# Joint Learning of Distance Metric and Query Model for Posteriorgram-Based Keyword Search

Batuhan Gündoğdu, *Student Member, IEEE*, Bolaji Yusuf, and Murat Saraçlar, *Member, IEEE*

***Abstract*—In this paper, we propose a novel approach to keyword search (KWS) in low-resource languages, which provides an alternative method for retrieving the terms of interest, especially for the out of vocabulary (OOV) ones. Our system incorporates the techniques of query-by-example retrieval tasks into KWS and conducts the search by means of the subsequence dynamic time warping (sDTW) algorithm. For this, text queries are modeled as sequences of feature vectors and used as templates in the search. A Siamese neural network-based model is trained to learn a frame-level distance metric to be used in sDTW and the proper query model frame representations for this learned distance. Experiments conducted on Intelligence Advanced Research Projects Activity Babel Program's Turkish, Pashto, and Zulu datasets demonstrate the effectiveness of our approach. In each of the languages, the proposed system outperforms the large vocabulary continuous speech recognition (LVCSR) based baseline for OOV terms. Furthermore, the fusion of the proposed system with the baseline system provides an average relative actual term weighted value (ATWV) improvement of 13.9% on all terms and, more significantly, the fusion yields an average relative ATWV improvement of 154.5% on OOV terms. We show that this new method can be used as an alternative to conventional LVCSR-based KWS systems, or in combination with them, to achieve the goal of closing the gap between OOV and in-vocabulary retrieval performances.**

***Index Terms*—Keyword search, low resource languages, out of vocabulary (OOV) terms, query modeling, distance metric learning, subsequence dynamic time warping.**

## I. Introduction

THE primary focus of this study is a subtask of spoken content retrieval called keyword search (KWS), also known as spoken term detection (STD), which has recently gained a great deal of interest within the speech processing community. KWS is defined as the retrieval of the locations of a user-provided term in an untranscribed speech archive. In contrast to a related task called *keyword spotting* [5], in KWS, the keywords are not known during model training and system set-up. Since the query is given in text form, a convenient approach to KWS would be using an LVCSR system and executing text retrieval on the LVCSR output [6]. Since LVCSR systems may fail to return reliable transcriptions under low resource and adverse recording conditions, the performance of the KWS systems that use them will be adversely effected. In order to alleviate this problem, the retrieval is conducted on stochastic structures like lattices or confusion networks obtained from the LVCSR systems instead of the single-best transcriptions. The detection scores for the query term are obtained from the lattices and returned to the user in an ordered manner. The hypotheses are then labeled as relevant or irrelevant based on these scores. However, if the term includes words that are not within the vocabulary of the LVCSR system, in other words, if the term is OOV for the LVCSR system, the lattice will not include such a term. In order to cope with this OOV problem, many approaches have been proposed in the literature. These approaches are discussed in detail in Section II. The method proposed in this paper also provides an alternative approach to the handling of OOV terms.

The presence of OOV terms constitutes a rather significant problem for retrieval in conversational speech, especially for low-resource languages. Recently, the MediaEval campaign [7] has made an effort to address the zero-resource keyword search with the query-by-example spoken term detection (QbE-STD) task. In QbE-STD, the query is also provided in audio form and searched in a speech archive. Out of the many approaches applied throughout the campaign, the frame level similarity matching-based approaches yielded the best results [8]. In the similarity matching-based search, versions of dynamic time warping (DTW) were applied on the speech document and the query. In this study, we aim to extend these successful QbE-STD techniques to KWS, by learning a query model for the text query and a distance metric to be used in DTW-based search.

In this paper, we follow a QbE-STD methodology similar to the work done in [9] on the KWS task and conduct a DTW-based similarity search on the speech document posteriorgram. Since in KWS, the queries are provided in text form and not as actual "examples", they are modeled as "*pseudo posteriorgrams*" to be used as templates in the DTW. Furthermore, we propose a Siamese neural network-based model to learn the frame repre-

Manuscript received April 1, 2017; revised July 25, 2017 and September 19, 2017; accepted September 24, 2017. Date of publication October 11, 2017; date of current version November 16, 2017. This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Project 116E076 and in part by the Intelligence Advanced Research Projects Activity via the Department of Defense U.S. Army Research Laboratory (DoD/ARL) under Contract W911NF-12-C-0012. This paper was presented in part at the IEEE 24th Signal Processing and Communication Application Conference, Zonguldak, Turkey, May 2016 [1], in part at the the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, March 2017 [2], in part at the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 2015 [3], and in part at the 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 2017 [4]. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Mary P. Harper. *(Corresponding author: Batuhan Gündoğdu.)*

The authors are with the Department of Electrical Engineering, Boğaziçi University, Istanbul 34470 Turkey (e-mail: batuhan.gundogdu@boun.edu.tr; bolaji.yusuf@boun.edu.tr; murat.saraclar@boun.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2017.2762080

sentations for these query *pseudo posteriorgrams* as well as a distance metric to be used with them in DTW. The experiments on IARPA Babel Program's Turkish, Pashto and Zulu datasets show that the proposed system yields promising results as a standalone system and outperforms the LVCSR-based baseline on OOV terms. Furthermore, when combined with the baseline, the proposed system provides a $13.9\%$ average relative ATWV improvement over all terms and, more significantly, the fusion system doubles the ATWV of the baseline on OOV terms on all of the languages, yielding a substantial average relative improvement of $154.5\%$.

This paper is organized as follows. In Section II, an outline of the related work on LVCSR-based KWS systems, OOV handling techniques and QbE-STD tasks is provided. In Section III, the novel application of DTW-based similarity search to KWS as well as the query generation are introduced. In Section IV, the distance metric learning and query modeling for KWS using the proposed asymmetric Siamese neural network are explained. In Section V, the experimental results are provided, and, finally in Section VI, we draw our conclusions.

## II. Related Work

In this section, we provide information on the previous research related to KWS and the OOV problem. We also include a brief review of QbE-STD in this section.

### *A. Keyword Search*

The main work on KWS was initiated with the 2006 NIST Spoken Term Detection (STD) evaluations [10]. Since then, the main approach adopted for KWS has been first obtaining lattices from speech through LVCSR systems, and subsequently running the search using weighted finite state transducers (WFST) [11]–[13]. Many successful applications of KWS were implemented on broadcast news [14], video lectures [15], and web-videos [16] with this approach. As an extension to the 2006 NIST STD evaluations, which worked on KWS systems in English, Mandarin, and Arabic, the IARPA Babel Program [17] was started in 2011 to specifically address low-resource languages. To simulate the low-resource set-up, the program included limited language pack (LLP) experiments with as low as 10 hours of training data per language. Throughout the program, various techniques were studied by the participants such as discriminative training for acoustic models [18], use of multilingual features [19], and data augmentation [20]. In conjunction with the Babel program, the OpenKWS evaluation [21] was initiated in 2013 to address new "surprise" languages in a short amount of time. These selected surprise languages were Vietnamese, Tamil, and Swahili during the campaign. With these, the focus was brought to low-resource languages, where there is insufficient labeled data available to train high performance LVCSR systems to apply the conventional KWS recipe.

### *B. The OOV Problem*

One major obstacle that the LVCSR-based KWS systems face especially in low-resource or morphologically rich languages is the OOV problem. Since the LVCSR systems utilize the lexicons and the language models (LM) obtained from the limited transcribed data to generate the lattices, any keyword containing an OOV term will not exist in the lattices. While the OOV problem arises chiefly from the scarcity of training resources, it is also compounded by the very nature of queries, in that users generally search for uncommon words such as names or technical terms.

The first approach to fixing the OOV issue is to conduct the search on subword-based lattices, such as phonemes, syllables or morphemes. Such systems can be considered "open vocabulary", even though the LVCSR system has a fixed and limited vocabulary [22]–[25]. For retrieval, keywords are represented as sequences of subword units and matched against the lattice. While subword-based systems make it possible to retrieve OOV terms, they often suffer from high false alarm rates due to the absence of the lexical constraints. Hybrid systems have also been proposed [26]. These address IV and OOV keywords with different level transducers, i.e. word level for IV and subword level for OOV, a computationally expensive procedure that carries the cost of running the decoding two times and increases the size of the document index drastically.

Another effort to address the OOV problem involves searching the web for text and filtering the obtained data to extend the LVCSR lexicon and the language model. In [27], the LLP experiments were conducted on Cantonese, Pashto, Tagalog, Turkish, and Vietnamese; across five languages the ATWV score was increased in average by 0.0243.

One other approach is to expand the LVCSR lexicon prior to the lattice generation using automatically generated pronunciations. In [28], the efficacy of lexicon expansion, given that an anticipation of the OOV terms are possible, is presented. However, in low-resource KWS, such knowledge is generally unavailable in advance. Hence, a cheap yet effective approach that involves using *proxy keywords* was proposed to utilize automatically generated pronunciations of the OOV keywords after the lattice has been generated. Since the OOV keywords do not exist in the lattice, the search is conducted on acoustically similar IV words instead of the OOV words [29]. For this, the phone confusion transducers are also integrated in the proxy keyword generation [30]. The proxy keyword approach not only provides a satisfactory performance in OOV handling, but it also alleviates the high false alarm rate and the computational expense issues associated with the subword or hybrid lattice-based KWS systems respectively. Based on a low-resource Tagalog KWS experiment, it was reported in [29], that using word proxies through a phone confusion transducer in the retrieval improved the OOV ATWV by $40\%$. In this paper, the proxy keyword approach is taken as the baseline OOV handling method.

As will be described in detail in Section III, we conduct the search on frame-level acoustic similarities independent of the LVCSR system and the LM. A similar approach to ours obtained a comparable performance to the state of the art proxy keyword approach. In [31], the keywords were modeled using their phonetic indices as point process models (PPM), and the search was conducted on the document posteriorgram without using WFST's. The experiments conducted on LLP sets of Haitian, Lao, Zulu, Assamese, and Bengali show that the complementary feature of the PPM approach yields significant improvements on OOV term detection, upon combination with the LVCSR and proxy keyword-based baseline. The average relative ATWV improvement after combination was reported to be $50.5\%$ across the languages.

### C. Query by Example Spoken Term Detection

QbE-STD is defined as the task of retrieving the locations of a query term, provided by the user as an audio snippet, in an audio archive [8]. We utilize the experience gained in the MediaEval campaign in KWS, based on the rationale that the zero-resource and multilingual nature of the tasks should make their solutions extensible to low-resource settings and "surprise" languages such as those in the recent IARPA Babel and OpenKWS evaluations.

The approaches adopted for QbE-STD during the MediaEval campaign can be grouped into two main categories: symbol-based approaches and frame-based approaches. Since the task is multi-lingual and there is no LM, the lattice-based approach of the state of the art KWS is not an option for QbE-STD. Hence, the symbol-based approaches utilize the HMM-based "query model vs. filler model" likelihood maximization technique [32].

Frame-based approaches, on the other hand, are centered around running versions of the DTW algorithm for an acoustic similarity search using the query as a template. Although some works used spectral features like MFCC's to represent query and document frames [33], Gaussian or phone posteriorgrams proved to be more appealing to many other groups due to their speaker independence [34]–[36]. There are two important conclusions of this campaign: (i) The frame-based systems yield better results than the symbol-based systems and (ii) the fusion of heterogeneous systems (symbol+frame) improves performance significantly [8], [37], [38].

One other way that the QbE-STD techniques have been utilized in KWS is to conduct pseudo relevance feedback, or blind relevance feedback, to re-score the hypotheses of the LVCSR-based KWS output [39]–[41]. After a first pass, the acoustic feature segments are used to obtain DTW distances and these similarity values are used to modify the system scores to reduce false alarms.

## III. Methodology

The proposed system is inspired by the success of the frame-based approaches of the QbE-STD task and applies these techniques to KWS. These systems use a version of the well-known DTW algorithm called subsequence DTW (sDTW), in which one of the time sequences (query) is significantly shorter than the other (document) and the shorter sequence is allowed to start and end at arbitrary locations of the longer sequence [42]. We conduct the search based on frame level similarities over the phone posteriorgram obtained from the document. The major difference from the QbE-STD task is that, whereas in QbE-STD the query is also an example of the document, i.e. an audio snippet, in KWS, the query is given in text form. Therefore, modeling the text query to be used as a template in the similarity search within the document posteriorgram is essential. The flowchart of the proposed system can be seen in Fig. 1. The methodology can be summarized thus:

1) The document posteriorgram and training data alignments are obtained through a HMM-DNN-based phone decoder trained using the limited data available.
2) From the alignment, phone duration statistics are obtained.
3) An asymmetric Siamese neural network is trained with the training posterior frames and their alignment labels to learn a distance to be used in sDTW.
4) Text queries are modeled as pseudo posteriorgrams with the duration statistics and the query model parameters learned from the alignment.
5) Finally, sDTW-based search is conducted on the document and query posteriorgrams.

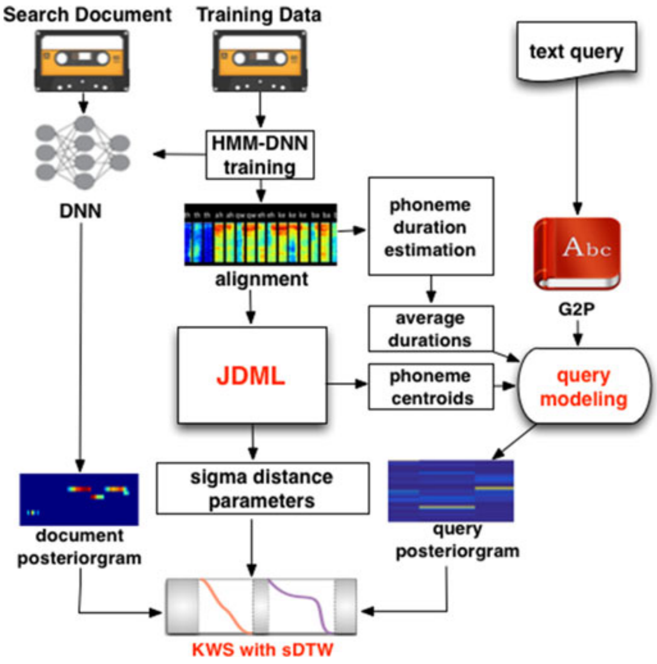


Fig. 1. Flowchart of the proposed system: A DNN is used the obtain document posteriorgram. From the training data alignment, the distance metric parameters and query model representations are learned jointly (JDML) and they are used to model the pseudo query posteriorgram to be used in sDTW-based search. The novelties are represented in red.

### A. Posteriorgram Generation

After acoustic feature extraction, the per-frame phone level posterior vectors are obtained using the Kaldi speech recognition toolkit[1] trained with the HMM-DNN[2] recipe [43]. The concatenation of these vectors is called *posteriorgram* which is simply a phone vs. time matrix. During training and decoding we used PLP features and the DNN[3] architecture of the Kaldi recipe based on p-norm activations [44]. We also store the posteriorgram and the alignment obtained from the training data to

[1] version 5.1.58
[2] babel/s5d
[3] tri6_nnet

be used in the query model and distance metric learning. We will henceforth be addressing the posteriorgrams of the search audio and the training data as document posteriorgram and training posteriorgram respectively.

### B. Query Modeling

As mentioned above, the queries need to be modeled artificially to conduct the sDTW-based similarity search within the document posteriorgram since they are given in text form. It was previously shown in [1]–[3] that the modeling of text queries as posteriorgrams using basic statistics makes the posteriorgram-based KWS feasible. After transforming the text query into a sequence of phonetic elements using the lexicon for IV words and a grapheme-to-phoneme (G2P) conversion system for OOV words, each phoneme index is replaced with its representation obtained from a look-up table. The two query modeling techniques proposed in [1] are called *binary modeling* and *average modeling*. In *binary modeling*, the frames are represented by concatenating one-hot vectors depicting the labels for the pertinent phoneme. *Average modeling* aims to incorporate the phone confusions into the model and uses means of the posterior vectors obtained from the training alignment as the phoneme representations. For the modeling of the phoneme durations, the one-hot or average vectors were repeated as many times as the average number of frames estimated from the training alignment.

In this work, on the other hand, we propose learning query frame representations that have better discriminative and representative features, rather than using one-hot vectors or the means. The details of obtaining these feature representations will be discussed in Section IV since they are learned jointly with the distance metric parameters.

### C. Subsequence DTW and Similarity Score

With the document and query posteriorgrams in hand, the sDTW algorithm [42] is used to obtain the alignments of the query with subsequences of the speech and the corresponding similarity scores. If we define the query,

$$\mathcal{Q} = \{\mathbf{q}_1, \cdots, \mathbf{q}_M\}$$

the document,

$$\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$$

and the optimal alignment path between $\mathcal{Q}$ and any subsequence of $\mathcal{X}$ as $\Phi$; the detection score for the subsequence (a hit hypothesis) is found from the average of the accumulated distance through the path using the frame-level distance measure of the sDTW algorithm,

$$\text{score} = 1 - \frac{1}{\text{length}(\Phi)} \sum_{(i,j)\in\Phi} d(\mathbf{q}_i, \mathbf{x}_j) \tag{1}$$

It is obvious from (1) that both the representation power of the frames of the query model ($\mathbf{q}_i$), and the distance metric $d(\mathbf{q}, \mathbf{x})$ are of paramount importance to the success of the KWS since the detection score is all we have to decide if a subsequence is a match or not. In the next section, we introduce the proposed distance metric learning (DML) scheme and compare it with the commonly used distance metrics.

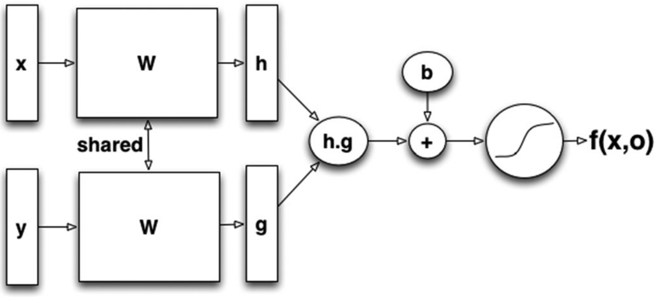


Fig. 2. The proposed DML model: The input frames are projected on the new space using the shared weight matrix $\mathbf{W}$ and the frame level similarity is obtained through a sigmoid non-linearity applied to their inner product plus a bias. The learned parameters are $\mathbf{W}$ and $b$.

## IV. Distance Metric Learning for KWS

As the distance metric of the sDTW algorithm, several measures can be used. In [45], the authors used the Euclidean distance measure with MFCC features. The cosine and logarithmic cosine distance measures are commonly used with posterior vectors [46], [47]. In [1], versions of cosine distance are compared for different query models.

Each metric has its own interpretations. The Euclidean and cosine distance metrics have geometric meanings. The Euclidean distance metric between $\mathbf{x}$ and $\mathbf{y}$ is the length of the shortest path between them in the Euclidean space. The cosine distance, on the other hand, is based on the angle between $\mathbf{x}$ and $\mathbf{y}$. The logarithmic cosine distance has been shown to be very useful in DTW-based QbE-STD tasks using posteriorgrams, since it has a probabilistic interpretation in that it can be considered as the negative log probability of the frames belonging to the same distribution [48].

As mentioned above, different distance metrics are useful for their corresponding vector spaces. In this work, we investigate learning a distance measure that works better for sDTW-based KWS. For this, we first propose the following neural network-based distance metric learning (DML) scheme similar to the Siamese networks used in signature [49] and face verification applications [50], [51]. In addition to learning the sub-space to be projected, we aim to obtain a measure that reflects the characteristics of the distribution of the data. We propose the model in Fig. 2, with the objective of obtaining a better discrimination in frame level - lower distance between examples of the same phone, higher distance between examples of different phones - by incorporating phone confusions into the distance value. We also aim to have the distance value be in range $[0, 1]$ so that it can be interpreted as the probability that two frames belong to the same phoneme. This range is also useful for pruning the scores of the hypotheses, since it facilitates meaningful thresholding. We call the distance obtained from this network the *sigma distance* [2] since a sigmoid non-linearity is used to map the weighted inner-product similarity to the desired range, with weights and the bias learned in training. Here, we use the term *distance metric* for our system loosely, since the output of the model does not satisfy the axioms of metric spaces (except non-negativity and symmetricity), and it only successfully provides a solution to our above stated objectives.

The input frames are projected onto a new space by $\mathbf{W}$ and the new sigma distance is obtained by

$$\mathbf{h} = \mathbf{W}\mathbf{x} \qquad \mathbf{g} = \mathbf{W}\mathbf{y}$$
$$f(\mathbf{x}, \mathbf{y}) = \sigma(\mathbf{h}^T \mathbf{g} + b)$$
$$d_\sigma(\mathbf{x}, \mathbf{y}) = 1 - f(\mathbf{x}, \mathbf{y}) \tag{2}$$

where

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{3}$$

### A. DML Training

As can be seen in Fig. 2, the network takes a pair of inputs and emits a scalar distance value. Hence, the training set comprises triplets $(\mathbf{x}_t, \mathbf{y}_t, r_t)$ where $r_t$ is the label indicating the kinship of the inputs $\mathbf{x}_t$ and $\mathbf{y}_t$. For the sake of simplicity, we call the pairs $(\mathbf{x}_t, \mathbf{y}_t)$ *friends* if they belong to the same phone, and *foes* otherwise. Then, as the labels $r_t$, we use 1 for friends and 0 for foes, in other words

$$r_t = \begin{cases} 1, \text{ if } & \text{class}(\mathbf{x}_t) = \text{class}(\mathbf{y}_t) \\ 0, \text{ if } & \text{class}(\mathbf{x}_t) \neq \text{class}(\mathbf{y}_t). \end{cases} \tag{4}$$

Since we have the objective of interpreting the system output as a probability value, we used the cross-entropy (CE) objective function in the back-propagation to increase the likelihood of friends outputting 1 and foes outputting 0. In our posteriorgram based KWS recipe, the pairs of friends and foes are obtained from the training posteriorgram. The objective function is then defined as:

$$J_{\text{DML}}(\mathbf{W}, b; \mathbf{x}_t, \mathbf{y}_t, r_t) = -r_t \log(f) - (1 - r_t)\log(1 - f) \tag{5}$$

where $f(\mathbf{x}_t, \mathbf{y}_t)$ is expressed as $f$ for the sake of simplicity. The gradient with respect to the parameters are found as:

$$\triangle b = r - f \tag{6}$$

and

$$\triangle \mathbf{W} = (r - f)\mathbf{W}(\mathbf{x}\mathbf{y}^T + \mathbf{y}\mathbf{x}^T). \tag{7}$$

With the gradients in hand and using learning rates $\mu$ and $\eta$, we update the parameters with on-line gradient descent as,

$$\mathbf{W} \leftarrow \mathbf{W} + \mu \triangle \mathbf{W} \quad \text{and} \quad b \leftarrow b + \eta \triangle b.$$

### B. DML as a New Kernel

Most of the known distance metrics are obtained from inner product similarities of the vectors. The distance values can be considered as applying a kernel on this inner product similarity value. If we use unit-norm vectors, these kernels can be seen below for the Euclidean (8), cosine (9) and logarithmic cosine (10) distance metrics.

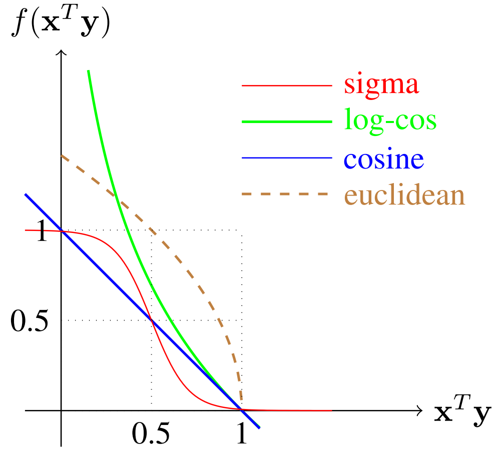


Fig. 3. Analysis of kernel functions for each distance metric: the $x$-axis represents the inner product and the $y$-axis represents the kernel that makes it a dissimilarity measure. The sigma distance parameters are taken to be $\mathbf{W} = 6I$ to fit in the same scale.

$$d_{\text{euc}}(\mathbf{x}, \mathbf{y}) = \sqrt{2 - 2\mathbf{x}^T\mathbf{y}} \tag{8}$$
$$d_{\text{cos}}(\mathbf{x}, \mathbf{y}) = 1 - \mathbf{x}^T\mathbf{y} \tag{9}$$
$$d_{\text{log-cos}}(\mathbf{x}, \mathbf{y}) = -\log(\mathbf{x}^T\mathbf{y}) \tag{10}$$

In Fig. 3, we provide a visualization of these kernel functions along with the proposed kernel of the sigma distance. The smooth and symmetric behavior of the sigmoid kernel function could be desirable for KWS. We see that for low similarity values, the log-cosine distance gives extremely high distance values. This may prevent us finding keywords when there is a lot of pronunciation variability and phone confusion. Furthermore, for the sigma distance, the inner product of two vectors may be increased or decreased with the $\mathbf{W}$ matrix to meet the desired distance value, whereas this is not an option for the other distance measures.

### C. Joint Query Modeling and DML

It was shown in [1] that use of different distance measures and query models in the posteriorgram based KWS affects the performance of the system. In this paper we aim to learn a distance metric and the appropriate query model for it. To incorporate the query modeling into DML, we add an extra layer at the query side and call it Joint DML (JDML). The proposed JDML model is shown in Fig. 4.

While the goal of DML is to learn a distance measure that models the frame level phone confusions, minimizes the average distance between friends and maximizes the average distance between foes, JDML is optimized for KWS purposes. Its goal is to simultaneously learn proper representations for each class (phoneme) and the distance parameters. Hence, now the input of the network is a pair, composed of the posterior vectors of the training alignment and a one-hot vector $(\mathbf{x}_t, \mathbf{o}_t)$. The output

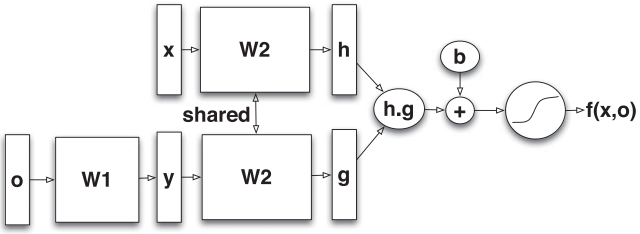


Fig. 4. JDML asymmetric Siamese Network Model: The document frame is used as $\mathbf{x}$ on the top of the network, and a one-hot vector depicting the phoneme index in the query model is inputted as $\mathbf{o}$. The $\mathbf{W}_1$ matrix gives the proper query frame representations to work with the sigma distance presented on the upper layer of the network.

**Algorithm 1:** JDML with Prior Equalization..

```
Separate the dataset into subset of classes:
    $Set-class(i) = \{\mathbf{x}_t|, \mathbf{x}_t \in C_i\}_{i=1}^K$
Initialize network parameters $\mathbf{W}_1, \mathbf{W}_2$ and $b$, set the
    learning rates $\mu_1, \mu_2, \eta$
repeat
    sample a class $C_i$ randomly $1 \leq i \leq K$
    sample $\mathbf{x}$ from $Set-class(i)$ randomly, set
        $\mathbf{o} = I(:, i)$
    Calculate $f(\mathbf{x}, \mathbf{o}) = \sigma(\mathbf{x}^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1\mathbf{o} + b)$ and the
        gradients (Fig. 4)
    Update network parameters for this pair of friends
        $\mathbf{W}_1 \leftarrow \mathbf{W}_1 + \mu_1\triangle\mathbf{W}_1$, $\mathbf{W}_2 \leftarrow \mathbf{W}_2 + \mu_2\triangle\mathbf{W}_2$ and
        $b \leftarrow b + \eta\triangle b$
    sample a class $C_i$ randomly $1 \leq i \leq K$
    sample $\mathbf{x}$ from $Set-class(i)$ randomly
    sample a class $C_j$ $1 \leq j \leq K, j \neq i$, (foes of $C_i$)
    set $\mathbf{o} = I(:, j)$
    Calculate $f(\mathbf{x}, \mathbf{o}) = \sigma(\mathbf{x}^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1\mathbf{o} + b)$ and the
        gradients
    Update network parameters for this pair of foes
        $\mathbf{W}_1 \leftarrow \mathbf{W}_1 + \mu_1\triangle\mathbf{W}_1$, $\mathbf{W}_2 \leftarrow \mathbf{W}_2 + \mu_2\triangle\mathbf{W}_2$ and
        $b \leftarrow b + \eta\triangle b$
until convergence
```

labels are the desired sigma similarities, given as

$$r_t = \mathbf{o}_t[\text{class}(\mathbf{x}_t)]$$

For any $\mathbf{x}_t$ in training, the prior of each class will be non-uniform. Also, for any $\mathbf{x}_t$, the number of foe centroids will be much higher than the number of the friend centroids. So, there is a risk of the system learning to favor 0 outputs. To overcome these two problems, we propose the JDML algorithm with prior equalization given in Algorithm-1. Prior equalization is very useful and can be utilized on other machine learning problems with unequal priors.

### *D. Interpretations of JDML*

In this section, we provide mathematical and intuitive interpretations and justifications for the proposed JDML recipe.

*1) JDML as a Representation Learner:* JDML can be considered to be a representation learner for query frames in the initial layer on the query side, and a distance metric learner on the second layer. We solve for the model in Fig. 4 again using the cross entropy cost function given in (5). The gradients are calculated following the on-line gradient descent recipe. Since the architecture of JDML is the same as DML after the first layer, the gradients for $b$ and $\mathbf{W}_2$ are given as in (6) and (7) respectively. The gradient for $\mathbf{W}_1$ is found by:

$$\triangle\mathbf{W}_1 = \frac{dJ}{d\mathbf{W}_1} = \frac{dJ}{df}\frac{df}{dz}\frac{dz}{d\mathbf{W}_1} = (r - f)(\mathbf{W}_2^T\mathbf{W}_2\mathbf{x}\mathbf{o}^T) \tag{11}$$

where $z = \mathbf{x}^T\mathbf{W}_2^T\mathbf{W}_2\mathbf{W}_1\mathbf{o} + b$ and $f(\mathbf{x}_t, \mathbf{o}_t)$ is denoted as $f$ for the sake of simplicity.

*2) JDML as Clustering:* We can also consider JDML as the task of finding the centroids of phoneme classes with respect to sigma distance $d_\sigma(\mathbf{x}, \mathbf{y})$ (2). Similar to the k-means clustering in Euclidean space, we take the derivative of the cost function and find the optimum. But this time the distance measure is sigma distance (instead of Euclidean distance) and the cost is total cross entropy (instead of MMSE). If we denote the class centroids as $\mathbf{m}_k, k = 1\cdots K$, and the class labels $r_{t,k} \overset{\Delta}{=} \delta(\text{class}(\mathbf{x}_t) = k)$, $k = 1\cdots K, t = 1\cdots N$.

$$J_{\text{clustering}}(\mathbf{m}_k, r_{t,k}) = \\ -\sum_t\sum_k\Big[r_{t,k}\text{log}(f_{t,k}) + (1 - r_{t,k})\text{log}(1 - f_{t,k})\Big] \tag{12}$$

where $f_{t,k} = \sigma(\mathbf{x}_t^T\mathbf{W}^T\mathbf{W}\mathbf{m}_k + b)$

If we take the derivative of $J$ with respect to $\mathbf{m}_k$ and equate to zero, we get

$$\begin{aligned} \frac{dJ}{d\mathbf{m}_k} &= 0 = -\sum_t(r_{t,k} - f_{t,k})\mathbf{W}^T\mathbf{W}\mathbf{x}_t \\ \sum_t r_{t,k}\mathbf{x}_t &= \sum_t f_{t,k}\mathbf{x}_t \\ \sum_{\forall\mathbf{x}_t\in C_k}\mathbf{x}_t &= \sum_t f_{t,k}\mathbf{x}_t \end{aligned} \tag{13}$$

This result makes sense, in that the optimal centroids are obtained at locations where $f_{t,k} = 1$ if $\mathbf{x}_t$ belongs to class $C_k$ and zero otherwise. Since there is no closed-form solution, we approach with gradient descent. The iteration of $\mathbf{m}_k$ is

$$\begin{aligned} \mathbf{m}_k &= \mathbf{m}_k - \eta\frac{dJ}{d\mathbf{m}_k} \\ \mathbf{m}_k &= \mathbf{m}_k + \eta\left(\sum_t(r_{t,k} - f_{t,k})\mathbf{W}^T\mathbf{W}\mathbf{x}_t\right) \end{aligned} \tag{14}$$

The equation (14) shows us that the class centroids $\mathbf{m}_k$ are the weighted and projected sum of all training samples. Weights are decided by the error; essentially, *friends are added and foes are subtracted.* This approach differs from the Euclidean distance in another aspect; while in Euclidean k-means, only the samples of the pertinent class are averaged, here all samples are taken into account in the calculation.

The update results obtained by the two approaches (representation learning and clustering) yield the same mathematical result. The equations (11) and (14) denote the same operations since the update on $\mathbf{W}_1$ takes place only on the pertinent column of the matrix (because of the one-hot vector on the outer product). Here the columns of $\mathbf{W}_1$ become the centroids of classes with respect to sigma distance ($\mathbf{m}_k$). One advantage of this approach is that we can update the sigma distance parameters ($\mathbf{W}_2$ and $b$) and the class centroids ($\mathbf{W}_1$) at the same time.

## V. Experimental Results

The proposed posteriorgram-based KWS system was tested on IARPA Babel Program's Turkish[4], Pashto[5] and Zulu[6] limited language pack (LLP) development and evaluation datasets. 10-hour transcribed training data is used both for the DNN and the DML/JDML training. The phone durations were estimated from the training alignment and the system development experiments for KWS were conducted on the 10-hour development data for each language. The development keyword list sizes were 307 (219 IV, 88 OOV), 2065 (1465 IV, 600 OOV) and 2000 (1194 IV, 806 OOV) for Turkish, Pashto and, Zulu respectively. We refer to this set as *dev-set*.

After running the search on the dev-set and learning the the KWS parameters (such as optimal decision threshold values, score normalization method, pruning thresholds etc.), we tested our system on the part of the evaluation data (evalpart 1), which was also provided by the Babel Program. The evaluation audio was about 5 hours for each language and the keyword lists had 3171 (1955 IV, 1216 OOV), 4203 (3232 IV, 971 OOV) and 3310 (2198 IV, 1112 OOV) for Turkish, Pashto, and Zulu respectively. We refer to this set as *eval-set*.

### A. Evaluation Metrics

We tested the performance of the proposed system using the widely used Term Weighted Value (TWV) metric [10] which is a linear combination of the precision and recall and defined as:

$$TWV(th) = 1 - [P_{miss}(th) + \beta P_{fa}(th)] \tag{15}$$

where $P_{miss}(th)$ and $P_{fa}(th)$ are probabilities of misses and false alarms respectively at the threshold $th$ and $\beta$ is the regularization constant to set the relative cost of false alarms versus misses ($\beta = 999.9$). The actual TWV (ATWV) is the system performance result for a set global threshold, the maximum TWV (MTWV) is defined as the ATWV obtained at the best global threshold, the optimum TWV (OTWV) is the MTWV obtained by applying the optimum threshold values for each keyword and the supremum TWV (STWV) is defined as the OTWV where the false alarms are not punished.

In addition to TWV, we also evaluated the normalized cross entropy cost ($C_{nxe}$) which was one of the main metrics in the MediaEval Campaign [8]. The $C_{nxe}$ is a measure of the distance of the system hypotheses from the ground truth (target vs. non-target) in an information theoretic sense. It measures not only the accuracy of the classifications, but also the confidence i.e. the separation of the scores of relevant and irrelevant hypotheses. Developing a system to minimize $C_{nxe}$ would allow for operation on a wider range of threshold values and, consequently, an improvement of the TWV is achieved. Similar to MTWV, $C_{nxe}^{\min}$ is the $C_{nxe}$ of an optimally calibrated system using an affine transform of the scores [52].

### B. System Set-Up

*1) Baseline System:* The baseline uses an LVCSR based system with a DNN acoustic model, as described in [53]. From the word lattices generated by the Kaldi speech recognition toolkit [43], a WFST based search is conducted on the keyword indices. The phonemic transcriptions of the OOV keywords are obtained using the Sequitur G2P system [54] and the OOV search is conducted using the proxy keyword method [29]. The baseline system uses KST normalization of scores [55].

*2) Posteriorgram Based KWS Systems:* For development, we tested 6 different posteriorgram-based KWS systems designed using different query models and distance measures:

1) CB: cosine distance on binary query model,
2) CA: cosine distance on average query model,
3) LB: log-cosine distance on binary query model,
4) LA: log-cosine distance on average query model,
5) DML: sigma distance on average query model,
6) **JDML**: simultaneous query modeling and distance metric learning system (proposed)

Starting from the Turkish dev-set, as we progressed on the system development, we eliminated the poor performing systems and continued with the more fruitful ones. The comparisons of the proposed system against the CB, CA, LA and, LB are useful to observe the efficacy of the proposed distance, while the comparison against the DML system showed the contribution of the simultaneous query modeling for the new distance measure.

### C. Individual System Results

In the individual development of the proposed KWS systems, we conducted the search using all of the systems and observed the MTWV, OTWV and STWV using sum-to-one (STO) [55] normalization. MTWV shows the systems' ability to discriminate correct hits from false alarms, while OTWV shows the potential for further score normalization and STWV signifies the detection power. The initial results can be seen in Fig. 5.

We observed that DML and JDML yield higher performances over all evaluation metrics compared to other systems. Another notable observation was that OTWVs of the DML based systems are significantly higher than those of other systems. Hence, as the next step, we tested the systems using various normalization techniques and continued the system development with CA, CB, DML and JDML systems.

On the individual systems, we applied various normalization techniques, such as histogram equalization [56], z-norm [57], m-norm [58] and the recently proposed b-norm, m2-norm and

[4] babel105b-v0.4 (dev:kwlist, eval:kwlist2)
[5] babel104b-v0.4bY (dev:kwlist3, eval:kwlist4)
[6] babel206b-v0.1e (dev:kwlist3, eval:kwlist4)

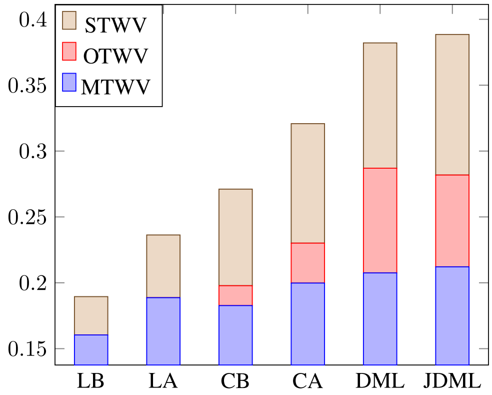


Fig. 5. Individual system results on the Turkish dev-set with STO.

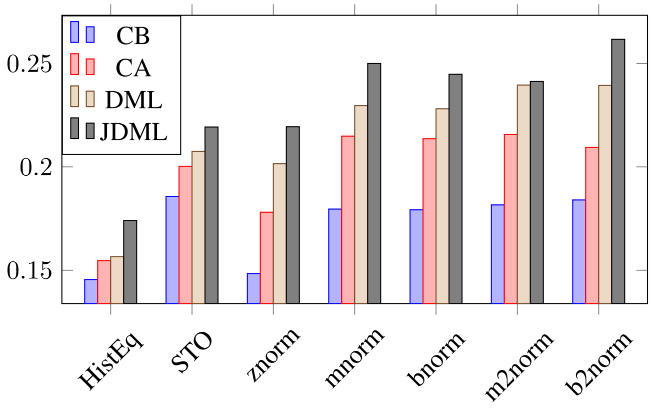


Fig. 6. MTWV performances of individual systems on Turkish dev-set under several normalizations.

b2-norm [59]. One of the most commonly used normalization techniques for KWS is keyword specific thresholding (KST). KST is based on the assumption that the raw score of a detection is the probability of it being correct. An estimate of the number of true occurrences in the document is obtained by summing the system scores. Since the posteriorgram based KWS systems can practically return scores for every subsequence of the document, such an estimate, and the KST normalization by extension, is not feasible. We observe that the proposed JDML system yields consistently better results than all other systems across the normalization techniques. The MTWV's after various normalizations can be seen in Fig. 6.

### *D. System Fusion Results*

As stated earlier, one of the most significant conclusions of the QbE-STD campaign and the Babel program was that the calibration and fusion of heterogeneous systems improve the overall performance. The proposed posteriorgram based KWS adopts a very different approach to the problem, hence the main intention was to improve the baseline performance in combination. Just like the individual system performances, the fusion of systems performed the best under the b2-norm normalization. The combination MTWV and $1 - C_{nxe}^{\min}$ scores obtained on the Turkish development set show a similar improvement trend as can be seen in Fig. 7.

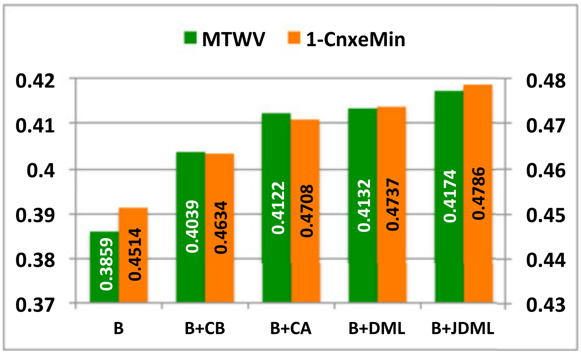


Fig. 7. MTWV and $1 - C_{nxe}^{\min}$ results after system fusion on Turkish dev-set.

TABLE I
THE DEV-SET SYSTEM FUSION RESULTS

| | | | B | JDML | B+JDML | Gain(%) |
|---|---|---|---|---|---|---|
| Turkish | MTWV | ALL | 0.3859 | 0.2617 | **0.4174** | 8.1 |
| | | IV | 0.4657 | 0.2723 | **0.4871** | 4.6 |
| | | OOV | 0.1880 | 0.2421 | **0.2538** | 35.0 |
| | OTWV | ALL | 0.4169 | 0.4175 | **0.5488** | 31.6 |
| | | IV | 0.4896 | 0.4200 | **0.5899** | 20.5 |
| | | OOV | 0.2317 | 0.4110 | **0.4440** | 91.6 |
| Pashto | MTWV | ALL | 0.2107 | 0.1101 | **0.2272** | 7.8 |
| | | IV | 0.2527 | 0.1009 | **0.2577** | 2.0 |
| | | OOV | 0.0842 | 0.1442 | **0.1591** | 89.0 |
| | OTWV | ALL | 0.3278 | 0.2305 | **0.3761** | 14.7 |
| | | IV | 0.3774 | 0.2144 | **0.3895** | 3.2 |
| | | OOV | 0.1786 | 0.2790 | **0.3355** | 87.9 |
| Zulu | MTWV | ALL | 0.2268 | 0.1816 | **0.3062** | 35.0 |
| | | IV | 0.3174 | 0.1776 | **0.3383** | 6.6 |
| | | OOV | 0.0915 | 0.2566 | **0.2597** | 183.7 |
| | OTWV | ALL | 0.3302 | 0.3409 | **0.4619** | 39.9 |
| | | IV | 0.4622 | 0.3380 | **0.4746** | 2.7 |
| | | OOV | 0.1333 | 0.3780 | **0.4428** | 232.2 |
| Average | MTWV | ALL | 0.2745 | 0.1845 | **0.3169** | 17.0 |
| | | IV | 0.3452 | 0.1836 | **0.3610** | 4.4 |
| | | OOV | 0.1212 | 0.2143 | **0.2242** | 102.5 |
| | OTWV | ALL | 0.3583 | 0.3296 | **0.4622** | 28.7 |
| | | IV | 0.4430 | 0.3241 | **0.4847** | 8.8 |
| | | OOV | 0.1812 | 0.3560 | **0.4074** | 137.2 |

Table I shows the improvements on MTWV and OTWV for IV and OOV terms separately. Although the improvement on IV terms after fusion is quite modest; on OOV terms, the proposed system outperforms the proxy keyword approach individually and provides significant improvements upon fusion.

The experiments conducted on the eval-set with a different keyword list proved the system's robustness. Using the optimal threshold values that produced the MTWV in the dev-set, we calculated the ATWV on the eval-set. The eval-set performances for each language are shown in Table II.

### *E. OOV Performance of the Proposed System*

The proposed posteriorgram-based KWS system aims to assist the LVCSR-based systems, especially on OOV terms. The experiments showed that not only did it improve the proxy keyword-based baseline ATWV by $154.5\%$ relative upon

TABLE II
THE EVAL-SET SYSTEM FUSION RESULTS

| | | | B | JDML | B+JDML | Gain(%) |
|---|---|---|---|---|---|---|
| Turkish | ATWV | ALL | 0.2848 | 0.1586 | **0.3219** | 13.0 |
| | | IV | 0.3595 | 0.1600 | **0.3791** | 5.4 |
| | | OOV | 0.0955 | 0.1551 | **0.1770** | 85.4 |
| | MTWV | ALL | 0.2861 | 0.1646 | **0.3286** | 14.8 |
| | | IV | 0.3725 | 0.1693 | **0.3862** | 3.7 |
| | | OOV | 0.0955 | 0.1608 | **0.1881** | 97.0 |
| | OTWV | ALL | 0.4126 | 0.3266 | **0.4997** | 21.1 |
| | | IV | 0.4955 | 0.3225 | **0.5422** | 9.4 |
| | | OOV | 0.2028 | 0.3370 | **0.3919** | 93.3 |
| Pashto | ATWV | ALL | 0.2458 | 0.1148 | **0.2588** | 5.3 |
| | | IV | 0.2775 | 0.1126 | **0.2791** | 0.6 |
| | | OOV | 0.0473 | 0.1289 | **0.1315** | 177.9 |
| | MTWV | ALL | 0.2459 | 0.1168 | **0.2597** | 5.6 |
| | | IV | 0.2775 | 0.1143 | **0.2846** | 2.6 |
| | | OOV | 0.0629 | 0.1353 | **0.1325** | 110.5 |
| | OTWV | ALL | 0.3857 | 0.2544 | **0.4229** | 9.6 |
| | | IV | 0.4220 | 0.2499 | **0.4408** | 4.5 |
| | | OOV | 0.1581 | 0.2824 | **0.3102** | 96.2 |
| Zulu | ATWV | ALL | 0.2459 | 0.1648 | **0.3033** | 23.3 |
| | | IV | 0.3132 | 0.1631 | **0.3452** | 10.2 |
| | | OOV | 0.0630 | 0.1693 | **0.1892** | 200.3 |
| | MTWV | ALL | 0.2462 | 0.1695 | **0.3056** | 24.1 |
| | | IV | 0.3132 | 0.1702 | **0.3458** | 10.4 |
| | | OOV | 0.0651 | 0.1775 | **0.2056** | 216.0 |
| | OTWV | ALL | 0.3106 | 0.3328 | **0.4608** | 48.4 |
| | | IV | 0.3800 | 0.3259 | **0.4898** | 28.9 |
| | | OOV | 0.1219 | 0.3516 | **0.3820** | 213.4 |
| Average | ATWV | ALL | 0.2588 | 0.1461 | **0.2946** | 13.9 |
| | | IV | 0.3167 | 0.1452 | **0.3345** | 5.4 |
| | | OOV | 0.0686 | 0.1511 | **0.1659** | 154.5 |
| | MTWV | ALL | 0.2594 | 0.1503 | **0.2979** | 14.9 |
| | | IV | 0.3211 | 0.1513 | **0.3389** | 5.5 |
| | | OOV | 0.0745 | 0.1578 | **0.1754** | 141.2 |
| | OTWV | ALL | 0.3696 | 0.3046 | **0.4611** | 26.4 |
| | | IV | 0.4325 | 0.2994 | **0.4909** | 14.3 |
| | | OOV | 0.1609 | 0.3236 | **0.3614** | 134.3 |

combination, but it also outperformed the baseline as an individual system on all experiments and all metrics for OOV terms. In Fig. 8, we plot the baseline and the proposed system's MTWV and ATWV numbers obtained from the three languages' dev-set and eval-set experiments. When the OOV performances are considered, we can observe the proposed system's superiority to the baseline. We also fit a line for both on the scatter plot and see that the line for the JDML system has a slope that almost equals to 1, where the baseline is far from that. We claim that a desired performance would be independent of the vocabulary and such a system should have similar IV and OOV performances. The proposed system yields such a performance given the experimental results.

### *F. Comparison With Similar Work*

In our survey of the existing approaches, we found the work on PPMs [31] to be the most comparable to ours, not only because of the approach but also because the authors worked with the same baseline Kaldi DNN architecture and OOV-handling method that we did[7]. After fusion with the baseline, the authors were able to obtain a significant $50.5\%$ average relative ATWV improvement on OOV terms over the languages they worked on. On the Zulu dev-set experiment, for which they reported a $111\%$ $(0.09 \rightarrow 0.19)$ relative gain on the baseline, we were able to obtain a $183.7\%$ improvement on MTWV. It should be noted that the authors reported the ATWV scores based on a KWS system whose hyper-parameters are learned from a subset of the dev-set.

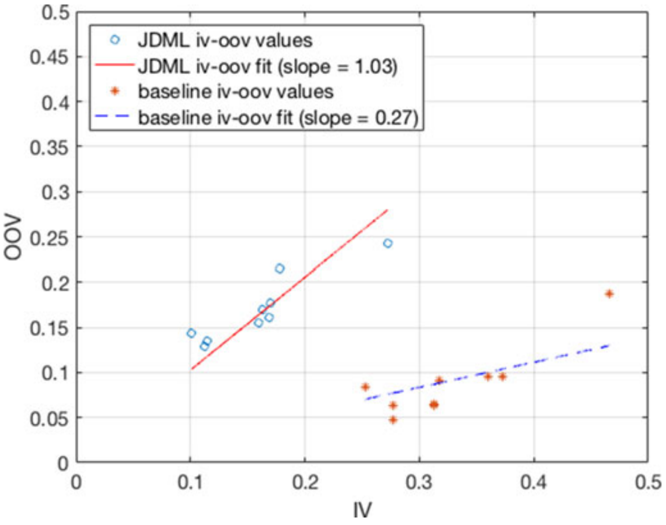


Fig. 8. The IV vs OOV ATWV and MTWV values of the baseline system and the proposed system. Each point on this scatter plot is an experiment (i.e. Pashto dev-set, Zulu eval-set etc.) and its corresponding IV vs OOV ATWV numbers.

## VI. CONCLUSION

This work has yielded many promising results and opened a leeway for further research on the topic. In this paper, we proposed a KWS system designed to perform well in low-resource conditions. For this, we conducted the search on the posteriorgram obtained from the document, based on frame-level similarities. The frame-level similarity search was conducted using the well known sDTW algorithm. An asymmetric Siamese neural network was trained to both learn a better distance measure to be used in sDTW, and frame-level representations to create the pseudo query posteriorgrams. We have tested the efficiency of the sub-parts of our system by creating various configurations of posteriorgram based KWS systems. Then, we compared the results as standalone systems and in combination with the LVCSR-based baseline system.

We have seen that learning a distance metric on the training posteriorgram not only helps incorporating the data confusion into the distance, but also yields a better KWS performance than the other distance measures as can be seen in Fig. 5. Furthermore, learning the frame representations using the JDML model makes the sigma distance more powerful since the

[7] Disclaimer: Since neither the version of Kaldi nor the keyword list used in the aforementioned work is provided, it was impossible to perform experiments under identical conditions.

representations are optimized for the new distance measure. The JDML-based systems consistently yield better KWS performances than other similar posteriorgram-based KWS systems.

As a standalone system, the proposed system outperforms the baseline on OOV terms. Furthermore, when the proposed system is combined with the baseline, the fusion system yields an average relative ATWV improvement of $13.9\%$ for all terms and a substantial average relative ATWV improvement of $154.5\%$ for OOV terms.

Another interesting observation can be made on ATWV of the proposed system averaged on 3 languages. In Table II, we see that the average ATWV's for JDML are 0.1461, 0.1452 and 0.1511 for all, IV and OOV terms, respectively. One of the goals we of this study was to close the gap between IV and OOV performances in KWS tasks. Both the average ATWV numbers' proximity to each other, and the regression line of Fig. 8 having a slope near 1 indicates that we achieved this goal.

## ACKNOWLEDGMENT

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. *Disclaimer:* The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. The authors would like to thank the reviewers of this paper for their invaluable input.

## REFERENCES

[1] B. Gündoğdu, L. Sarı, G. Çetinkaya, and M. Saraçlar, “Template-based keyword search with pseudo posteriorgrams,” in *Proc. IEEE 24th Signal Process. Commun. Appl. Conf.*, 2016, pp. 973–976.

[2] B. Gündoğdu and M. Saraclar, “Distance metric learning for posteriorgram based keyword search,” in *Proc. 42nd IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 5–9, 2017, pp. 5660–5664.

[3] L. Sarı, B. Gündoğdu, and M. Saraçlar, “Fusion of LVCSR and posteriorgram based keyword search,” in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 824–828.

[4] B. Gündoğdu and M. Saraçlar, “Similarity learning based query modeling for keyword search,” in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 3617–3621.

[5] D. A. James and S. J. Young, “A fast lattice-based approach to vocabulary independent wordspotting,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1994, vol. 1, pp. 377–380.

[6] C. Chelba, T. J. Hazen, and M. Saraçlar, “Retrieval and browsing of spoken content,” *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 39–49, May 2008.

[7] N. Rajput and F. Metze, “Spoken web search,” in *Proc. MediaEval 2011 Workshop*, 2011, pp. 1–2.

[8] X. Anguera, L. J. Rodriguez-Fuentes, I. Szoke, A. Buzo, and F. Metze, “Query-by-example spoken term detection evaluation on low-resource languages,” in *Proc. Int. Workshop Spoken Lang. Technol. Underresourced Lang.*, 2014, pp. 24–31.

[9] T. J. Hazen, W. Shen, and C. White, “Query-by-example spoken term detection using phonetic posteriorgram templates,” in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2009, pp. 421–426.

[10] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddingtion, “Results of the 2006 spoken term detection evaluation,” in *Proc. ACM SIGIR Workshop Search. Spontaneous Conversational*, 2007, pp. 51–55.

[11] C. Allauzen, M. Mohri, and M. Saraçlar, “General indexation of weighted automata: Application to spoken utterance retrieval,” *Proc. Workshop Interdiscip. Approaches Speech Indexing Retrieval HLT-NAACL 2004*, 2004, pp. 33–40, 2004.

[12] D. Can and M. Saraclar, “Lattice indexing for spoken term detection,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 8, pp. 2338–2347, Nov. 2011.

[13] M. Saraçlar and R. Sproat, “Lattice-based search for spoken utterance retrieval,” in *Proc. HLT-NAACL 2004*, 2004, vol. 51, pp. 129–136.

[14] D. Wang, J. Frankel, J. Tejedor, and S. King, “A comparison of phone and grapheme-based spoken term detection,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 4969–4972.

[15] A. Garcia and H. Gish, “Keyword spotting of arbitrary words using minimal speech resources,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 1, pp. 949–952.

[16] S.-w. Lee, K. Tanaka, and Y. Itoh, “Combining multiple subword representations for open-vocabulary spoken document retrieval,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 1, pp. 505–508.

[17] M. Harper, “IARPA Babel program,” 2014.

[18] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 2345–2349.

[19] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, “Very deep multilingual convolutional neural networks for LVCSR,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4955–4959.

[20] X. Cui, V. Goel, and B. Kingsbury, “Data augmentation for deep neural network acoustic modeling,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1469–1477, Sep. 2015.

[21] “OpenKWS14 keyword search evaluation plan.” 2013. [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf

[22] Y. He *et al.*, “Using pronunciation-based morphological subword units to improve OOV handling in keyword search,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 1, pp. 79–92, Jan. 2016.

[23] J. Mamou, B. Ramabhadran, and O. Siohan, “Vocabulary independent spoken term detection,” in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval.*, 2007, pp. 615–622.

[24] D. Karakos and R. Schwartz, “Subword and phonetic search for detecting out-of-vocabulary keywords,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2469–2473.

[25] S. Parlak and M. Saraclar, “Performance analysis and improvement of Turkish broadcast news retrieval,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 731–741, Mar. 2012.

[26] L. Burget, “Hybrid word-subword decoding for spoken term detection,” in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 42–48.

[27] A. Gandhe, L. Qin, F. Metze, A. Rudnicky, I. Lane, and M. Eck, “Using web text to improve keyword spotting in speech,” in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2013, pp. 428–433.

[28] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, and O. Yilmaz, “Quantifying the value of pronunciation lexicons for keyword search in lowresource languages,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 8560–8564.

[29] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, “Using proxies for OOV keywords in the keyword search task,” in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, Dec. 2013, pp. 416–421.

[30] M. Saraclar *et al.*, “An empirical study of confusion modeling in keyword search for low resource languages,” in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding.*, Dec. 2013, pp. 464–469.

[31] C. Liu, A. Jansen, G. Chen, K. Kintzley, J. Trmal, and S. Khudanpur, “Low-resource open vocabulary keyword search using point process models,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2789–2793.

[32] I. Szöke, M. Skácel, and L. Burget, “BUT QUESST 2014 system description,” in *Proc. CEUR Workshop*, 2014, pp. 1–2.

[33] J. A. Gómez, L.-F. Hurtado, M. Calvo, and E. Sanchis, “ELiRF at mediaeval 2013: Spoken web search task,” in *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18–19, 2013.

[34] C. Gracia, X. Anguera, and X. Binefa, “The CMTECH spoken web search system for mediaeval 2013,” in *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18–19, 2013.

[35] L. J. Rodríguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, “GTTS systems for the SWS task at mediaeval 2013,” in *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18–19, 2013.

[36] H. Wang and T. Lee, “The CUHK spoken web search system for MediaEval 2013,” in *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18–19, 2013.

[37] R. F. A. Alberto Abad and I. Trancoso, “The L2F spoken web search system for MediaEval 2013,” in *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18–19, 2013.

[38] F. G. Igor Szoke, Lukas Burget, and L. Ondel, “BUT SWS 2013—Massive parallel approach,” in *Proc. MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18–19, 2013.

[39] H.-Y. Lee, Y. Zhang, E. Chuangsuwanich, and J. R. Glass, "Graph-based re-ranking using acoustic feature similarity between search results for spoken term detection on low-resource languages," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2479–2483.

[40] H.-Y. Lee, P.-W. Chou, and L.-S. Lee, "Improved open-vocabulary spoken content retrieval with word and subword lattices using acoustic feature similarity," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1045–1065, 2014.

[41] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for OOV terms," in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2009, pp. 404–409.

[42] M. Mueller, "Dynamic timewarping," in *Information Retrieval for Music and Motion*. New York, NY, USA: Springer, 2007.

[43] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Automat. Speech Recognit. Underst.*, 2011, pp. 18–23.

[44] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 215–219.

[45] A. Muscariello, G. Gravier, and F. Bimbot, "Zero-resource audio-only spoken term detection based on a combination of template matching techniques," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 921–924.

[46] S. Laguna, M. Calvo, L.-F. Hurtado, and E. Sanchis, "ELiRF at mediaeval 2015: Query by example search on speech task (QUESST)," in *Proc. MediaEval 2015 Workshop*, 2015, pp. 1–3.

[47] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 5157–5160.

[48] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *Proc. IEEE Workshop Automat. Speech Recognit. Understanding*, 2009, pp. 398–403.

[49] J. Bromley *et al.*, "Signature verification using a "siamese" time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.

[50] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis Pattern Recognit.*, 2005, vol. 1, pp. 539–546.

[51] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 815–823.

[52] L. J. Rodriguez-Fuentes and M. Penagarikano, "MediaEval 2013 spoken web search task: System performance measures," Dept. Elect. Electron., University of the Basque Country, Leioa, Spain, Tech. Rep. TR-2013–1, 2013.

[53] J. Trmal *et al.*, "A keyword search system using open source software," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2014, pp. 530–535.

[54] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Commun.*, vol. 50, no. 5, pp. 434–451, 2008.

[55] Y. Wang and F. Metze, "An in-depth comparison of keyword specific thresholding and sum-to-one score normalization," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2474–2478.

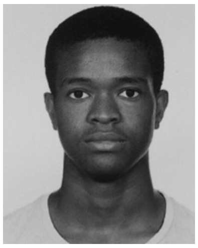
[56] M. Montague and J. A. Aslam, "Relevance score normalization for metasearch," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, 2001, pp. 427–433.

[57] C. Gracia, X. Anguera, and X. Binefa, "Combining temporal and spectral information for query-by-example spoken term detection," in *Proc. IEEE 22nd Eur. Signal Process. Conf.*, 2014, pp. 1487–1491.

[58] I. Szöke, L. Bürget, F. Grézl, J. H. Černockỳ, and L. Ondel, "Calibration and fusion of query-by-example systems—BUT SWS 2013," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 7849–7853.

[59] B. Gündoğdu and M. Saraçlar, "Novel score normalization techniques for posteriorgram based keyword search," in *Proc. IEEE 25th Signal Process. Commun. Appl. Conf.*, 2017, pp. 1–4.

**Batuhan Gündoğdu** received the B.S. degree from the Department of Electrical and Electronics Engineering, Turkish Naval Academy, Istanbul, Turkey, in 2006, the M.S. degree from the Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA, in 2012, where he studied with the special fellowship granted from the Turkish Naval Forces. He is currently working toward the Ph.D. degree in the Department of Electrical and Electronic Engineering, Boğaziçi University, Istanbul, Turkey, and serving as an instructor in the Department of Electrical and Electronics Engineering, National Defense University, Naval Academy. His research interests include machine learning, speech retrieval, and statistical signal processing.



**Bolaji Yusuf** received the B.S. degree from the Department of the Electronics Engineering, Kültür University, Istanbul, Turkey, in 2015. He is currently working toward the M.S. degree in the Department of Electrical and Electronics Engineering, Boğaziçi University, Istanbul, Turkey. His research interests include speech and language processing, particularly cross-lingual and multilingual speech retrieval.

**Murat Saraçlar** received the B.S. degree from the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey, in 1994, the M.S.E. and Ph.D. degrees from the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA, in 1997 and 2001, respectively. From 2000 to 2005, he was with the Multimedia Services Department, AT&T Labs Research. In 2005, he joined the Department of Electrical and Electronic Engineering, Boğaziçi University, Istanbul, Turkey, where he is currently a Full Professor. He was a Visiting Research Scientist at Google Inc, New York, NY, USA (2011–2012) and an Academic Visitor at IBM T.J. Watson Research Center (2012–2013). He has more than 100 publications in journals and conference proceedings. He received the AT&T Labs Research Excellence Award in 2002, the Turkish Academy of Sciences Young Scientist (TUBA-GEBIP) Award in 2009, and the IBM Faculty Award in 2010. He served as an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (2009–2012) and the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2012–2016), was on the editorial board of *Language Resources and Evaluation* (2012–2016) and is currently serving on the editorial board of *Computer Speech and Language*. He is also a member of the IEEE Signal Processing Society Speech and Language Technical Committee (2007–2009, 2015–2018).