DISTANCE METRIC LEARNING FOR POSTERIORGRAM BASED KEYWORD SEARCH

Batuhan Gündoğdu *Murat Saraçlar*

Boğaziçi University
Department of Electrical and Electronics Engineering
34342 Bebek, Istanbul, Turkey

ABSTRACT

In this paper, we propose a neural network based distance metric learning method for a better discrimination in the sequence-matching based keyword search (KWS). In this technique, we conduct a version of Dynamic Time Warping (DTW) based similarity search on the speaker independent posteriorgram space. With this, we aim to compensate for the scarcity of the resources and overcome the out-of-vocabulary (OOV) term problem, which is one of the main issues for KWS on low-resource languages. This distance measure is then used in the DTW-based similarity search, as an alternative and in comparison to the widely and generally used distance metrics. **The experiments ran on IARPA Babel Program's Turkish search data show that, the proposed system outperforms the baseline by 6.3% and when combined with the baseline system, the improvement reaches 44.9%.**

***Index Terms*—** keyword search, spoken term detection, distance metric learning, low-resource languages, dynamic time warping

1. INTRODUCTION

The main focus of this paper is the keyword search task, also called spoken term detection (STD), with a special focus on addressing the problem of retrieving OOV terms. KWS is defined as the task of retrieving the occurrences of a keyword given by the user in text form, in an audio archive. In contrast to keyword spotting [1], in KWS keywords are known only when the user enters a query, possibly resulting in OOV query words. This application has been becoming increasingly interesting and prominent for military, intelligence and civilian usage as the need of obtaining specific parts of an audio archive grow, with the ever-increasing amount of untranscribed digital speech data. KWS can be applied in conference recordings, telephone and radio conversations, audio/video lectures, broadcast news and many other areas with speech data involvement [2, 3, 4]. The general and contemporary approach to KWS is to pass the audio to a Large Vocabulary Speech Recognition (LVCSR) system and to apply index search on the pre-indexed data [5, 6, 7]. However, for low-resource languages, where limited or no transcribed data is available, the LVCSR systems have high word error rates and the performances of the KWS systems that depend on them are limited by imperfections of the LVCSR systems. Furthermore, for these languages, most of the keywords fall out of the vocabulary of the LVCSR systems, for which we can not obtain an LVCSR output resulting in a low KWS performance for such terms.

Hence, in this paper, we will be addressing the OOV and low-resource problem; proposing a methodology to overcome the shortcomings of LVCSR based KWS systems and introducing a KWS system to perform instead of or in combination with them. The work in [8] followed a similar goal to address the same problems using point process models. Our proposed system is based on a sequence-matching based methodology using a version of DTW, called subsequence DTW (sDTW), inspired by query-by-example spoken term detection (QbE-STD) tasks where the keyword is also an audio snippet, by modeling the text query as a posteriorgram to be used in frame level similarity search in the posteriorgram obtained from the audio archive. In [9] and [10] we showed that the pseudo query modeling and posteriorgram based KWS helps improve the LVCSR based system performance and in [11] we analyzed the effects of different query modeling techniques and different distance measures used in frame-level similarity search.

The novelty of this paper is the introduction of a neural-network based distance metric learning (DML) method to be used in frame-level similarity search. We show that, when the new similarity measure learned from a very little data used, the sequence-matching based KWS outperforms the LVCSR based KWS by 6.3% and improves the KWS performances by 44.9% when combined with the LVCSR based system. In Section 2 we introduce the methodology the sequence matching based search and then in Section 3 we provide mathematical background and training of the proposed DML model. In Section 4 we present the experimental results, consisting of the experiments on discriminative power of the new distance model and the results of the KWS experiments conducted using the model.

2. SEQUENCE MATCHING BASED KWS

Sequence matching has been used in early speech recognition systems [12], music retrieval (query by humming) [13] and QbE-STD tasks [14, 15] using versions of DTW algorithms. In this work, we use sDTW in which the keyword, henceforth called the query, is compared with each subsequence of the audio utterance(s) dynamically and the subsequence(s) yielding an average distance value lower than a given threshold is returned as a match. If we call the query, $\mathcal{Q} = \{\mathbf{q}_1, \cdots, \mathbf{q}_M\}$ and the utterance $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$,

This study uses the IARPA Babel Program base period language collection release babel105b-v0.4, supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF- 12-C-0012. 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.

sDTW dynamically finds the optimal alignment path between $\mathcal{Q}$ and the most similar subsequence of $\mathcal{X}$ denoted as $\hat{\Phi}$. Then, the detection score of the search is decided from the accumulated distance through the optimal path using a frame-level distance measure $d(\mathbf{q}, \mathbf{x})$.

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

The flowchart of our sDTW based KWS system proposed in [10] and also used in this work can be seen on Figure 1.

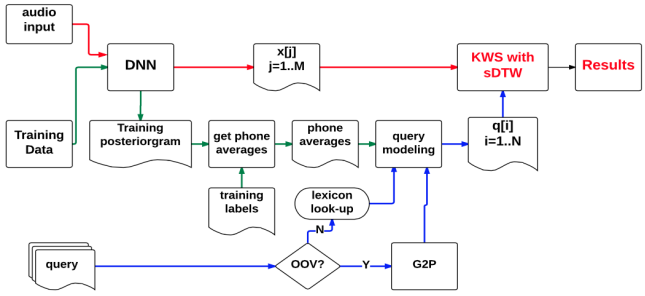

Fig. 1. Flowchart of the sDTW based KWS system

3. DISTANCE METRIC LEARNING

As can be seen in (1), the distance measure used between the frames holds a significant importance on the detection score. Since the phone posterior vectors are used as the representation of frames, the most widely used distance metrics are cosine distance and logarithmic cosine distance [16]. Both of these metrics, along with the euclidean distance, use the inner-product of the two vectors as a similarity measure and apply a kernel to convert it into a distance value. The kernels applied to the inner-product similarity values, for vectors with unit norms, can be seen on Figure 2 for euclidean (2), cosine (3) and logarithmic cosine (4) distance measures.

$$d_{\text{euc}}(\mathbf{x}, \mathbf{y}) = \sqrt{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2 < \mathbf{x}, \mathbf{y} >} \tag{2}$$

$$d_{\cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{< \mathbf{x}, \mathbf{y} >}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{3}$$

$$d_{\text{log-cos}}(\mathbf{x}, \mathbf{y}) = -\log(\frac{< \mathbf{x}, \mathbf{y} >}{\|\mathbf{x}\|\|\mathbf{y}\|}) \tag{4}$$

Although the above mentioned metrics are useful in certain applications and provide mathematically valuable similarity values, none reflect the characteristics of the distribution of the data. One very intuitive solution to this would be to use weighed inner-products with covariances estimated from data, yet the question of which dissimilarity kernel to use still remains. As a better distance value we propose a neural network based model, similar to the siamese networks used in signature [17] and face [18] verification applications.

3.1. DML Neural Network Model

We propose the neural network based DML model in Figure 3, built on the objectives 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 to be in $[0, 1]$ such that the it can be interpreted as the probabilities that two frames belong to the same phoneme or different phonemes. We call the distance obtained from this network *sigma distance* since 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, yet it solely and successfully provides a solution to our above stated objectives.

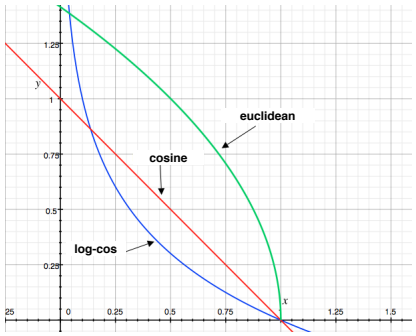

Fig. 2. Comparison on the kernel functions: It can be observed that logarithmic cosine distance metric yields very high distance for low similarities, in other words, it is less tolerant than euclidean to dissimilar samples, i.e pronunciation variations.

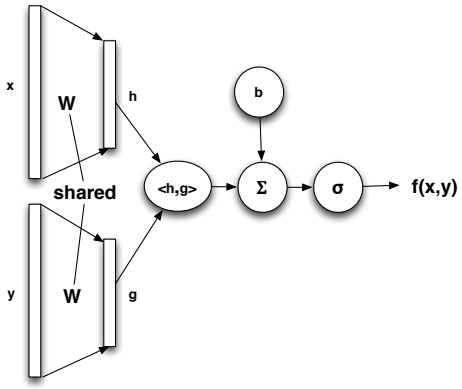

Fig. 3. New distance value $f(\mathbf{x}, \mathbf{y})$ will be the output of this network

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

$$\begin{aligned} \mathbf{h} = \mathbf{W}\mathbf{x} \qquad \mathbf{g} = \mathbf{W}\mathbf{y} \\ f(\mathbf{x}, \mathbf{y}) = \sigma(< \mathbf{h}, \mathbf{g} > + b) \\ d_\sigma(\mathbf{x}, \mathbf{y}) = 1 - f(\mathbf{x}, \mathbf{y}) \end{aligned} \tag{5}$$

where

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{6}$$

3.2. DML Training

As can be seen on Figure 3, the network takes a pair of inputs and emits a scalar distance value. So, we can consider the training set as

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

$$r_t = \begin{cases} 1, & \text{if} \quad \text{class}(\mathbf{x}_t) = \text{class}(\mathbf{y}_t) \\ 0, & \text{if} \quad \text{class}(\mathbf{x}_t) \neq \text{class}(\mathbf{y}_t) \end{cases} \tag{7}$$

Clearly, for a training set of more than two classes, the size of the foes class will be significantly larger than the size of friends. In order to solve this problem, we separated the training dataset into friends and foes. Then in each epoch, we trained the network using the whole friends set in random order and a random subset of foes with the size of the friends set. We also applied prior equalization on the phoneme classes by taking the same number of samples from each phoneme class into the distance learning training.

Since we have the objective of interpreting the system output as a probability value, we used the cross-entropy (CE) objective function in the backpropagation. It can be considered as an objective of increasing the likelihood of friends to output 1 and vice versa. The objective function is then defined as:

$$\begin{aligned} J_{CE}(\mathbf{W}, b; \mathbf{x}_t, \mathbf{y}_t, r_t) = \\ r_t \log(f(\mathbf{x}_t, \mathbf{y}_t)) + (1 - r_t)\log(1 - f(\mathbf{x}_t, \mathbf{y}_t)) \end{aligned} \tag{8}$$

If we express $f(\mathbf{x}_t, \mathbf{y}_t)$ as f for simplicity, the gradient with respect to the parameters are found as follows:

$$\begin{aligned} \triangle b = \frac{dJ}{db} &= \frac{dJ}{df}\frac{df}{dz}\frac{dz}{db} \\ &= (\frac{r - f}{f(1 - f)})(f(1 - f))(1) = r - f \end{aligned} \tag{9}$$

and

$$\begin{aligned} \triangle \mathbf{W} = \frac{dJ}{d\mathbf{W}} &= \frac{dJ}{df}\frac{df}{dz}\frac{dz}{d\mathbf{W}} \\ &= (\frac{r - f}{f(1 - f)})(f(1 - f))(\frac{d}{d\mathbf{W}}(\mathbf{x}^T\mathbf{W}^T\mathbf{W}\mathbf{y})) \\ &= (r - f)\mathbf{W}(\mathbf{x}\mathbf{y}^T + \mathbf{y}\mathbf{x}^T) \end{aligned} \tag{10}$$

where $z = < \mathbf{h}, \mathbf{g} > + b = \mathbf{x}^T\mathbf{W}^T\mathbf{W}\mathbf{y} + b$.

With the gradients in hand and using learning rates μ and η, we update the parameters with online gradient descent as,

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

The flowchart of the DML training is given in Figure 4.

4. EXPERIMENTAL RESULTS

The experiments for this work were conducted on two stages. First, the discriminative power of sigma distance was measured on the trained network and compared with other distance metrics. Then, an sDTW based KWS was run using sigma distance for frame-level similarity check and the results were compared with the baseline and systems using other distance metrics.

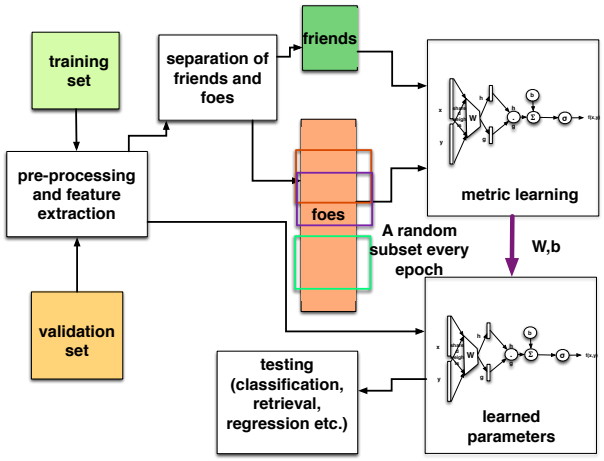

Fig. 4. Flowchart of DML Training

4.1. Datasets and System Set-up

In our system, we used IARPA Babel limited language pack (LimitedLP) Turkish conversational telephone speech data (babel105b-v0.4) [19]. Both the 10-hour training posteriorgram with its alignments and the 10-hour search posteriorgram was obtained using Kaldi Speech Recognition Toolkit [20].

DML : In DML training, only a very little portion of the training alignment is used. We used only 200 random frames per phoneme class, from different speakers. When these frames are combined into DML training pairs, a training set of approximately 36 million pairs was obtained. As a preliminary set-up we used a square weight matrix initialized with a uniform noise added identity matrix. For validation, we used another random subset of the training data, consisting of 100 random samples per phoneme, about 8.8 million pairs.

KWS : The search was conducted on the 88 OOV terms over the 10-hour audio document. The text queries were modeled artificially as posteriorgrams using the average phone posterior-vectors and average phone durations obtained from a subset of the training data. This pseudo-query modeling was explained in [10].

4.2. Sigma Distance Discrimination Performance

To see the dicriminative power of the DML network, we observed the statistics of the distances between friends and foes; and compared them with the common distance metric measures. We have seen that the Gini mean [21] of friends (mean of distances between friends - also called the statistical dispersion) and the Gini mean of foes get farther from each other with sigma distance. In other words, *friends get closer as foes get farther*. This was one of our key goals for achieving discrimination in detection. The normalized histograms of the distances between friends and foes using different distance measures can be seen on Figure 5. We see that euclidean distance and logarithmic cosine distance does not perform well on discrimination of friends and foes. Histogram of cosine distance looks similar to that of sigma distance, however, while in cosine distance the distance histogram of friends looks a flat, it decays as required in sigma distance. The first and second order statistics of the dispersions can be seen on Table 1.

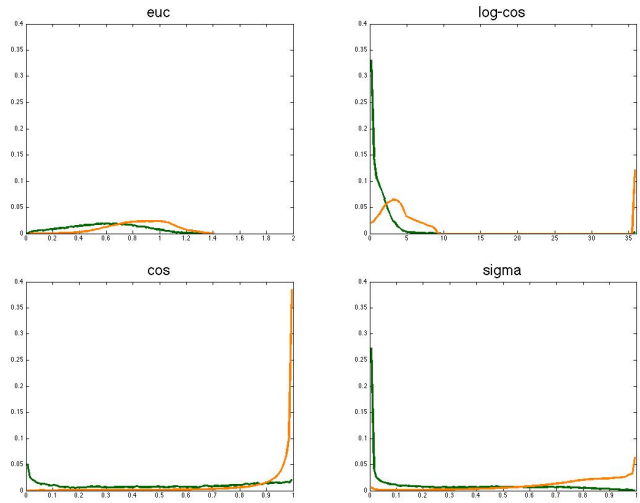

Fig. 5. Dispersion histograms: orange lines denote the histogram of foes and the green lines denote friends.

DISTANCE/SUBSET	friends		foes	
	mean	variance	mean	variance
euclidean	0.5930	0.0685	0.8603	0.0455
log-cos	1.2514	4.4896	7.6956	114.9505
cos	0.5130	0.1122	0.9123	0.0263
initial sigma	0.3313	0.0019	0.3714	0.0001
sigma (converged)	0.2760	0.0849	0.7559	0.0489

Table 1. Dispersion statistics of different distance metrics, initial sigma is calculated with $\mathbf{W} = I$ and $b = -0.5$

4.3. KWS Experiments

4.3.1. Evaluation Metrics and the Baseline

As the evaluation metric, the maximum term weighted value (MTWV) given in (11) is used. MTWV yields a performance score based on a balanced evaluation between correct detections and false alarms [22].

$$\text{MTWV} = 1 - \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} P_{\text{miss}}(q) + \beta P_{\text{FA}}(q) \qquad (11)$$

Using the same threshold for all queries in set $\mathcal{Q}$, $P_{\text{miss}}(q)$ and $P_{\text{FA}}(q)$ denote the probabilities of miss and false alarm for query q using the optimal threshold value. β is a constant deciding the cost and award value between false alarms and correct hits. Given this definition a system returning all queries with no false alarms will yield a MTWV of $\mathbf{1}$. Similarly, a system with no outputs will yield a $\mathbf{0}$ MTWV and hence it is possible to have negative MTWVs for those systems having more false alarms than correct decisions. The baseline system has an MTWV performance of 0.1887 and it uses the LVSCR based Babel KWS setup in Kaldi toolkit [23, 24]. This pipeline uses proxy keywords to handle OOV queries by searching for acoustically similar in-vocabulary (IV) words instead of the OOV keywords. The proxy keyword generation was introduced in [25]. A similar approach based on confusion modeling was also proposed in [26].

4.3.2. Sigma Distance used in KWS

The proposed sigma distance was used in the frame-level discrimination in the sDTW based search. Speech activity detection was applied to the audio posteriorgram using the phonetic posterior values, and silence/non-speech parts having a posterior probability value of 0.5 or more were filtered-out. Also, a keyword specific detection thresholding was applied on the system output by keeping detections holding a score greater than 97% of the best scoring detection for that keyword. It was observed that the common sum-to-one (STO) normalization is more successful after this pruning. The sigma distance was compared with systems using different query modeling techniques (binary vs average) and other commonly used distance metrics (cosine vs logarithmic cosine) that were proposed in [11]. The experiments show that, as an individual system, the system using sigma distance is better than all systems using other distance values and outperforms the baseline by $\mathbf{6.3}\%$. When combined with the baseline system, the improvement reaches $\mathbf{44.9}\%$. The results of individual systems and the merged systems can be seen on Figure 6.

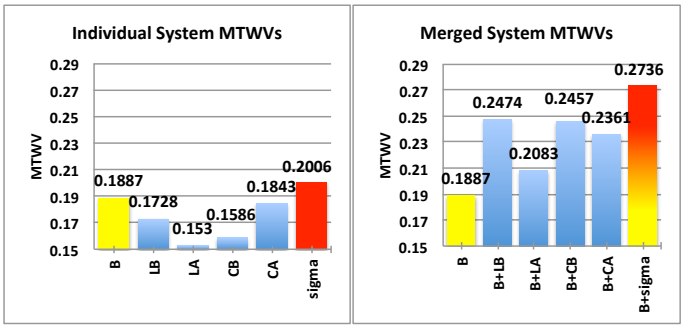

Fig. 6. Individual and merged systems MTWV results, baseline (B), logarithmic cosine (distance)-average query (modeling) (LA),logarithmic cosine-binary query (LB), cosine-average query (CA), cosine-binary query (CB) and the new sigma distance which uses average query.

5. CONCLUSION AND FURTHER WORK

We have seen that the the distance measure learned using the proposed system, discriminates frames better than the common distance metrics and can successfully be used in sDTW based KWS. Not only does the new model provide a more discriminative means of frame similarity comparison and project the phone confusions into the distance value, but also contributes to obtaining a better threshold value to be used in KWS with its interpretable range in $[0, 1]$. It should be noted that in this work, we simulated the low-resource set-up by using very little (84 seconds) of the training data in DML. Learning may get better with more data. The aim of this work was to address the OOV problem and we have seen this methodology improves the LVCSR based system by 44.9% on OOV terms. The proposed system ran on 219 IV terms has a similar performance to the OOV MTWV (0.2066 vs 0.2006), a slight improvement due to the usage of lexicon. As a future work, the deeper models can be trained using additional hidden layers before or after the layer emitting the posteriorgram space, in order to model non-linear dependencies between phonemes, or states. Similar to the work in [27] the distance can be learned on non-linear functions between frames, or even sequences if we use recurrent nets.

6. REFERENCES

[1] G. Chen, C. Parada, and G. Heigold, "Small-footprint keyword spotting using deep neural networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 4087–4091.

[2] C. Chelba, T. J. Hazen, B. Ramabhadran, and M. Saraçlar, "Speech Retrieval," in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, 2011.

[3] M. Larson and G. J F Jones, "Spoken Content Retrieval : A Survey of Techniques and Technologies," vol. 5, no. 2011, pp. 235–422, 2012.

[4] S Parlak and M Saraclar, "Spoken term detection for Turkish broadcast news," *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5244–5247.

[5] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata: application to spoken utterance retrieval," *SpeechIR '04 Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004*, pp. 33–40, 2004.

[6] Murat Saraclar and Richard Sproat, "Lattice-based search for spoken utterance retrieval," *HLT-NAACL 2004: Main Proceedings*, vol. 51, pp. 61801, 2004.

[7] Ciprian Chelba, Timothy J Hazen, and Murat Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 39–49, 2008.

[8] Chunxi Liu, Aren Jansen, Guoguo Chen, Keith Kintzley, Jan Trmal, and Sanjeev Khudanpur, "Low-resource open vocabulary keyword search using point process models," in *2014 15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2789–2793.

[9] L. Sarı, B. Gündoğdu, and M. Saraçlar, "Posteriorgram based approaches in keyword search," in *2015 IEEE 23rd Signal Processing and Communications Applications Conference (SIU)*, pp. 1183–1186.

[10] L. Sarı, B. Gündoğdu, and M. Saraçlar, "Fusion of LVCSR and posteriorgram based keyword search," in *2015 Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

[11] B. Gündoğdu, L. Sarı, G. Çetinkaya, and M. Saraçlar, "Template-based keyword search with pseudo posteriorgrams," in *2016 IEEE 24th Signal Processing and Communication Application Conference (SIU)*, pp. 973–976.

[12] 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.

[13] J-S R. Jang and M-Y Gao, "A query-by-singing system based on dynamic programming," in *Proceedings of international workshop on intelligent system resolutions (8th bellman continuum), Hsinchu*, 2000, pp. 85–89.

[14] C. Chan and L. Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *2010 Eleventh Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

[15] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for OOV terms," in *2009 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 404–409.

[16] Xavier Anguera, Luis J Rodriguez-Fuentes, Igor Szoke, Andi Buzo, and Florian Metze, "Query-by-example spoken term detection evaluation on low-resource languages," in *Proceedings of the International Workshop on Spoken Language Technologies for Underresourced Languages (SLTU)*, vol. 24, p. 31.

[17] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[18] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546.

[19] "OpenKWS14 keyword search evaluation plan," http://www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf.

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

[21] A. Čiginas and D. Pumputis, "Gini's mean difference and variance as measures of finite populations scales," *Lithuanian Mathematical Journal*, vol. 55, no. 3, pp. 312–330, 2015.

[22] J. G. Fiscus, J. Ajot, J. S Garofolo, and G. Doddingtion, "Results of the 2006 spoken term detection evaluation," in *Proceedings of ACM SIGIR Workshop on Searching Spontaneous Conversational*, 2007, pp. 51–55.

[23] J. Trmal, G. Chen, D. Povey, S. Khudanpur, P. Ghahremani, X. Zhang, V. Manohar, C. Liu, A. Jansen, D. Klakow, et al., "A keyword search system using open source software," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, pp. 530–535.

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

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

[26] Murat Saraclar, Abhinav Sethy, Bhuvana Ramabhadran, Lidia Mangu, Jia Cui, Xiaodong Cui, Brian Kingsbury, and Jonathan Mamou, "An empirical study of confusion modeling in keyword search for low resource languages," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Dec 2013*, 2013, pp. 464–469.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.