system was around 78 hours of material extracted from the Spanish material used in the 2006 TC-STAR automatic speech recognition evaluation campaign[1] [29]. The language model was trained using a text database of 160 MWords composed of material from several sources (transcriptions of European and Spanish Parliaments from the TC-STAR database, subtitles, books, newspapers, online courses and the transcriptions of the MAVIR sessions included in the training set[2] [23]).

6. EXPERIMENTAL RESULTS

Table 3 presents the results obtained on the training set of Albayzin 2014 SOS evaluation. For each set of phoneme posteriorgrams, its ATWV when using all the phonetic units was used, and then the technique for unit selection proposed in Section 3 was applied in order to reduce the set of phonetic units into 10, 20, 30, 40 and 50 units, where possible. The Table shows that, for the four phone decoders used in these experiments, the phonetic relevance technique achieved an improvement in the QbESTD task with respect to using the whole set of phonemes. This improvement is specially noticeable in the case of Hungarian, where the ATWV highly increased when reducing the set of phonemes.

Table 4 shows the results obtained on the test experiment of Albayzin 2014 SOS evaluation, once the decision threshold and the number of phonetic units of each system were tuned as described in Section 5. These results reinforce the validity of the unit selection strategy as, as happened in the training set, results achieved when only keeping the most relevant units are better than those achieved when using the whole set of units. In addition, paired t-tests suggest that the achieved

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

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

Table 3. Results on the training experiment of Albayzin 2014 SOS evaluation: numbers in bold show the best result achieved by each phone decoder.

	All	10	20	30	40	50
CZ	0.118	0.092	0.138	0.142	**0.164**	-
EN	0.174	0.156	**0.196**	0.183	-	-
HU	0.062	0.086	0.157	0.164	**0.177**	0.173
RU	0.128	0.063	0.110	0.150	**0.157**	0.147

improvement is relevant specially in the case of the Hungarian. The best results were achieved when using the English models, but the improvement achieved by the unit selection approach is not very significant in this case.

Table 4. Results on the test experiment of Albayzin 2014 SOS evaluation: numbers in bold show the best result achieved by each phone decoder.

	All	Reduced
CZ	0.068	**0.148**
EN	0.202	**0.214**
HU	0.087	**0.142**
RU	0.107	**0.126**

6.1. Comparison with a language-dependent approach

Table 5 shows the results achieved by the language-dependent approach for QbE STD described in Section 4. Comparing these results with those shown in Tables 3 and 4, it can be observed that this approach is outperformed by the system using English phonetic posteriorgrams. The language-dependent system outperformed the results achieved when using the Czech, Hungarian and Russian phone decoders, but the difference in performance is very slight, suggesting that cross-lingual approaches are able to obtain competitive performance in QbE STD tasks.

Table 5. Results on Albayzin 2014 SOS evaluation when using the language-dependent contrastive system.

Partition	ATWV
Train	0.182
Test	0.151

7. CONCLUSIONS AND FUTURE WORK

This paper described an approach for selecting relevant phonetic units when dealing with cross-lingual query-by-example spoken term detection, based on the premise that the phonetic units in one language are not always relevant when representing spoken documents in a different target language. Given a DTW-based system that uses phonetic posteriorgrams to represent the queries and the documents, the proposed technique computes the relevance of each phonetic unit by obtaining its contribution to the cost of the best alignment path in the DTW algorithm. The experimental validation of this technique was performed in the framework of Albayzin 2014 search on speech evaluation: QbE STD was performed on a Spanish database using phonetic posteriorgrams in Czech, English, Hungarian and Russian. The results show that reducing the set of phonetic units enhanced the performance of the proposed system, specially when dealing with phone decoders with a large number of phonetic units, as it increases the probability of having phonetic units that are irrelevant in the target language. These results were also compared with a language-dependent approach based on large vocabulary continuous speech recognition and lattice search, showing that the performance gap between language-dependent and cross-lingual approaches is steadily decreasing.

In future work, new techniques for selecting the most relevant phonetic units in the context of QbE STD will be explored. In the approach proposed in this paper, the number of phonetic units was selected in a training dataset, but we plan to design an objective criterion to decide which of the phonetic units are relevant without having to empirically tune the number of selected phonetic units.

Having a criterion to decide whether a phonetic unit is relevant for QbE STD in a given target language opens new research possibilities as, instead of considering the different phone decoders individually, it is possible to think of a large pool of phonetic units, in which the most suitable units are used depending on the target language.

Lastly, we also plan to extend the proposed approach to feature selection for zero-resource QbE STD: having the possibility to perform feature selection in this task reduces the computational load of assessing the performance of different features that are not commonly used in QbE STD, making it possible to explore new options for zero-resource approaches.

8. ACKNOWLEDGEMENTS

This research was funded by the Spanish Government under project TEC2012-38939-C03-01, the Galician Government through the research contract GRC2014/024 and ‘AtlantTIC Project’ CN2012/160, and also by the Spanish Government and the European Regional Development Fund (ERDF) under project TACTICA.

9. REFERENCES

[1] C. Chelba, T.J., Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39–49, May 2008.

[2] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. Van Heerden, G. Mantena, A. Muscariello, K. Pradhallad, I. Szöke, and J. Tejedor, "The spoken web search task at MediaEval 2011," in *Proceedings of the 37th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[3] F. Metze, E. Barnard, M. Davel, C. Van Heerden, X. Anguera, G. Gravier, and N. Rajput, "The spoken web search task," in *Proceedings of the MediaEval 2012 Workshop*, 2012.

[4] X. Anguera, F. Metze, A. Buzo, I. Szöke, and L.J. Rodriguez-Fuentes, "The spoken web search task," in *Proceedings of the MediaEval 2013 Workshop*, 2013.

[5] X. Anguera, L.J. Rodriguez-Fuentes, I. Szöke, A. Buzo, and F. Metze, "Query by example search on speech at Mediaeval 2014," in *Proceedings of the MediaEval 2014 Workshop*, 2014.

[6] J. Tejedor, D.T. Toledano, X. Anguera, A. Varona, L.F. Hurtado, A. Miguel, and J. Colás, "Query-by-example spoken term detection ALBAYZIN 2012 evaluation: overview, systems, results, and discussion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 23, 2013.

[7] J. Tejedor, D.T. Toledano, L.J. Rodriguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel, "The ALBAYZIN 2014 search on speech evaluation plan," 2014.

[8] P. Yang, H. Xu, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S.J. Leow, B. Ma, E.S. Chng, and H. Li, "The NNI query-by-example system for MediaEval 2014," in *Proceedings of the MediaEval 2014 Workshop*, 2014.

[9] M. Martinez, P. Lopez-Otero, R. Varela, A. Cardenal-Lopez, L. Docio-Fernandez, and C. Garcia-Mateo, "GTM-UVigo systems for Albayzin 2014 search on speech evaluation," in *Iberspeech 2014: VIII Jornadas en Tecnologa del Habla and IV SLTech Workshop*, 2014.

[10] Y. Zhang and J.R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams.," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2009, pp. 398–403.

[11] G.V. Mantena, S. Achanta, and K. Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 5, pp. 944–953, 2014.

[12] X. Anguera, M. Skácel, V. Vorwerk, and J. Luque, "The Telefonica Research spoken web search system for MediaEval 2013," in *Proceedings of the MediaEval 2013 Workshop*, 2013.

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

[14] L.J. 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.

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

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

[17] L.J. Rodriguez-Fuentes, A. Varona, and M. Penagarikano, "GTTS-EHU systems for QUESST at MediaEval 2014," in *Proceedings of the MediaEval 2014 Workshop*, 2014.

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

[19] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, , no. 1, 1978.

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

[21] Xavier Anguera and Miquel Ferrarons, "Memory Efficient Subsequence DTW for Query-by-example Spoken Term Detection," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2013, pp. 1–6.

[22] Xavier Anguera, "Information retrieval-based dynamic time warping.," in *INTERSPEECH*, 2013, pp. 1–5.

[23] A. Moreno Sandoval and L. Campillos Llanos, “MAVIR: a corpus of spontaneous formal speech in Spanish and English,” in *Iberspeech 2012: VII Jornadas en Tecnologa del Habla and III SLTech Workshop*, 2012.

[24] National Institute of Standards and Technology (NIST), “The spoken term detection (STD) 2006 evaluation plan,” 2006.

[25] A. Varona, M. Penagarikano, L.J. 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 *INTERSPEECH*, 2011, pp. 2901–2904.

[26] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. 2011, IEEE Signal Processing Society.

[27] Daniel Povey, Mirko Hannemann, Gilles Boulianne, Luks Burget, Arnab Ghoshal, Milos Janda, Martin Karafit, Stefan Kombrink, Petr Motlcek, Yanmin Qian, Korbinian Riedhammer, Karel Vesel, and Ngoc Thang Vu, “Generating exact lattices in the wfst framework.,” in *ICASSP*. 2012, pp. 4213–4216, IEEE.

[28] Dogan Can and Murat Saraclar, “Lattice indexing for spoken term detection.,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 8, pp. 2338–2347, 2011.

[29] Antonio Cardenal-Lopez Laura Docio-Fernandez and Carmen Garcia-Mateo, “Tc-star 2006 automatic speech recognition evaluation: The uvigo system,” in *TC-STAR Workshop on Speech-to-Speech Translation*, 2006.