ZERO RESOURCE GRAPH-BASED CONFIDENCE ESTIMATION FOR OPEN VOCABULARY SPOKEN TERM DETECTION

Atta Norouzian[1], Richard Rose[1], Sina Hamidi Ghalehjegh[1], Aren Jansen[2,3]

[1]Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
[2]Human Language Technology Center of Excellence,
[3]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland

ABSTRACT

In this paper the use of acoustic similarity of speech intervals for generating improved confidence scores for spoken term detection (STD) is investigated. A procedure based on acoustic dotplots which requires no training data is deployed for discovering similar speech intervals. A graph based random walk algorithm incorporates acoustic similarity of hypothesized term occurrences for improving the corresponding confidence scores. The proposed approach is evaluated in an open vocabulary STD task defined on a lecture domain corpus. It is shown that updating the confidence scores in this fashion results in a significant increase in term detection performance of out of vocabulary search terms. A relative improvement of 12.9% in figure of merit was gained relative to that obtained from a baseline lattice based STD system.

Index Terms— Open vocabulary spoken term detection, Dotplot, Random walk on directional graphs

1. INTRODUCTION

There is a wide variety of techniques that are currently being used for spoken term detection (STD) from repositories of spoken audio recordings. The most common approaches rely on lattices generated by a large vocabulary continuous speech recognition (LVCSR) system [1, 2, 3, 4, 5]. In addition, there are many alternative STD approaches in use that require fewer linguistic resources than those required by LVCSR systems [6, 7]. All STD systems produce hypothesized occurrences of query terms in response to queries submitted by users of a search engine. These hypotheses are generally accompanied by a score that provides a measure of confidence that the hypothesized query term corresponds to an actual occurrence of the term in the spoken audio.

The goal of the techniques presented in this paper is to generate a new set of term confidence measures that incorporate a measure of acoustic similarity between hypothesized term occurrences. The similarity measure is obtained from acoustic dotplots described in Section 2. This measure is important since it is generated using a non-parametric system that requires no speech or language resources for training. The updated confidence measures are obtained by forming a directed graph whose vertices correspond to the acoustic intervals containing term hypotheses and whose edges correspond to similarity measures derived from the acoustic dotplots. A random walk on this graph, described in Section 3.2, is performed to update the confidence measures associated with hypothesized terms.

The process of generating term confidence measures presented in this paper is very general in that it can be applied to hypothesized acoustic intervals generated by any STD system. However, its application is particularly appropriate to the case where lattice based STD is applied to search terms that are out of vocabulary (OOV) for the underlying LVCSR system. The experimental study in Section 5 presents the application of the graph based approach to updating confidence measures for detecting OOV search terms in a lecture speech domain. Moreover, in Section 5 the impact of the graph based approach on STD performance for OOV terms will be presented.

This work is related to previous work in using dotplots for measuring acoustic similarity and graph based rescoring of STD hypotheses. Techniques for generating acoustic dotplots have been investigated and applied to term discovery [8, 9] and topic detection [10]. The graph based techniques have also been applied to rescoring STD hypotheses in lattice based STD systems [3]. However, this previous work relied on measures of acoustic similarity derived from the ASR lattices and therefore was only applicable to in vocabulary (IV) search terms. The important aspect of the techniques presented here is that they are implemented completely separate from any STD system and require no training resources whatsoever.

2. DISCOVERING SIMILAR SPEECH INTERVALS

The function of this system component is to discover intervals in the speech signal that are acoustically similar. This process starts by using an energy based voice activity detector to extract a collection of speech segments. Next, a d-dimensional feature vector is extracted from each frame of each segment. For a segment containing o frames this leads to a feature vector time series representation of the form $\{l_1, l_2, \ldots, l_o\}$, where l_i corresponds to the feature vector extracted from frame i. All pairs of speech segments are then searched for acoustically similar intervals using a zero resource word discovery system [8] based on the graphical method of acoustic dotplots.

An acoustic dotplot derived from a feature vector time series $\{l_1, l_2, \ldots, l_o\}$ is an $o \times o$ matrix denoted by $\mathbf{M}$ whose elements $m_{i,j}$ are defined by the cosine similarity between feature vector l_i and l_j:

$$m_{ij} = \frac{1}{2}\left[1 + \frac{\langle l_i, l_j \rangle}{\|l_i\|\|l_j\|}\right]. \quad (1)$$

The similarity measure in Equation 1 takes a value of 1 when l_i and l_j point in the same direction, 0.5 when they are orthogonal, and 0 when they point in opposite directions.

An example dotplot is depicted in Figure 1 where the brightness of the pixels is proportional to the similarity between the corresponding feature vectors as computed in Equation 1. The acoustic feature vectors used for plotting this figure are the standard 39-dimensional perceptional linear prediction (PLP) features. The main diagonal in this figure corresponds to the self similarity of the feature vectors

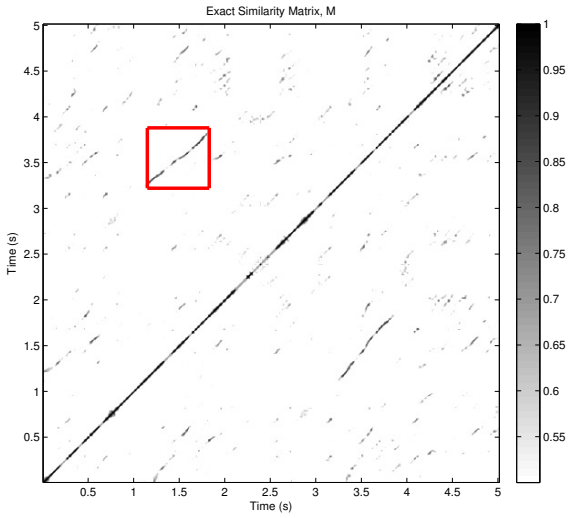

Fig. 1. An example acoustic dotplot obtained from a 5 second long speech utterance.

and any diagonal line segment off the main diagonal represents a repetition of a term (word or phrase). For example the boxed line in the figure corresponds to the repetition of the term *one million dollars*. Having used voice activity detection to restrict our search to speech regions, we prevent the discovery algorithms from returning a flood of silence repetitions. Thus, given our set of speech segments, we need only construct the acoustic dotplot between all pairs and search each for diagonal lines. Constructing the acoustic dotplot is efficiently performed using randomized algorithms described in [8]. The search for approximate diagonal line structures is accomplished using a series of post processing algorithms defined in [11]. Finally, diagonal line matches are further refined using segmental dynamic time warping (SDTW) [9]. The outputs of this system is a collection of similar interval pairs from the speech segments of the form (y_i, y_j), where each interval y_i is represented by its start and end time. The system also generates a similarity measure between each interval pair (y_i, y_j) as $k_{i,j}$ which is derived from match probability from a logistic regression on a collection of generic dotplot features.

3. UPDATING STD CONFIDENCE SCORES

In this section the graph based framework for updating STD confidence scores using the information derived from the zero resource acoustic dotplots is presented. First, the process of constructing a directed graph from the discovered acoustically similar intervals is explained. Next, the random walk algorithm on directed graphs is reviewed and the process of incorporating hypothesized term occurrences and their corresponding confidence scores into the graph is described. Finally, the issues involved in using the random walk for updating term confidence scores are addressed.

3.1. Graph Construction

The output obtained from the zero resource acoustic dotplot in Section 2 is composed of two components, a list of interval pairs, (y_i, y_j), that are acoustically similar and a match probability, $k_{i,j}$, between each pair. Let $G = (V, E, W)$ be a directed graph representing the acoustically similar intervals with a set of vertices, V, connected by directed edges, E, whose corresponding weights are denoted by W. Each interval pair (y_i, y_j) is represented in G with a pair of vertices (v_i, v_j) that are connected by two directed edges $e_{i,j}$ and $e_{j,i}$. While, in general the weights on two edges connecting two vertices of a directed graph can have different values, in our scenario both weights, $w_{i,j}$ and $w_{j,i}$ are set to the match probability $k_{i,j}$ associated with the corresponding vertices and hence are equal. The process of constructing the graph from the similar intervals discovered using the acoustic dotplot is illustrated in Figure 2.

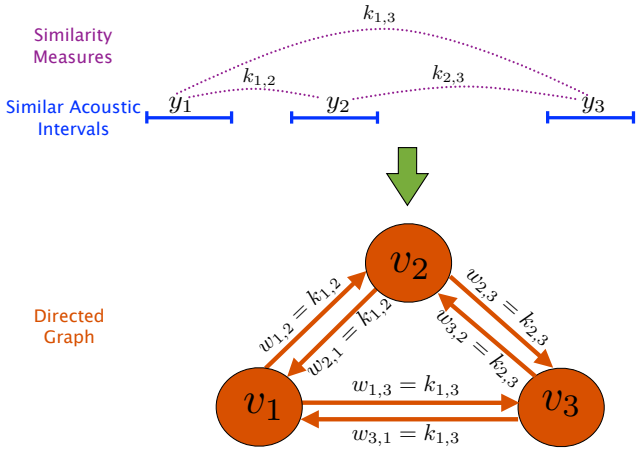

Fig. 2. The process of constructing a directed graph from acoustically similar intervals.

3.2. Random Walk on Directed Graphs

In a random walk process on a directed graph such as G, the system starts from some vertex and at each time step, selects one of the edges directed from the current state with a probability proportional to the weight of the edge, and moves to another vertex. Hence, the probability of being at a given vertex, v_i, at time step t depends on the probability of being at vertices that have edges directed to v_i at the previous time $t-1$. This probability can be computed as

$$p_t(v_i) = \sum_{j \in B_i} \frac{w_{j,i}}{r_j} p_{t-1}(v_j), \tag{2}$$

where B_i is the set of vertices with edges directed to v_i and the normalization factor $r_j = \sum_z w_{z,j}$ is the sum of weights of all the edges directed from v_j. The probability of each of the vertices is updated at each time step according to Equation 2 until it reaches its stationary value. Assuming that the stationary value of all the probabilities is reached at time step π, we have

$$p_\pi(v_i) = \sum_{j \in B_i} \frac{w_{j,i}}{r_j} p_\pi(v_j). \tag{3}$$

Representing the probability distribution of the vertices in a column vector, $\mathbf{p}_\pi = [p_\pi(v_1), p_\pi(v_2), \ldots, p_\pi(v_m)]^T$, and incorporating the normalized weights in an m by m matrix $\mathbf{N}$ with elements $n_{j,i} = \frac{w_{j,i}}{r_j}$, Equation 3 can be rewritten in matrix form as

$$\mathbf{p}_\pi = \mathbf{N}\mathbf{p}_\pi. \tag{4}$$

Based on Equation 4 the stationary probability distribution at time π is the eigenvector of $\mathbf{N}$ whose corresponding eigenvalue is 1.

3.3. Random Walk with Bias

In some scenarios there are sources that provide additional information about the probability distribution for the vertices of the graph. This probability distribution can be incorporated into the random walk process by adding an additional term in Equation 2,

$$p_t(v_i) = \alpha \sum_{j \in B_i} \frac{w_{j,i}}{r_j} p_{t-1}(v_j) + (1-\alpha)c(v_i), \tag{5}$$

where $c(v_i)$ represents the additional probability of the vertex v_i [12]. The additional information in our case is derived from an STD system. For a query term Q, an STD system outputs a list of hypothesized term occurrences accompanied by confidence scores in the form of $(x_Q^1, f_Q^1), (x_Q^2, f_Q^2), \ldots, (x_Q^u, f_Q^u)$, where x_Q^i corresponds to the ith hypothesized occurrence of Q with confidence score f_Q^i. Each hypothesized occurrence of the query term x_Q^i is assigned to a vertex in the graph v_z if there is a substantial overlap between the underlying interval y_z and x_Q^i. In the cases where no such vertex is found in the graph, x_Q^i is assigned a new vertex. Having assigned vertices of the graph to all the hypothesized occurrences, the corresponding confidence scores are normalized and regarded as additional probabilities of the vertices. For the vertex v_z corresponding to hypothesized occurrence x_Q^i, the additional probability $c(v_z)$ is obtained by

$$c(v_z) = \frac{f_i}{\sum_{j=1}^{u} f_j}. \tag{6}$$

Representing the additional probabilities in a column vector $\mathbf{c} = [c_{v_1}, c_{v_2}, \ldots, c_{v_n}]^T$ and using an auxiliary m-dimensional row vector of ones, $\mathbf{e} = [1, 1, \ldots, 1]$, the stationary probability distribution of (4) can be written as

$$\begin{aligned} \mathbf{p}_\pi &= \alpha \mathbf{N}\mathbf{p}_\pi + (1-\alpha)\mathbf{c} \\ &= (\alpha \mathbf{N} + (1-\alpha)\mathbf{c}\ \mathbf{e})\mathbf{p}_\pi = \mathbf{H}\mathbf{p}_\pi. \end{aligned} \tag{7}$$

Equation (7) indicates that the stationary probability distribution $\mathbf{p}_\pi$ is the eigenvector of the m by m matrix $\mathbf{H}$ whose corresponding eigenvalue is 1. The stationary probability of the vertices corresponding to the hypothesized term occurrences are then extracted from the graph and are regarded as new confidence scores. The detection performance obtained using the new confidence scores is evaluated in Section 5 for different threshold values.

4. A LATTICE BASED STD SYSTEM

The lattice based STD system used in this work is based on a hybrid two pass approach [13]. The offline process of configuring this system starts by segmenting the audio repository into short segments and feeding them to an LVCSR system. The LVCSR system in turn generates a word lattice for each audio segment. An efficient indexing technique described in [1] is then applied to the collection of lattices and an inverted index is constructed. In response to a query term typed by a user, search is performed in two passes.

In the first pass search, a subword based approach is deployed for identifying audio segments likely to contain occurrences of the query term using the index. The second pass search begins by producing a phoneme representation for the candidate segments obtained in the first pass. Depending on the query term, two strategies are considered for generating the phonemic representation. For the IV query terms the phonemic representation is obtained by expanding the corresponding word lattices into phone lattices using baseform pronunciations obtained from the ASR lexicon. For the OOV query terms, the 1-best hypotheses of an unconstrained hybrid HMM/NN phone decoder is used as the phonemic representation of the candidate segments. In either case, once a phonemic representation for the candidate segments is obtained, it is matched against the phonemic expansion of the query term Q. A score is then computed for all possible alignments using a phone edit distance [1]. Finally, phoneme sequences with scores higher than a threshold are identified and the corresponding intervals in the audio segments, x_Q^i, and the corresponding confidence scores, f_Q^i, are returned.

5. EXPERIMENTAL STUDY

This section presents an experimental study for evaluating the performance of the confidence score updating technique described in Section 3.2. The study evaluates the ability of this technique to improve the detection performance for OOV query terms produced by the lattice based STD system described in Section 4. The evaluation is performed on a lecture speech task domain which is described in Section 5.1. Section 5.2 describes the configuration of the baseline STD system. In Section 5.3, the detection performance obtained from the original confidence scores generated by the baseline STD system are compared to that obtained using the updated confidence scores.

5.1. Task Domain and Graph Construction

A repository of audio recordings of McGill course lectures available at [14] is used for this study. A number of lectures from this repository were randomly selected and manually transcribed for evaluation and development purposes. The evaluation set consists of two lectures containing 17914 words with a total duration of 131 minutes. An automatic segmentation is performed on the test lectures and segments with an average duration of 2 seconds are generated and acoustic features are extracted from them. The acoustic feature vectors include 12 perceptual linear prediction (PLP) features and an energy feature plus their first and second derivatives amounting to 39 features. After performing mean and variance normalization of the feature vectors they are used for generating the acoustic dotplots. The algorithm for extracting lines in the dotplots is tuned to return similar intervals of about half a second which is a rough estimate of a word duration. The directed graph constructed from these intervals contains 4500 vertices with 83000 edges. The random walk algorithm of Equation 7 was performed on the graph for all the query terms for a range of values of α and best average STD performance was obtained with α equal to 0.6.

5.2. Baseline STD system

The LVCSR system used in the baseline STD system described in Section 4 is a GMM/HMM recognizer originally developed for the AMI project [15]. A series of acoustic and language modeling techniques described in [13] were applied to this recognizer for configuring it for the lecture speech task. A word accuracy of 56.5 and a language model test set perplexity of 143 were measured on the test set. With a vocabulary of 52,800 words, the rate of occurrence of OOV words in the test lectures is a relatively high 11.2 percent.

To evaluate STD performance, a set of 175 query terms were chosen from the most frequent content words in the transcriptions of the test set. The set of query terms consists of 141 IV words and 34 OOV words. The length of the phonemic expansion of the query terms ranges from as few as 2 phonemes for "ear" to 17 phonemes

for the term "phenylpropanolamine". There are a total of 1442 occurrences of query terms in the test set transcriptions out of which 1199 are IV and 243 are OOV occurrences. A more detailed description of the STD system and its performance can be obtained in [1, 13].

5.3. Performance Evaluation

The performance of the STD systems is measured in terms of probability of detection and the number of false alarms per query term per hour. The probability of detection is defined as the number of correctly detected query terms normalized by the total number of actual occurrences of the query term in the test set. A query term detection is considered to be "correct" if its starting time is within 2 seconds of the starting time of the labeled actual occurrence of the term in the test data. Otherwise it is labeled as a false alarm

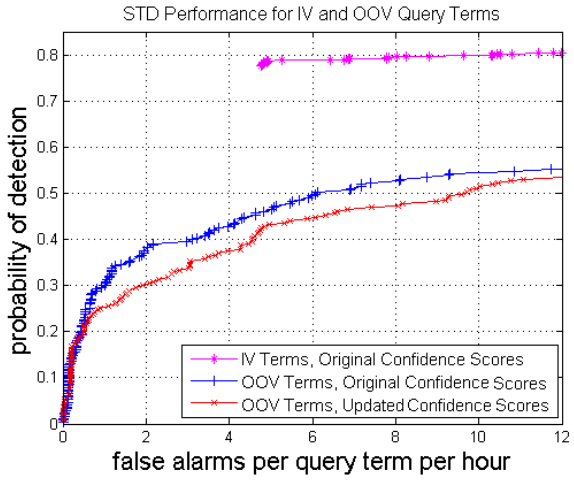

Fig. 3. Detection performance obtained for IV and OOV query terms using original and updated confidence scores.

The data points used for plotting the curves in Figure 3 are obtained by varying the threshold applied to the confidence scores associated with the hypothesized occurrences of the query terms. With a reasonably low threshold, there are in total 149 correctly hypothesized occurrences of the OOV query terms and 1037 correctly hypothesized occurrences of the IV query terms derived from the STD system. Hence, the maximum recall rate achievable for OOV query terms is 61% and for the IV query terms is 86%. The top curve in Figure 3 represents the performance of the lattice based STD system for IV query terms. The two bottom curves correspond to detection performances of OOV query terms obtained using the STD confidence scores and updated scores derived from the graph based procedure in Section 3.2.

Two important observations can be made from this figure. First, despite using subword based algorithms to detect occurrences of the OOV query terms in the lattice based STD system, the detection rate for OOV query terms is still substantially lower than that of the IV query terms. Second, comparing the two curves plotted for the OOV query terms indicates that a significant improvement in detection performance is gained by updating the confidence scores using the information extracted from the zero resource acoustic dotplots. It can be seen that confidence score updating results in a relative improvement of up to 26% in recall rate at 2 false alarms per query term per hour.

For evaluating the detection performance over a range of false alarms a single figure of merit (FOM) can be derived from the plot in Figure 3. The FOM values reported in Table 1 are obtained by averaging the area under the bottom two curves in Figure 3 over a range from 0 to 10 false alarms per query term per hour. Table 1 shows that using the updated confidence scores leads to a relative improvement of 12.9% in FOM compared to the original confidence scores.

Confidence Scores	FOM
Original	38.67%
Updated	43.65%

Table 1. Figure of merit obtained from original and updated confidence scores averaged over all OOV query terms.

6. SUMMARY & CONCLUSION

The use of a graph based technique for exploiting the acoustic similarity of speech intervals for improving the performance of open vocabulary STD systems was investigated here. The similarity of speech intervals was discovered with no training or vocabulary using zero resource acoustic dotplots obtained directly from acoustic signals. The performance of the proposed system was evaluated in a lecture domain task for a set of OOV query terms. It was shown that for the hybrid STD system used in this study, the detection figure of merit for OOV query terms improved by 12.9% when the STD confidence scores were updated by the zero resource system using the graph based approach.

It is important to note that this procedure for updating confidence scores is completely independent from the STD system used to generate the hypothesized term occurrences. Hence, it can be applied without requiring any additional resources to any task domain and any STD system.

7. REFERENCES

[1] A. Norouzian and R. Rose, "An efficient approach for two-stage open vocabulary spoken term detection," in *IEEE Workshop on Spoken Language Technology Proc., 194-199, 2010.* IEEE, 2010, pp. 194–199.

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

[3] Y.N. Chen, C.P. Chen, H.Y. Lee, C.A. Chan, and L.S. Lee, "Improved spoken term detection with graph-based re-ranking in feature space," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5644–5647.

[4] D. Miller, M. Kleber, C. Kao, O. Kimball, T. Colthurst, S.A. Lowe, R.M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *In Proceedings of Interspeech*, 2007, pp. 314–317.

[5] Tsung-Wei Tu, Hung-Yi Lee, and Lin-Shan Lee, "Improved spoken term detection using support vector machines with acoustic and context features from pseudo-relevance feedback," in *Proc. ASRU*, 2011.

[6] Keith Kintzley, Aren Jansen, Kenneth Church, and Hynek Hermansky, "Inverting the point process model for fast phonetic

keyword search," in *International Speech Communication Association*, 2012.

[7] I. Szöke, P. Schwarz, P. Matějka, L. Burget, M. Karafiát, and J. Černockỳ, "Phoneme based acoustics keyword spotting in informal continuous speech," in *Text, Speech and Dialogue*. Springer, 2005, pp. 746–746.

[8] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 401–406.

[9] A. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE T-ASLP*, vol. 16, no. 1, pp. 186–197, 2008.

[10] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 460–470.

[11] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Interspeech*, 2010.

[12] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The pagerank citation ranking: Bringing order to the web.," Technical report, Stanford InfoLab, November 1999.

[13] A. Norouzian and R. Rose, "Facilitating open vocabulary spoken term detection using a multiple pass hybrid search algorithm," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5169–5172.

[14] COOL, "Mcgill online course repository," in *http://cool.mcgill.ca*, 2012.

[15] T. Hain, L. Burget, J. Dines, M. Karafiat, D. van Leeuwen, M. Lincoln, G. Garau, and V. Wan, "The 2007 AMI (DA) system for meeting transcription," in *Proc. NIST RT07 Workshop*, 2008.