

Neural Network based End-to-End Query by Example Spoken Term Detection

Dhananjay Ram, Lesly Miculicich and Hervé Bourlard, *Fellow, IEEE*

Abstract—This paper focuses on the problem of query by example spoken term detection (QbE-STD) in zero-resource scenario. State-of-the-art approaches primarily rely on dynamic time warping (DTW) based template matching techniques using phone posterior or bottleneck features extracted from a deep neural network (DNN). We use both monolingual and multilingual bottleneck features, and show that multilingual features perform increasingly better with more training languages. Previously, it has been shown that the DTW based matching can be replaced with a CNN based matching while using posterior features. Here, we show that the CNN based matching outperforms DTW based matching using bottleneck features as well. In this case, the feature extraction and pattern matching stages of our QbE-STD system are optimized independently of each other. We propose to integrate these two stages in a fully neural network based end-to-end learning framework to enable joint optimization of those two stages simultaneously. The proposed approaches are evaluated on two challenging multilingual datasets: Spoken Web Search 2013 and Query by Example Search on Speech Task 2014, demonstrating in each case significant improvements.

Index Terms—Spoken term detection, query by example, deep neural network, bottleneck features, end-to-end, subsequence detection.

I. INTRODUCTION

Query-by-example spoken term detection (QbE-STD) [1] is defined as the task of detecting all files from an audio archive which contain a spoken query provided by a user (see Figure 1). It enables users to search through multilingual audio archives using their own speech. The primary difference from keyword spotting is that QbE-STD relies on spoken queries instead of textual queries making it a language independent task. In general, the queries and test utterances are generated by different speakers in different languages with varying acoustic conditions and without constraints on vocabulary, pronunciation lexicon, accents etc. Thus, the search is performed relying only on acoustic data of the query and test utterances with no language specific resources, as a zero-resource task. It is essentially a pattern matching problem in the context of speech data where the targeted pattern is the information represented using speech signal and given to the system as a spoken query.

A QbE-STD system finds great application in searching through multimedia content produced by news agencies, radio broadcast channel, internet, social media etc. These contents are massive and are generally produced by a large diverse

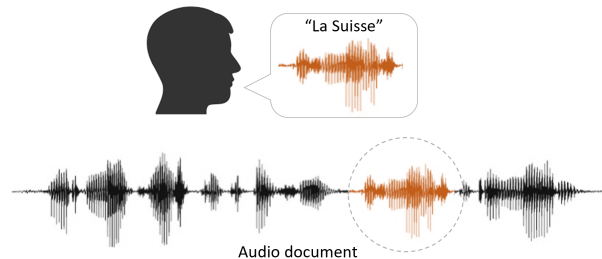


Fig. 1. Query-by-Example Spoken Term Detection

group of people in multiple different languages. The search through this data still relies on its textual description which may not be always available or may be insufficient for representing the complete contents of data. Therefore, text based retrieval algorithms give very limited search results. Moreover, it is desirable to search through those contents using speech as a natural and generic medium of communication.

State-of-the-art QbE-STD systems primarily rely on DTW based template matching techniques to find the spoken queries in test utterances. This approach involves the following two steps: (i) extraction of suitable feature vectors from the queries and test utterances, (ii) employing those features to estimate the likelihood of the query occurring somewhere in the test utterance as a sub-sequence. Spectral features [2], [3], posterior features (posterior probability vector for phone or phone-like units) [4], [5] as well as bottleneck features (representation obtained from the bottleneck layer of a neural network) [6], [7] have been used for this task. The matching likelihood is generally obtained using a dynamic time warping (DTW) algorithm on the frame-level similarity matrix computed from the feature vectors of the query and each audio document. Several variants of DTW have been proposed to deal with sub-sequence detection problem: Segmental DTW [2], [4], Slope-constrained DTW [8], Sub-sequence DTW [9], Subspace-regularized DTW [10], [11] etc.

Previously in [12], we proposed to cast the template matching problem as binary classification of images. Feature vectors from the spoken query and test utterances are used to compute frame-level similarities in a matrix form. This matrix contains a quasi-diagonal pattern if the query occurs in the test utterance. A convolutional neural network (CNN) based classifier is trained to identify the pattern and make a decision about the occurrence of the query. This approach is shown to perform significantly better than the best DTW based system using concatenation of multiple monolingual phone posteriors.

In this work, we use bottleneck feature representation in-

D. Ram is with Idiap Research Institute, Martigny 1920, Switzerland (e-mail: dhananjay.iem@gmail.com)

L. Miculicich and H. Bourlard are with Idiap Research Institute, Martigny 1920, Switzerland and École Polytechnic Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland. (email: firstname.lastname@idiap.ch)

stead of posterior features as it has been shown to perform better with DTW based matching [6]. The monolingual features used in those cases suffer from the language mismatch problem during DNN based feature extraction. To deal with this problem, we train multilingual networks aimed at obtaining language independent representation. These multilingual bottleneck features are used for both DTW and CNN based matching. Finally, we integrate the representation learning and CNN-based matching to jointly train and further improve the QbE-STD performance. Different components of this system are implemented separately to analyze their performance before building the end-to-end system. The contributions of this paper are summarized in the following:

- **Representation Learning** (Section III): In contrast to using several language dependent bottleneck features for QbE-STD, here we propose to train multilingual bottleneck networks to estimate language independent representation of the query and test utterances. This is achieved by using multitask learning principle [13] to jointly classify phones from multiple languages and the shared network is able to learn language independent representation. These representations are employed to estimate the query detection likelihood using both DTW (Section IV) and CNN based Matching.
- **CNN based Matching** (Section V): The DTW based template matching is applied on a frame-level similarity matrix computed from the feature vectors of the query and the test utterance to estimate the likelihood score of occurrence. Unlike DTW, we view the similarity matrix as an image and propose to approach the QbE-STD problem as an image classification task. We observe that the similarity matrix contains a quasi-diagonal pattern if the query occurs in the test utterance. Otherwise, no such pattern is observed. Thus for each spoken query, a test utterance can be categorized as an example of positive or negative class depending on whether the query occurs in it or not.
- **End-to-End QbE-STD System** (Section VI): The proposed neural network based end-to-end system takes spectral features (MFCC) corresponding to a query and a test utterance as input, and the output indicates whether the query occurs in the test utterance. It has three components: (i) Feature extraction, (ii) Similarity matrix computation and (iii) CNN based matching, combined into one architecture for end-to-end training. The feature extractor aims at obtaining language independent representation to produce better score for similarity matrix which in turn improves the CNN based matching.

The proposed end-to-end QbE-STD system has the following advantages over the baseline DTW based approach: (i) the CNN based matching provides a learning framework to the problem (ii) the CNN considers the whole similarity matrix at once to find a pattern, whereas the DTW algorithm takes localized decisions on the similarity matrix to find a warping path, (iii) the CNN based matching introduces a discrimination capability in the system and (iv) the end-to-end training enables joint optimization of the representation

learning and the matching network.

The proposed methods are evaluated on SWS 2013 database and their generalization ability is analyzed on QUESST 2014 database as described in Section VIII. The significant improvements obtained using these approaches show the importance of a learning framework for QbE-STD. Finally, we present the conclusions in Section IX.

II. PRIOR WORKS

We summarize various approaches for spoken query detection in this section. Most of the successful approaches can be combined into a category called template matching, consisting of two primary steps: (i) feature extraction and (ii) matching likelihood computation. Generally, suitable feature vectors are estimated from both the spoken queries and test utterances before computing the matching likelihood between them using some variation of dynamic programming algorithm. Spectral features like Mel frequency cepstral coefficient (MFCC) or perceptual linear prediction (PLP) have been used with limited success. Posterior features estimated from Gaussian mixture model (GMM) [4] as well as deep neural network (DNN) [5], [14] yield better performance. The GMMs are generally trained in an unsupervised way where the output indicates posterior probabilities of different Gaussian components in the model [2], [4]. On the other hand, the DNNs are trained in a supervised manner using labeled data from several well resourced languages and the outputs can be posteriors of monophones, context dependent phones or senones [5], [8]. The output of the DNN is considered as an instantaneous characterization of the speech signal irrespective of the input language. Enhanced phone posteriors and phonological posteriors have also been used as speech representation [15], [16]. Both supervised and unsupervised bottleneck features from DNNs have been used for query detection [6], [7], [17], [18].

The features extracted from the spoken query and test utterance are used to compute a frame-level distance matrix (using a suitable distance metric e.g. euclidean, cosine etc). A dynamic time warping (DTW) algorithm can be used to find the least cost path through this distance matrix to determine a frame level mapping between the query and test utterance, and the accumulated cost indicates the degree of match. However, this standard DTW performs matching between two complete temporal sequences making it unsuitable for subsequence matching in our case. Segmental DTW [2], [4] deals with this problem by constraining the warping path in a predefined window. But it cannot handle utterances with large speaking rate variation, which can be solved using slope-constrained DTW [8]. It penalizes the slope of the warping path by limiting the number of frame mappings between the query and the test utterance. In sub-sequence DTW [9], the algorithm forces the cost of insertion at the beginning and end of the query to be 0, thus encouraging the warping path to start and end at any frame of the test utterance and gives us a sub-sequence matching the spoken query.

Alternative to DTW, subspace modeling of queries are used to compute a frame level score for faster detection [10], [19], [20]. These subspace scores are also used to regularize

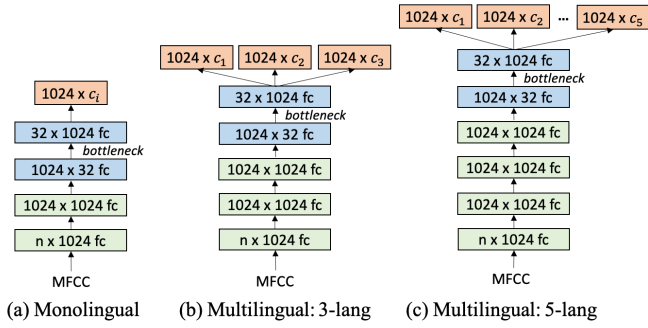


Fig. 2. Monolingual and multilingual DNN architectures for extracting bottleneck features using multiple languages. c_i is the number of classes for the i -th language and n is the size of input vector.

the distance matrix for DTW to boost the performance [10], [11]. The template matching can also be performed with a Convolutional Neural Network (CNN) while using the distance matrix as an input image to find the warping path and achieve higher accuracy [12]. Neural networks with attention mechanism has been used for this task as well [21], [22]. Additionally, the problem of acoustic and speaker mismatch is mitigated using model based approaches. These methods use hidden Markov models (HMM) to model acoustic units which are derived in an unsupervised manner. The queries and test utterances are represented using those HMMs and symbolic search techniques are used to retrieve test utterances containing the query [3], [23].

III. REPRESENTATION LEARNING

In this section, we discuss different monolingual and multilingual bottleneck features used for spoken query detection. Bottleneck features are low-dimensional representation of data generally obtained from a hidden bottleneck layer of a DNN [6], [24], [25]. This layer has a smaller number of hidden units compared to other layers, which constrains the information flow through the network during training. It enables the network to focus on the essential information from data for minimizing the final loss function. In the following, we present the DNN architectures used to obtain different types of bottleneck features.

A. Monolingual Neural Network

We train DNNs for phone classification using five languages to estimate five distinct monolingual bottleneck features. The DNN architecture consists of 3 fully connected layers of 1024 neurons each, followed by a linear bottleneck layer of 32 neurons, and a fully connected layer of 1024 neurons. The final layer feeds to the output layer of size c_i corresponding to number of classes (e.g. phones) of the i -th language. The architecture is presented in Figure 2.

The monolingual bottleneck features have previously been shown to provide good performance for this task [6]. Here, we analyze their performance and further train multilingual networks to estimate better features for QBE-STD.

B. Multilingual Neural Network

Multilingual neural networks have been studied in the context of ASR in order to obtain language independent representation of speech signal [25]. Those networks are trained using multitask learning principle [13] which aims at exploiting similarities across tasks resulting in an improved learning efficiency when compared to training each task separately. Generally, the network architecture consists of a shared part and several task-dependent parts. In order to obtain multilingual bottleneck features we model phone classification for each language as different tasks, thus we have a language independent part and a language dependent part. The language independent part is composed of the beginning and intermediate layers of the network which are shared by all languages forcing the network to learn common characteristics. The language dependent part is modeled by the output layers (marked in orange in Figure 2), and enables the network to learn particular characteristics of each language.

In this work, we train two different multilingual networks using 3 languages and 5 languages respectively in order to analyze the effect of training with additional languages. The architecture of these networks are presented in Figure 2 and described in the following.

- **Multilingual (3 languages):** this architecture consists of 4 fully connected layers having 1024 neurons each, followed by a linear bottleneck layer of 32 neurons. Then, a fully connected layer of 1024 neurons feeds to 3 output layers corresponding to the different training languages. The 3 output layers are language dependent while the rest of the layers are shared among the languages.
- **Multilingual (5 languages):** this architecture is similar to the previous one except it uses an additional fully connected layer of 1024 neurons, and two extra output layers corresponding to the 2 new languages. The increased number of layers is intended at modeling the extra training data gained by adding new languages.

All neural networks discussed in this section have rectifier linear unit (ReLU) as non-linearity used after each linear transform except in the bottleneck layer and the output layer. The output layer has multiple softmax layers corresponding to each language. Note that, we also experimented with more and less number of layers compared to the ones presented above and the architecture discussed here yields the best performance.

IV. DTW BASED TEMPLATE MATCHING

The trained neural networks discussed in previous section is used to extract different types of bottleneck features for DTW based template matching. The features from the query examples are used to construct reference templates to match with the test utterances as discussed below.

A. Query Template Construction

We construct a query template in two ways depending on the number of examples available: (i) one, (ii) more than one. In the first case, the feature vectors constitute the reference

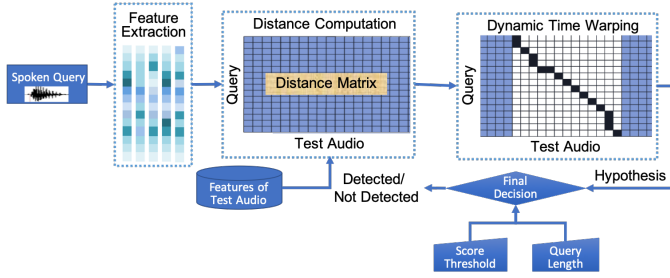


Fig. 3. Block diagram of the baseline system. We extract features vectors from a query and a test utterance to compute the corresponding distance matrix, and apply DTW to obtain the best matching sub-sequence. If the length of the hypothesis is smaller than half the query length, it is discarded to reduce false alarm rate. Otherwise, its score is compared to a threshold to yield a final decision.

template. In the other case, we construct an average template using different examples of the same query. For this purpose, we select the example with highest number of frames as reference template and use DTW [26] to obtain a frame level mapping between the reference and rest of the examples. The frames mapped together are averaged to compute the final template for matching [5], [27].

B. Template Matching

The DTW algorithm proposed in [5] for query detection is used as our baseline system. It was the best system for Spoken Web Search (SWS) in MediaEval challenge 2013 [28]. The basic framework of the system is presented in Figure 3 and is briefly discussed below.

The features of queries and test utterances are used to compute a frame level distance matrix using cosine distance [6]. A DTW algorithm (similar to the slope-constrained DTW [8]) is performed on this distance matrix to find the optimal cost path. The cost is normalized at each step using the partial path length and constraints are imposed to let the warping path begin and end at any point in the test utterance. It gives us a sub-sequence of the test utterance that optimally matches the query and the corresponding likelihood score. The resulting sub-sequences are filtered depending on their lengths to reduce the false alarms. The likelihood scores are compared with a predefined threshold to make final decision.

V. CNN BASED MATCHING

The DTW based template matching for query detection can be replaced with a CNN by casting the problem as a binary classification of images [12]; where the images are similarity matrices between the queries and test utterances. In the following, we describe this method, including the process of image construction and our CNN architecture.

A. Image Construction

The input to the CNN is composed of similarity matrices calculated between the queries and test utterances. These matrices form a quasi-diagonal pattern in the regions where a query and a test utterance match. This is caused by the high similarity values that such regions represent (see the yellow

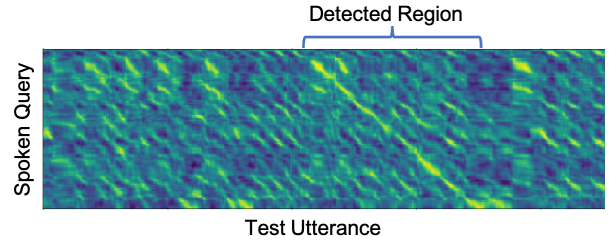


Fig. 4. Positive case: the query occurs in the test utterance

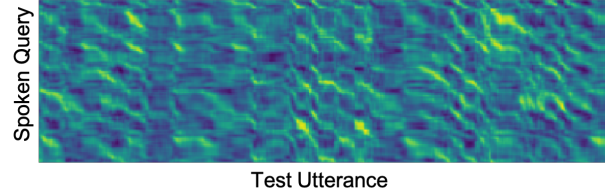


Fig. 5. Negative case: the query does not occur in the test utterance

pattern in Fig 4). To calculate the similarity matrices, first we extract bottleneck features (Section III) from both spoken queries and test utterances using MFCC features as input. Let us consider, $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m]$ and $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n]$ representing the features of a spoken query and a test utterance respectively, where m and n are the number of frames in each case. We compute cosine similarity [6] between two feature vectors \mathbf{q}_i and \mathbf{t}_j , as follows:

$$s(\mathbf{q}_i, \mathbf{t}_j) = \frac{\mathbf{q}_i \cdot \mathbf{t}_j}{\|\mathbf{q}_i\| \cdot \|\mathbf{t}_i\|} \quad (1)$$

Then, we apply a range normalization to constrain the values in the range $[-1, 1]$.

$$s_{norm}(\mathbf{q}_i, \mathbf{t}_j) = -1 + 2 \cdot \frac{(s(\mathbf{q}_i, \mathbf{t}_j) - s_{min})}{(s_{max} - s_{min})} \quad (2)$$

$$\text{where } s_{min} = \min_{i,j} (s(\mathbf{q}_i, \mathbf{t}_j)) \quad (3)$$

$$s_{max} = \max_{i,j} (s(\mathbf{q}_i, \mathbf{t}_j)) \quad (4)$$

We define two categories of images: (i) positive class, when the query occurs in the utterance, and (ii) negative class otherwise. Figures 4 and 5 show examples of these classes. The vertical and horizontal axis represent the frames of the query and test utterance respectively. The strength of values are shown with colors, yellow for high values and blue for low ones.

B. Methodology

Here we present a CNN architecture used to classify the similarity matrices defined in the previous section. The architecture is similar to a VGG network [29] that performs well in image classification task. It consists of a series of convolution and max-pooling layers with fixed sized filters and numbers of feature maps for all layers, simplifying the hyperparameter selection process. We have one channel similarity matrix as input instead of the three channel RGB color images generally used in standard image classification tasks. The detailed architecture is described in Table I where

TABLE I
CNN ARCHITECTURE

| Layer | Description |
|---------|--|
| Input | $100 \times 800 \times 1$ |
| Maxpool | Channel: in=1, out=1, Filter: 2x2, Stride: 2 |
| Conv | Channel: in=1, out=30, Filter: 3x3, Stride: 1 |
| Conv | Channel: in=30, out=30, Filter: 3x3, Stride: 1 |
| Maxpool | Channel: in=30, out=30, Filter: 2x2, Stride: 2 |
| Conv | Channel: in=30, out=30, Filter: 3x3, Stride: 1 |
| Conv | Channel: in=30, out=30, Filter: 3x3, Stride: 1 |
| Maxpool | Channel: in=30, out=30, Filter: 2x2, Stride: 2 |
| Conv | Channel: in=30, out=30, Filter: 3x3, Stride: 1 |
| Conv | Channel: in=30, out=30, Filter: 3x3, Stride: 1 |
| Maxpool | Channel: in=30, out=30, Filter: 2x2, Stride: 2 |
| Conv | Channel: in=30, out=30, Filter: 3x3, Stride: 1 |
| Conv | Channel: in=30, out=15, Filter: 3x3, Stride: 1 |
| Maxpool | Channel: in=15, out=15, Filter: 2x2, Stride: 2 |
| FC | Input: $1 \times 23 \times 15$, Output=60 |
| FC | Input:60, Output=2 |
| SM | Input:2, Output=2 |

Conv: Convolution; FC: Fully connected; SM: Softmax

convolution layers use ReLU [30] as activation function. The number of channels and dropout were optimized to 30, and 0.1 respectively with a development set. The training label for the network indicates whether the query occurs in a test utterance corresponding to a input similarity matrix. The training data can be constructed from any pair of spoken queries and test utterances from any language with minimal supervision, as we only need the information if a query is part of the test utterance, without requiring the full transcription. Note that, we also performed experiments with simpler architectures and expected good performance due to the simplicity of the task. However, those experiments with less number of layers failed to outperform the baseline system.

The training of CNN for query detection poses the following two main challenges:

- **Variable size input:** The CNN architecture discussed earlier requires fixed size input, but our similarity matrices have variable lengths and widths due to the varying duration of corresponding spoken queries and test utterances. We solve this problem by fixing the size of all input matrices to a predetermined length and width (in our case 100×800). Bigger matrices are down-sampled by deleting its rows and/or columns in regular intervals. On the other hand, smaller matrices are increased in size by filling the gap with the lowest value from the corresponding similarity matrices. The down-sampling step does not affect the desired quasi-diagonal pattern severely as the deleted rows and columns are spread throughout the similarity matrix. Also, we did not segment the test utterances in fixed size intervals to perform detection on each segment separately, as it requires the region of occurrence of a query in a test utterance as ground-truth label, which is not available for QbE-STD.
- **Unbalanced data:** The number of positive and negative samples is highly unbalanced for the query detection task (in our training data is 0.1% to 99.9% respectively), due to the very small frequency of occurrence of a given query in the test utterances. We solve this problem by creating a balanced training set for each training epoch. We choose

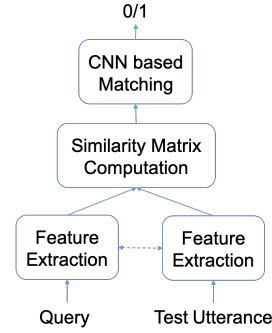


Fig. 6. Neural network based end-to-end architecture for QbE-STD. The two feature extraction blocks share the same set of parameters.

all positive examples and randomly sample the same number of negative examples from the corresponding set. We also considered using weighted loss function for training, however the experiments showed that our strategy yields better performance.

VI. END-TO-END QBE-STD SYSTEM

In this section, we propose a novel neural network based end-to-end architecture to perform QbE-STD. We combine the representation learning network with the CNN based matching network in one architecture such that the input to the network are MFCC features corresponding to a query and a test utterance, and the output indicates whether the query occurs in the test utterance. We discuss this architecture and the training procedure in the following sections.

A. Architecture

The end-to-end architecture has 3 components as shown in Figure 6: (i) Feature extraction, (ii) Similarity matrix computation and (iii) CNN based matching. The feature extraction block is used to obtain a frame-level representation using MFCC features as input for both the query and test utterance. The goal of this block is to obtain a language independent representation which produces better frame-level similarity score to construct the similarity matrix. This block can be implemented using DNN, CNN or long short term memory (LSTM) network, and we use DNN for this purpose.

We can use any of the 3 architectures presented in Section III as our feature extraction block. However, we observe that the multilingual network trained using 5 languages generates the best bottleneck features for query detection (see Section VIII). Thus we use this architecture as feature extraction block for our end-to-end system. We use the language independent part of the network (first 5 layers, until bottleneck layer) to extract features from both the query and test utterance which feeds to the second block of our architecture.

The second block of our architecture computes a frame-level similarity matrix between the query and the test utterance using cosine similarity as described in Section V-A. This similarity matrix is input to the CNN to produce a matching score as discussed in Section V-B. This whole network is jointly optimized by training it in an end-to-end manner as discussed in the following section.

B. Training Challenges

The end-to-end network faces same challenges as the CNN based matching network due to the nature of the problem as discussed in Section V-B. In addition, we do not have sufficient data to train this network from scratch. Thus, we use the principle of transfer learning [31] to initialize different blocks of this network using previously trained network instead of random initialization. The CNN based matching block is initialized with the trained network from Section V and the feature extraction block is initialized with the first 5 layers of the 5 language neural network presented in Section III-B. The weight matrices corresponding to CNN based matching block can be frozen during training to enable the system to only train the feature extraction block. In this setting, the CNN based matching block can be viewed as a loss function to extract better features. These feature vectors should be able to produce more discriminative quasi-diagonal patterns (as discussed in Section V-A) required to classify the positive examples from the negative ones.

VII. EXPERIMENTAL SET-UP

In this section, we describe the databases used to train and evaluate different systems. Then, we discuss the training procedure for *Representation Learning*, *CNN based Matching* and the *End-to-End* system. We also present the preprocessing steps to perform the experiments and different evaluation metrics used to test and compare our systems.

A. Databases

We use GlobalPhone database [32] to train the monolingual as well as multilingual models presented in Section III. The QbE-STD experiments are performed on Spoken Web Search (SWS) 2013 [28] and Query by Example Search on Speech Task (QUESST) 2014 [33] databases using *DTW based Template Matching*. Then, we use the SWS 2013 dataset to train the *CNN based Matching* network as well as the *End-to-End* network and evaluate the corresponding models. We use the QUESST 2014 dataset to show the generalization ability of those models.

- (i) **GlobalPhone Corpus:** GlobalPhone [32] is a multilingual speech database consisting of high quality recordings of read speech with corresponding transcription and pronunciation dictionaries in 20 different languages. It was designed to be uniform across languages in terms of audio quality (type of microphone, noise condition, channel), the collection scenario (task, setup, speaking style), phone set conventions (IPA-based naming of phone) etc. In this work, we use French (FR), German (GE), Portuguese (PT), Spanish (ES) and Russian (RU) to train monolingual as well as multilingual networks and extract the corresponding bottleneck features for QbE-STD experiments. We have an average of ~ 20 hours of training and ~ 2 hours of development data per language.
- (ii) **Spoken Web Search (SWS) 2013:** The SWS 2013 database is part of the MediaEval challenge 2013 [28] for evaluating QbE-STD systems. It consists of speech data from 9 different low-resourced languages: Albanian,

TABLE II
NUMBER OF DIFFERENT TYPES OF QUERIES AVAILABLE IN SWS 2013, PARTITIONED ACCORDING TO THE NUMBER OF EXAMPLES PER QUERY.

| Query Set | Examples per query | | |
|-------------|--------------------|-----|----|
| | 1 | 3 | 10 |
| Development | 311 | 100 | 94 |
| Evaluation | 310 | 100 | 93 |

Basque, Czech, non-native English, Isixhosa, Isizulu, Romanian, Sepedi and Setswana. It was collected from different sources with varying acoustic conditions and in different amounts from each languages. The variety of data reduces the possibility of over-fitting. There are 505 queries in the development set and 503 queries in the evaluation set. The queries are categorized into 3 types depending on the number of examples available per query, as shown in Table II. The search space consists of 20 hours of audio with 10762 utterances.

- (iii) **Query by Example Search on Speech Task (QUESST) 2014:** The QUESST 2014 database is part of the MediaEval challenge 2014 [33] that we use to evaluate the generalizability of different approaches. It consists of ~ 23 hours of speech data (12492 files) in 6 languages as search corpus: Albanian, Basque, Czech, non-native English, Romanian and Slovak. The development and evaluation set has 560 and 555 queries respectively which were separately recorded than the search corpus. We did not use this dataset for training or tuning our models. Unlike SWS 2013 dataset, all queries have only one example available. There are three types of occurrences of a query defined as a match in this dataset. Type 1: exact matching of the lexical representation of a query (same as in SWS 2013), Type 2: slight lexical variations at the start or end of a query, and Type 3: multiword query occurrence with different order or filler content between words (Refer [33] for more details).

B. Bottleneck Feature Extraction

We use Kaldi toolkit [34] to extract MFCC features with corresponding ‘delta’ and ‘delta-delta’ coefficients, and generate the target labels for training different neural networks presented in Section III. MFCC features with a context of 6 frames (both left and right) constitutes the input vector of size 507. The context size is optimized using the development queries in SWS 2013. The outputs are monophone states (also known as pdfs in kaldi) corresponding to each language in GlobalPhone corpus. These training labels are generated using a GMM-HMM based speech recognizer [14]. The number of classes corresponding to French, German, Portuguese, Spanish and Russian are 124, 133, 145, 130, 151 respectively. Note that, we also trained these networks using senone classes, however they perform worse than the monophone based training.

We apply layer normalization [35] before the linear transforms and use rectifier linear unit (ReLU) as non-linearity after each linear transform except in the bottleneck layer for both monolingual and multilingual networks. We train those networks with batch size of 255 samples and dropout of

0.1. In case of multilingual training, we use equal number of samples from each language under consideration. Adam optimization algorithm [36] is used with an initial learning rate of 10^{-3} to train all networks by optimizing cross entropy loss. The learning rate is halved every time the development loss increases compared to the previous epoch until the learning rate reaches a value of 10^{-4} . All the networks were trained for 50 epochs.

We extract bottleneck features from these trained networks and apply speech activity detection (SAD) before using them for DTW as well as CNN based matching. The SAD relies on the silence and noise class posterior probabilities obtained from three different phone classifiers (Czech, Hungarian and Russian) [37] trained on SpeechDAT(E) database [38]. These probabilities are averaged and compared with rest of the phone class probabilities to identify and remove the noisy frames. Audio files with less than 10 frames after SAD are not used for detection experiments, however those are considered during evaluation [5], [10], [12].

C. CNN Training

The search space for QbE-STD in SWS 2013 database is shared between the development and evaluation queries. The labels for these queries indicate whether a query occurs in a test utterance or not. There is no training set available, thus we only have these queries to train our CNN. We split the 505 development queries in two sets of 495 and 10 queries respectively for training and tuning the model. Due to the multiple examples available for a subset of queries, we effectively have 1551 query examples. Our experiments are designed in this manner to follow the setup of SWS 2013 task and make a fair comparison.

We filter the queries and test utterances using a SAD discussed in previous section to obtain 1488×10750 training example pairs. It constitutes 24118 positive examples, and rest are negative examples. We balance the data for each training epoch by following the strategy presented in Section V-B. We shuffle the training example pairs and use a batch size of 20 samples. We use the Adam optimization algorithm [36] with an initial learning rate of 10^{-4} to optimize cross entropy loss.

D. End-to-End Training

The training and development sets for the network presented in Section VI-A consists of the same pairs of queries and test utterances as used to train the CNN in previous section. The difference is: the CNN uses bottleneck features, whereas the end-to-end network uses the corresponding MFCC features. We attempt to train the network by randomly initializing the weight matrices of the whole network. However those trained models yield very poor detection performance. This can be attributed to the limited training data as well as the complexity of the problem. Thus, we begin the training by initializing different blocks of the model with corresponding pre-trained networks as discussed in Section VI-B. In order to limit the trainable parameters, we progressively freeze the first few layers of the feature extraction block and train separate networks. In this case of end-to-end training, the

frame-level speech activity detection (SAD) (as discussed in Section VIII-A) is performed on the output of feature extraction network before using them to compute the similarity matrix. It is not applied on the MFCC features in order to avoid discontinuities in the contextual input vectors. In both CNN and End-to-End training, we use loss on the development set to avoid over-fitting. The End-to-End networks converge faster, within ~ 20 epochs, whereas the CNN takes 356 epochs to converge because we train it from scratch.

Finally, we normalize the score outputs from all the systems to have zero-mean and unit-variance per query in order to reduce variability across different queries [5], [10], [12] for evaluation. All neural network architectures presented in this work are implemented using Pytorch [39] *.

E. Evaluation Metric

We use minimum normalized cross entropy (C_{nxe}^{min}) as primary metric and maximum Term Weighted Value ($MTWV$) secondary metric to evaluate the performance of different systems [40]. C_{nxe} quantifies the knowledge that a QbE-STD system has on the ground truth, and C_{nxe}^{min} is obtained by calibrating the system scores. More precisely, it computes the information provided by the detection scores of a given system such that $C_{nxe}^{min} \approx 0$ indicates a perfect system and $C_{nxe}^{min} = 1$ shows a non-informative system. $MTWV$ is computed by taking into account the miss and false alarm rates as well as the corresponding costs. We consider cost of false alarm (C_{fa}) to be 1 and cost of missed detection (C_m) to be 100. We also perform one-tailed paired samples t-test to compute the significance of performance improvement in any comparison.

VIII. EXPERIMENTAL ANALYSIS

We conducted extensive experiments to evaluate and compare the query detection performance of different systems presented in this paper: (i) *DTW based Matching*, (ii) *CNN based Matching* and (iii) *End-to-End* neural network model.

A. DTW based Template Matching

We perform DTW based template matching using bottleneck features extracted from the monolingual and multilingual networks discussed in Section III and present their detection performance on SWS 2013 and QUESST 2014 databases.

1) *Performance on SWS 2013*: We consider two cases depending on the number of examples per query to evaluate different bottleneck features for QbE-STD. In case of a single example per query, the corresponding features constitute the template. On the other hand, with multiple examples per query we compute an average template before performing the detection experiment as discussed in Section IV-A. The C_{nxe}^{min} and $MTWV$ scores for query detection using both monolingual and multilingual bottleneck features are shown in Table III. We can see that the Portuguese (PT) feature performs the best among the monolingual features with very close performance from Spanish (ES) feature.

*https://github.com/idiap/E2E_QbE_STD

TABLE III

PERFORMANCE OF THE DTW BASED TEMPLATE MATCHING APPROACH IN SWS 2013 USING MONOLINGUAL AND MULTILINGUAL BOTTLENECK FEATURES FOR SINGLE AND MULTIPLE EXAMPLES PER QUERY USING ALL EVALUATION QUERIES.

| Training Language | Single Example | | Multiple Examples | |
|-------------------|-----------------------------|-----------------|-----------------------------|-----------------|
| | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ |
| Portuguese (PT) | 0.6771 | 0.3786 | 0.6478 | 0.3963 |
| Spanish (ES) | 0.6776 | 0.3754 | 0.6501 | 0.3967 |
| Russian (RU) | 0.7035 | 0.3184 | 0.6767 | 0.3383 |
| French (FR) | 0.7021 | 0.333 | 0.6757 | 0.3511 |
| German (GE) | 0.7503 | 0.2643 | 0.7257 | 0.2919 |
| PT-ES-RU | 0.6330 | 0.4305 | 0.6023 | 0.4478 |
| PT-ES-RU-FR-GE | 0.6204 | 0.4358 | 0.5866 | 0.4580 |

TABLE IV

PERFORMANCE OF THE DTW BASED TEMPLATE MATCHING APPROACH IN QUESST 2014 USING MONOLINGUAL AND MULTILINGUAL BOTTLENECK FEATURES FOR DIFFERENT TYPES OF QUERIES IN EVALUATION SET.

| Training Language(s) | T1 Queries | | T2 Queries | | T3 Queries | |
|----------------------|-----------------------------|-----------------|-----------------------------|-----------------|-----------------------------|-----------------|
| | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ |
| Portuguese (PT) | 0.5582 | 0.4671 | 0.6814 | 0.3048 | 0.8062 | 0.1915 |
| Spanish (ES) | 0.5788 | 0.4648 | 0.7074 | 0.2695 | 0.8361 | 0.1612 |
| Russian (RU) | 0.6119 | 0.4148 | 0.7285 | 0.2434 | 0.8499 | 0.1385 |
| French (FR) | 0.6266 | 0.4242 | 0.7462 | 0.2086 | 0.8522 | 0.1249 |
| German (GE) | 0.6655 | 0.3481 | 0.7786 | 0.1902 | 0.8533 | 0.1038 |
| PT-ES-RU | 0.4828 | 0.5459 | 0.6218 | 0.3626 | 0.7849 | 0.2057 |
| PT-ES-RU-FR-GE | 0.4606 | 0.5663 | 0.6013 | 0.3605 | 0.7601 | 0.2138 |

The 3 language and 5 language network, as discussed in Section III-B are trained using (PT, ES, RU) and (PT, ES, RU, FR, GE) languages respectively. The 3 language network uses the 3 best performing monolingual training languages. The results in Table III show that both multilingual features perform significantly better than the best monolingual feature. We also observe that PT-ES-RU-FR-GE features significantly outperform PT-ES-RU features indicating that additional languages for training provide better features for QbE-STD.

2) *Performance on QUESST 2014*: We have only one example per query in case of QUESST 2014 dataset, thus the corresponding bottleneck features constitute the template. It has three different types of queries as discussed in Section VII-A. Similar to [41], we did not employ any specific strategies to deal with those different types of queries. The C_{nxe}^{\min} and $MTWV$ scores corresponding to different types of queries using both monolingual and multilingual features is shown in Table IV. We can see that the bottleneck feature from Portuguese (PT) performs the best among monolingual features for all three types of queries. We have a similar observation as in SWS 2013 that PT-ES-RU-FR-GE network performs better than PT-ES-RU network indicating that more languages for training helps in obtaining better features for DTW.

B. CNN based Matching

We use the best performing features (PT-ES-RU-FR-GE) in the previous set of experiments to train a *CNN based Matching* of queries and test utterances and compare the results.

TABLE V

PERFORMANCE COMPARISON OF *DTW based Matching*, *CNN based Matching* AND *End-to-End* NEURAL NETWORK MODEL FOR QbE-STD IN SWS 2013 USING SINGLE AND MULTIPLE EXAMPLES OF ALL EVALUATION QUERIES.

| System | Single Example | | Multiple Examples | |
|--------------|-----------------------------|-----------------|-----------------------------|-----------------|
| | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ |
| DTW Matching | 0.6204 | 0.4358 | 0.5866 | 0.4580 |
| CNN Matching | 0.6078 | 0.3986 | 0.5767 | 0.4115 |
| End-to-End | 0.5339 | 0.4412 | 0.5207 | 0.4654 |

TABLE VI

PERFORMANCE COMPARISON OF *DTW based Matching*, *CNN based Matching* AND *End-to-End* NEURAL NETWORK MODEL FOR QbE-STD IN QUESST 2014 USING DIFFERENT TYPES OF QUERIES IN EVALUATION SET.

| System | T1 Queries | | T2 Queries | | T3 Queries | |
|--------------|-----------------------------|-----------------|-----------------------------|-----------------|-----------------------------|-----------------|
| | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ |
| DTW Matching | 0.4606 | 0.5663 | 0.6013 | 0.3605 | 0.7601 | 0.2138 |
| CNN Matching | 0.4121 | 0.6103 | 0.5235 | 0.4375 | 0.6569 | 0.3603 |
| End-to-End | 0.3796 | 0.6499 | 0.5158 | 0.4433 | 0.6278 | 0.3617 |

1) *Performance on SWS 2013*: We present the performance of *CNN based Matching* and compare it with the corresponding *DTW based Matching* in Table V. Similar to DTW based system, we use template averaging for queries with multiple examples. This method was followed during test time, however the training samples were formed using only single example per query. We observe from Table V that the *CNN based Matching* performs significantly better in terms of C_{nxe}^{\min} score but worse in terms of $MTWV$ score for both single and multiple examples per query case. This shows that for a specific operating point, the DTW matching performs better than CNN matching (as $MTWV$ score corresponds to an operating point), however the CNN produces more informative scores about the ground-truth than the DTW.

2) *Performance on QUESST 2014*: We use the model trained on SWS 2013 for testing on QUESST 2014 evaluation set to analyze the generalizability of CNN based matching system. We compare the performance of DTW and CNN based matching in Table VI. As discussed earlier, it has three types of queries and we do not apply any specific strategy to deal with them. We can clearly see that CNN performs significantly better than DTW for all 3 types of queries. The performance gets increasingly worse from Type 1 to Type 2 and from Type 2 to Type 3. This can be attributed to the training of our system using only queries from SWS 2013 which are similar to Type 1 queries from QUESST 2014. However the consistency in performance improvement for all kinds of queries shows that CNN based matching system is generalizable to new datasets.

C. End-to-End QbE-STD System

We utilize the bottleneck feature extractor and *CNN based Matching* network to construct the *End-to-End* QbE-STD system as discussed in Section VI and analyze its performance on both SWS 2013 and QUESST 2014 databases. We also discuss that the *CNN based Matching* network can be used

TABLE VII

PERFORMANCE OF THE *End-to-End* NEURAL NETWORK BASED APPROACH IN SWS 2013 FOR SINGLE AND MULTIPLE EXAMPLES PER QUERY USING ALL EVALUATION QUERIES. DIFFERENT NUMBER OF LAYERS IN THE FEATURE EXTRACTOR BLOCK WERE FROZEN TO TRAIN WITH LIMITED DATA.

| # of layers frozen | Single Example | | Multiple Examples | |
|-----------------------|-----------------------------|-----------------|-----------------------------|-----------------|
| | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ |
| 3 | 0.5555 | 0.4328 | 0.5396 | 0.4552 |
| 2 | 0.5637 | 0.4417 | 0.5541 | 0.4557 |
| 1 | 0.5522 | 0.4461 | 0.5395 | 0.4682 |
| 0 | 0.5339 | 0.4412 | 0.5207 | 0.4654 |

TABLE VIII

PERFORMANCE OF THE END-TO-END NEURAL NETWORK BASED APPROACH IN QUESST 2014 FOR DIFFERENT TYPES OF QUERIES IN EVALUATION SET. DIFFERENT NUMBER OF LAYERS IN THE FEATURE EXTRACTOR BLOCK WERE FROZEN TO TRAIN WITH LIMITED DATA.

| # of layers frozen | T1 Queries | | T2 Queries | | T3 Queries | |
|-----------------------|-----------------------------|-----------------|-----------------------------|-----------------|-----------------------------|-----------------|
| | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ |
| 3 | 0.3881 | 0.6395 | 0.5238 | 0.4362 | 0.6254 | 0.3669 |
| 2 | 0.3796 | 0.6499 | 0.5158 | 0.4433 | 0.6278 | 0.3617 |
| 1 | 0.3888 | 0.6309 | 0.5124 | 0.4513 | 0.6148 | 0.3793 |
| 0 | 0.4268 | 0.6190 | 0.5338 | 0.4338 | 0.6591 | 0.3646 |

as a loss function to obtain better features for DTW based template matching.

1) *Performance on SWS 2013*: We follow the procedure described in Section VII-D to train the *End-to-End* network using SWS 2013 database. We freeze the first few layers of the feature extractor while keeping the rest of network trainable and show the corresponding results in Table VII. Similar to previously presented systems, we use template averaging to obtain the template for queries with multiple examples. However, the template averaging is performed after the query examples are forward passed through the feature extractor. We can see from Table VII that the best performance is obtained by training all layers of the feature extractor. It shows that the problem of limited training data can be alleviated by pre-training different parts of the network before end-to-end training.

2) *Performance on QUESST 2014*: The generalization ability of the models trained on SWS 2013 is evaluated using QUESST 2014 database and the results are presented in Table VIII. We observe that T1 queries perform best with the model trained using 2 frozen layers, whereas T2 and T3 queries perform best with the model trained using 1 frozen layer. It can be attributed to the training of the models using SWS 2013, which enables the network to optimize for that database when fine-tuning all layers of the feature extractor.

3) *CNN based Matching as Loss Function*: In the *End-to-End* model, we can freeze the parameters of the *CNN based Matching* network and consider it as a loss function for fine tuning the feature extraction network. This loss function enables the feature extractor to learn and generate features which produce more discriminative similarity matrices to be classified by the CNN. It can be observed through the performance of the system. We use the features obtained

TABLE IX

PERFORMANCE OF THE DTW BASED TEMPLATE MATCHING APPROACH USING MULTILINGUAL BOTTLENECK FEATURES WHICH ARE FINE TUNED USING CNN BASED LOSS FUNCTION. THE EXPERIMENTS WERE PERFORMED USING EVALUATION QUERIES IN SWS 2013 FOR SINGLE AND MULTIPLE EXAMPLES PER QUERY.

| # of layers frozen | Single Example | | Multiple Examples | |
|-----------------------|-----------------------------|-----------------|-----------------------------|-----------------|
| | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ | $C_{nxe}^{\min} \downarrow$ | $MTWV \uparrow$ |
| 3 | 0.5788 | 0.4633 | 0.5521 | 0.4888 |
| 2 | 0.5705 | 0.4708 | 0.5539 | 0.4914 |
| 1 | 0.5607 | 0.4719 | 0.5429 | 0.4894 |
| 0 | 0.5718 | 0.4597 | 0.5593 | 0.4738 |
| Bottleneck | 0.6204 | 0.4358 | 0.5866 | 0.4580 |

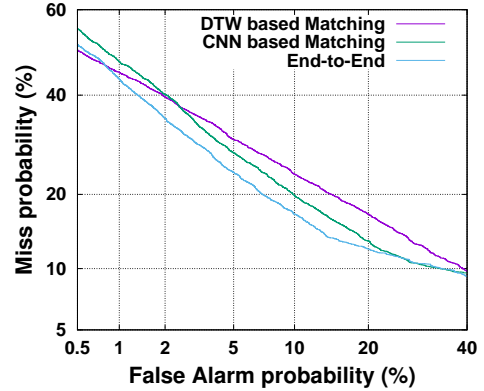


Fig. 7. DET curves comparing the performance of *DTW based Matching*, *CNN based Matching* and *End-to-End* system on SWS 2013 database using evaluation queries with single example.

after fine-tuning the network to perform *DTW based Matching* and compare it with the best performance obtained using bottleneck features as shown in Section VIII-A. Similar to previous experiment, we progressively freeze different number of layers of the feature extractor and the results are presented in Table IX. We observe that the feature extractor retrained with 1 frozen layer gives the best results which is significantly better than the bottleneck features indicating the importance of CNN based loss function.

D. System Comparisons

Here, we present a final comparison of different systems discussed in this work.

1) *C_{nxe}^{\min} and $MTWV$ scores*: The comparisons corresponding to SWS 2013 and QUESST 2014 databases are presented in Tables V and VI respectively. We observe that the *CNN based Matching* performs significantly better than the *DTW based Matching* in both metrics for QUESST 2014, but for SWS 2013 the improvement is observed only in terms of C_{nxe}^{\min} . The *End-to-End* system performs significantly better than other systems in both databases, in both metrics. However, the performance gain of *End-to-End* system over *CNN based Matching* is smaller in QUESST 2014 compared to SWS 2013 database. It could be due to the fine tuning of *End-to-End* system on SWS 2013 database.

2) *DET curves*: We present the same system comparison using DET curves in Figures 7 and 8 respectively. In case

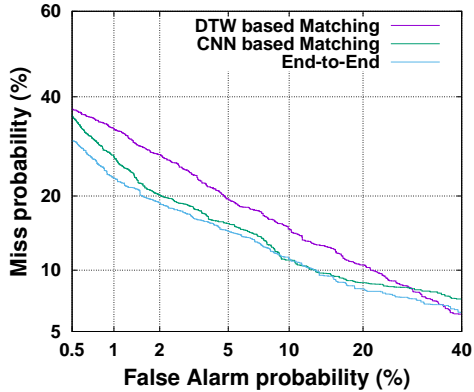


Fig. 8. DET curves comparing the performance of *DTW based Matching*, *CNN based Matching* and *End-to-End* system using T1 evaluation queries of QUESST 2014 database.

of SWS 2013 database, we compare the performance using single example per query, and for QUESST 2014 database, we compare T1 query performance. In both databases the *CNN based Matching* and *End-to-End* system performs better than the *DTW based Matching* except for low false alarm rates.

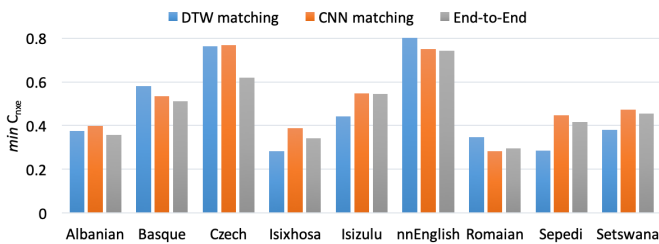


Fig. 9. Comparison of QbE-STD performance of language specific evaluation queries (single example per query) of SWS 2013 using C_{nxe}^{\min} values (lower is better)

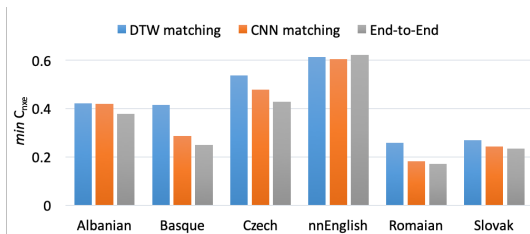


Fig. 10. Comparison of QbE-STD performance of language specific evaluation queries (T1 query) of QUESST 2014 using C_{nxe}^{\min} values (lower is better)

3) *Language Specific Performance*: We compare the language specific query performance using C_{nxe}^{\min} values in Figures 9 and 10 respectively. In SWS 2013 database, the experiments are performed using single examples per query. This comparison shows that the performance of *CNN based Matching* and *End-to-End* system are worse than the *DTW based Matching* for ‘Isixhosa’, ‘Isizulu’, ‘Sepedi’ and ‘Setswana’ indicating that the performance gains are not uniform through-

out different languages. This is due to the considerably less amount of training data corresponding to those languages.

In QUESST 2014 database, we compare the T1 query performances. Similar to SWS 2013 database, non-uniform performance improvement is observed for queries of different languages. The performance is marginally worse only for ‘non-native English’ queries in *End-to-End* system.

IX. CONCLUSION

In this paper, we implemented several monolingual as well as multilingual neural networks to extract bottleneck features for QbE-STD and show that more training languages give better performance. We implemented a *CNN based Matching* approach for QbE-STD using those bottleneck features. It enables discriminative learning between positive and negative classes, which is not featured in *DTW based Matching* systems. It gives significant improvement over the best DTW system with bottleneck features. Then, we proposed to integrate the bottleneck feature extractor with the *CNN based Matching* network to provide an end-to-end learning framework for QbE-STD. It gives further improvement over the CNN based matching approach. Both the *CNN based Matching* and *End-to-End* system are generalizable to other database, giving significant improvement over the *DTW based Matching*. We also show that the CNN matching block in the *End-to-End* system can be used as a loss function to obtain better language independent features which can be useful for other tasks e.g. unsupervised unit discovery. In future, we plan to extend our approach using triplet loss, and use better up-sampling and down-sampling methods to resize the similarity matrices. We also intend to investigate our techniques using bigger databases and include more languages. The codes are available at: https://github.com/idiap/E2E_QbE_STD

ACKNOWLEDGMENT

The research leading to these results has received funding from the Swiss NSF projects on ‘‘Parsimonious Hierarchical Automatic Speech Recognition and Query Detection (PHASER-QUAD)’’, grant agreement number 200020-169398, and ‘‘Learning Representations of Abstraction for Opinion Summarisation (LAOS)’’, grant number FNS-30216. We also thank the reviewers for their insightful remarks to improve this manuscript.

REFERENCES

- [1] D. Ram, ‘‘Language independent query by example spoken term detection,’’ Ph.D. dissertation,  cole Polytechnique F d rale de Lausanne, 2019.
- [2] A. S. Park and J. R. Glass, ‘‘Unsupervised pattern discovery in speech,’’ *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.
- [3] C.-a. Chan and L.-s. Lee, ‘‘Model-based unsupervised spoken term detection with spoken queries,’’ *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1330–1342, 2013.
- [4] Y. Zhang and J. R. Glass, ‘‘Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams,’’ in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

- [5] L. J. Rodríguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "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, pp. 7819–7823.
- [6] I. Szöke, M. Skácel, L. Burget, and J. Cernocký, "Coping with channel mismatch in query-by-example-but QUESST 2014," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5838–5842.
- [7] H. Chen, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Unsupervised bottleneck features for low-resource query-by-example spoken term detection," in *INTERSPEECH*, 2016, pp. 923–927.
- [8] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.
- [9] M. Müller, *Information retrieval for music and motion*. Springer, 2007, vol. 2.
- [10] D. Ram, A. Asaei, and H. Bourlard, "Sparse subspace modeling for query by example spoken term detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1130–1143, June 2018.
- [11] —, "Subspace regularized dynamic time warping for spoken query detection," in *Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, 2017.
- [12] D. Ram, L. Miculicich, and H. Bourlard, "CNN based query by example spoken term detection," in *Nineteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.
- [13] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
- [14] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.
- [15] D. Ram, A. Asaei, and H. Bourlard, "Phonetic subspace features for improved query by example spoken term detection," *Speech Communication*, vol. 103, pp. 27–36, 2018.
- [16] A. Asaei, D. Ram, and H. Bourlard, "Phonological posterior hashing for query by example spoken term detection," *Proc. Interspeech 2018*, pp. 2067–2071, 2018.
- [17] H. Chen, C. C. Leung, L. Xie, B. Ma, and H. Li, "Multitask feature learning for low-resource query-by-example spoken term detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1329–1339, Dec 2017.
- [18] D. Ram, L. Miculicich, and H. Bourlard, "Multilingual bottleneck features for query by example spoken term detection," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec 2019, pp. 621–628.
- [19] D. Ram, A. Asaei, P. Dighe, and H. Bourlard, "Sparse modeling of posterior exemplars for keyword detection," in *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.
- [20] D. Ram, A. Asaei, and H. Bourlard, "Subspace detection of DNN posterior probabilities via sparse representation for query by example spoken term detection," in *Seventeenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2016.
- [21] C.-W. Ao and H.-y. Lee, "Query-by-example spoken term detection using attention-based multi-hop networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6264–6268.
- [22] K. Zhang, Z. Wu, J. Jia, H. Meng, and B. Song, "Query-by-example spoken term detection using attentive pooling networks," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 1267–1272.
- [23] C.-y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 2012, pp. 40–49.
- [24] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *Twelfth annual conference of the international speech communication association*, 2011.
- [25] K. Veselý, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 336–341.
- [26] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [27] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
- [28] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodríguez-Fuentes, "The Spoken Web Search task," in *the MediaEval 2013 Workshop*, 2013.
- [29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [31] L. Y. Pratt, J. Mostow, C. A. Kamm, and A. A. Kamm, "Direct transfer of learned information among neural networks," in *AAAI*, vol. 91, 1991, pp. 584–589.
- [32] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8126–8130.
- [33] X. Anguera, L. J. Rodríguez-Fuentes, I. Szoke, A. Buzo, F. Metze, and M. Penagarikano, "Query-by-example spoken term detection evaluation on low-resource languages," in *The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)*, 2014.
- [34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.
- [35] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [36] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [37] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology BUT, 2008.
- [38] P. Pollák, J. Boudy, K. Choukri, H. Van Den Heuvel, K. Vicsi, A. Virag, R. Siemund, W. Majewski, P. Staroniewicz, H. Tropf *et al.*, "Speechdat (e)-eastern european telephone speech databases," in *the Proc. of XLDB 2000, Workshop on Very Large Telephone Speech Databases*. Citeseer, 2000.
- [39] A. Paszke, S. Gross, and S. Chintala, "Pytorch," 2017, [online] <http://pytorch.org/>.
- [40] L. J. Rodríguez-Fuentes and M. Penagarikano, "Mediaeval 2013 spoken web search task: system performance measures," *n. TR-2013-1, Department of Electricity and Electronics, University of the Basque Country*, 2013.
- [41] L. J. Rodríguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "GTTS-EHU systems for QUESST at mediaeval 2014," in *MediaEval*, 2014.



Dhananjay Ram received his MTech in Electrical Engineering from Indian Institute of Technology Kanpur, India. He held a research engineer position at the same institute before pursuing PhD. He received his PhD in 2019 from École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. His PhD thesis focused on building Language Independent Query by Example Spoken Term Detection system. He has more than 8 years of research experience on application of machine learning for speech processing, natural language processing as well as computer vision related problems. He has done several research internships in both industry and academia. He was a PostDoc scholar for several months at Idiap Research Institute, Switzerland and, currently he is a scientist in Amazon.com Inc working in natural language understanding towards improving Amazon's digital assistant Alexa. He has several publications in various international conferences and journals. His research interests lie in the areas of machine learning, deep learning and Bayesian inference.



Lesly Miculicich received her Master of Science (MS) in Computer Science from University of Bern, Fribourg and Neuchatel in Switzerland. Currently, she is a PhD student at École Polytechnic Fédérale de Lausanne (EPFL) in Electrical Engineering and research assistant at Idiap Research Institute, Switzerland. Her PhD thesis focuses on building Document Level Neural Machine Translation. She has several years of work experience in the area of computer security for the banking industry. She has also done several research internships in GoogleAI

and Microsoft AI & Research, as well as in academia. Her research interests lie in the areas of machine learning and deep learning with application in natural language understanding.



Hervé Bourlard received the Electrical and Computer Science Engineering degree and the PhD degree in Applied Sciences both from “Faculté Polytechnique de Mons”, Mons, Belgium. After having been a member of the Scientific Staff at the Philips Research Laboratory of Brussels and an R&D Manager at L&H SpeechProducts, he is now Director of the Idiap Research Institute, Full Professor at the Swiss Federal Institute of Technology Lausanne (EPFL), and Founding Director of the Swiss NSF National Centre of Competence in Research on “Interactive Multimodal Information Management (IM2)” (2001-2013). Having spent (since 1988) several long-term and short-term visits (initially as a Guest Scientist) at the International Computer Science Institute (ICSI), Berkeley, CA, he is now an ICSI External Fellow and a member of its Board of Trustees. His research interests mainly include statistical pattern classification, signal processing, multi-channel processing, artificial neural networks, and applied mathematics, with applications to a wide range of Information and Communication Technologies, including spoken language processing, speech and speaker recognition, language modeling, multimodal interaction, augmented multi-party interaction, and distant group collaborative environments.

H. Bourlard is the author/coauthor/editor of 8 books and over 330 reviewed papers (including one IEEE paper award) and book chapters. He is a Fellow of IEEE and ISCA, and a Senior Member and Member of the European Council of ACM. He is (or has been) a member of the program/scientific committees of numerous international conferences (e.g., General Chairman of IEEE Workshop on Neural Networks for Signal Processing 2002, Co-Technical Chairman of IEEE ICASSP 2002, General Chairman of Interspeech 2003) and on the Editorial Board of several journals (e.g., past co-Editor-in-Chief of “Speech Communication”). He is the recipient of several scientific and entrepreneurship awards.