A K-nearest neighbours approach to unsupervised spoken term discovery
Alexis Thual, Corentin Dancette, Julien Karadayi, Juan Benjumea, Emmanuel Dupoux

To cite this version:

HAL Id: hal-01947953
https://hal.archives-ouvertes.fr/hal-01947953
Submitted on 7 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
A K-NEAREST NEIGHBOURS APPROACH TO UNSUPERVISED SPOKEN TERM DISCOVERY

Alexis Thual, Corentin Dancette, Julien Karadayi, Juan Benjumea, Emmanuel Dupoux

EHESS, ENS, PSL University, CNRS, INRIA

ABSTRACT

Unsupervised spoken term discovery is the task of finding recurrent acoustic patterns in speech without any annotations. Current approaches consist of two steps: (1) discovering similar patterns in speech, and (2) partitioning those pairs of acoustic tokens using graph clustering methods. We propose a new approach for the first step. Previous systems used various approximation algorithms to make the search tractable on large amounts of data. Our approach is based on an optimized k-nearest neighbours (KNN) search coupled with a fixed word embedding algorithm. The results show that the KNN algorithm is robust across languages, consistently outperforms the DTW-based baseline, and is competitive with current state-of-the-art spoken term discovery systems.

Index Terms— spoken term discovery, word discovery, unsupervised, word segmentation

1. INTRODUCTION

Current automatic speech recognition pipelines require considerable textual resources to train acoustic and language models. Several approaches have been pursued to address the challenge of building useful technology for languages where few textual resources are available. The “zero-resource” setting [1] tackles the most extreme case, where no textual resource is available at all, and aims at extracting useful linguistic units from raw speech in an unsupervised way. Spoken term discovery focuses on discovering word-sized units; it can be used for document classification [2], segmentation [3] or retrieval [4, 5], and idiolect recovery [6]. It can also help other zero resource objectives, such as learning phonetic unit or representations [7, 8, 9].

Searching for all matching motifs in a speech corpus scales quadratically and can be impractical for large corpora. Here, we separate spoken term discovery into two subproblems: representing candidate motifs as a fixed-sized vector, and similarity-based search. For the latter, we use a high performance k-nearest neighbours library developed by Facebook [10], which can scale to billion of search terms. We test our systems on datasets from the Zero Resource Challenges [11, 12], and compare it to the state-of-the-art.

2. RELATED WORK

Spoken term discovery systems can be sorted into two classes. The first class attempts to discover recurring motifs in the speech signal by using a DTW-based distance metric over motifs, and clustering these motifs to form a lexicon [13, 14, 15, 16, 17, 18]. The focus of these algorithms is not to discover all of the words, but to discover a subset of good candidates. This is why they typically set a lower bound on the duration of the motifs, in order to avoid short words, which tend to have many neighbors and be hypo-articulated. The second class attempts to exhaustively parse the input signal into words, thereby leveraging the additional constraints that word candidates cannot occupy the same portion of speech. These algorithms both segment the speech and construct a lexicon at the same time [19, 20]. They are variants of text-based algorithms based on the idea of minimizing the length (or maximizing the probability) of the corpus and its accompanying lexicon [21, 22, 23].

In the present work, we aim at adding a new element to the first class, but simplify the pipeline by getting rid of the DTW step, which requires to compute a very large matrix of frame-wise similarity of the corpus by itself. The system proposed by Jansen and van Durme [16], which was used as baseline of the Zerospeech Challenge 2015 [11], proposes to avoid computing the whole matrix by binarizing it using random projections, an indexation technique yielding a large gain in speed, but a decrement in performance due to the approximations performed by the algorithm [16]. Here, we propose to use another indexation tool, based on product quantization [24], and specifically the FAISS library which can perform efficient exact or approximate search on CPUs or GPUs [10].

3. DESCRIPTION OF THE PIPELINE

Our pipeline is presented in Figure 1. After extracting speech features, we segment the input speech and construct a fixed-length representation of each segmented term. We then use a KNN algorithm to systematically find a set of matching pairs, which we narrow down by a variant of Non Maximal Suppression to avoid overlap. This is followed by an optional step of graph clustering, before the output is evaluated.
works for deriving such embeddings [26].

Here, we use a simple Gaussian smoothed down-sampling method, typically used in [20] and studied in [26]. The idea is to represent each segmented term using only \(d\) frames: in order to do so, one averages adjacent frames using a gaussian kernel. Let \(f_i\) be the \(i\)-th frame of the output, then

\[
f_i = N \left( \frac{i \cdot n_f}{d}; r \cdot n_f + s \cdot g_d(i) \right) \cdot t
\]

where \(g_d : i \rightarrow \frac{d}{2} - \frac{d - i}{\alpha}, t\) is the feature representation of the term being down-sampled, \(n_f\) is the number of frames of \(t\), \(r\) and \(s\) two parameters which we here set to 0.07 and 0.1 respectively, as recommended in [26]. We set the value of \(d\) to 20 for all experiments reported in this paper. We had tested several values of \(d\) (from 10 to 30), \(r\) (from 0 to 1) and \(s\) (from 0 to 1), but none of these changes had strong positive or negative impact on our results.

3.4. KNN search

The pre-segmented embedded terms are stored in a large index. Then, for each term, we search the entire index for its \(k\) nearest neighbours using FAISS [10], where the similarity is based on the dot product of the embedded terms. In our pipeline, we use an exact search method, parallelized on 10 CPUs, with \(k\) set to 200. Searching the index against itself takes 98% of the overall run-time.

We tested several values for \(k\), ranging from 50 to 500. The idea is using a reasonably large value for \(k\) (instead of \(k = 1\)) is to take into account self-overlapping pairs which may also tend to be very self-similar. In addition, if a good match is found between two different parts of the corpus, several overlapping matches will also have a good similarity match. Therefore our strategy was to allow a large number of matches to be returned, and prune this number down in a selection step. Experiments showed that all values above 100 yield the same results for a dataset of the size of a few hours. We set the value of \(k\) to 200 in order to handle larger datasets while keeping the run-time reasonable.

3.5. Selection of top pairs

Generating \(T \times k\) pairs, where \(T\) is the number of previously segmented terms, obviously yields too many pairs, most of which are of poor quality. In addition, as we said above, many pairs are redundant and partially overlapping. In order to reduce the number of pairs, we perform the following 4 steps for each of the input speech files:

1. Sorting: we recover the totality of the KNNs for each segmented term an input file, and sort the resulting pairs of terms by decreasing similarity.
2. Thresholding: we keep only the top $\delta \%$ of the pairs for a given file; in this paper, $\delta$ is the main control parameter and manipulated systematically (typically between 0.5% and 3%).

3. Self-overlapping pairs removal: given a pair $p = (t_1, t_2)$, we discard $p$ if $t_1$ and $t_2$ overlap more than a given threshold $\tau_{self}$ where $\tau_{self}$ equals 50% in our experiments. This selection phase is very efficient and discards a large proportion of the pairs (around 90%).

4. Non Maximal Suppression (NMS): to remove the remaining overlapping pairs, we perform a variant of NMS by sequentially inspecting the pairs starting from the most similar ones. We only keep a new pair if it is not overlapping with an already selected pair. We use a threshold $\tau_{other}$ of 50%. NMS helps discarding around 50% of the remaining pairs. Practically, we distribute the selected pairs into slots, one for each by pair of utterances in the original corpus (defined as contiguous stretches of speech) and perform NMS within each slot only. This allows for faster computations in our implementation.

In this paper, we define temporal overlap between two pairs of terms $p = (t_1, t_2)$ and $p' = (t'_1, t'_2)$ as follows:

$$o(p, p') = \frac{l(t_1 \cap t'_1) \cdot l(t_2 \cap t'_2)}{l(t_1) \cdot l(t_2)}$$

where $l$ yields the length of a given term and $t \cap t'$ designates the overlapping part of terms $t$ and $t'$. It is important to notice that $o(p, p')$ is not necessarily equal to $o(p', p)$ and that the order in which pairs are seen during the NMS actually has an impact. Self-overlapping of a given pair $p = (t_1, t_2)$, with $t_1 = [a_1, b_1]$ and $t_2 = [a_2, b_2]$, is computed using the following formula:

$$so(p) = \frac{\min(b_1, b_2) - \max(a_1, a_2)}{\min(l(t_1), l(t_2))}$$

3.6. Clustering

Once pairs of word-units have been selected, one can perform a clustering step in order to group those pairs together. The goal is that each cluster would account for one word-unit of the discovered lexicon, and that each member of a given cluster represents one occurrence of the aforementioned word-unit. Several clustering methods were documented [17]. We did not try to use them here and rather focused on the quality of the output pairs of our pipeline. This can be interpreted as if each pair was a cluster on its own.

4. EXPERIMENT I

The aim of this Experiment is to test systematically the effect of two important parameters in the KNN pipeline described above: $(l_{min}, l_{max})$ the range of durations selected in the presegmentation step, and $\delta$, the percentage of pairs that are kept after the KNN search. The range of duration is important because too long terms can lead to useless computations, while too short terms could give rise to very noisy results. To choose the range of durations to study, we had a peek of the distribution of word durations across the languages of the Zero Resource Challenges, and chose the following partition of durations: 100-200ms, 200-300ms, 300-500ms, 500-1000ms, each of which approximately accounts for 25% of the word tokens across these languages. We also used 200-500ms, which covers the middle range, and about 50% of the word tokens. The $\delta$ threshold commands a tradeoff between the number of pairs and the quality of these pairs, and we vary this threshold systematically to explore the whole range of this tradeoff.

This is conducted in the Mandarin dataset of the Zero Ressource Speech Challenge 2017 [12], which we considered to be our development dataset to setup and test our KNN pipeline. Previously mentioned parameters remained fixed to the values described in the pipeline description.

4.1. Dataset

The dataset, available in the Zero Resource Speech Challenge 2017, includes 12 Mandarin speakers, each of whom speaks for a duration from 10 to 25 minutes. The overall duration of this dataset is 2 hours and 30 minutes. It has been forced aligned to enable phoneme-based evaluation metrics.

4.2. Evaluation

To study the effect of duration and $\delta$, we only focus on the subset of the Zero Resource Speech Challenge 2015 metric devoted to measuring the quality of the similarity-based search, namely:

- Normalized Levenshtein Distance (NED): given a pair $p = (t_1, t_2)$, one can read the phonetic transcriptions of $t_1$ and $t_2$ using the gold temporal alignments; the normalized Levenshtein distance between these two transcriptions yields the NED for this pair
- Coverage (COV): it indicates the proportion of input speech which was discovered; for instance, a coverage of 0.5 indicates that half the input phones were involved in computed pairs of terms
- Number of pairs: simply the number of pairs of terms yielded by the pipeline

4.3. Results and discussion

Results are reported in Figure 2. They confirm the expected effects of duration range on the tradeoff between number and quality of the pairs. Articulation for short words is on average worse than it is for longer terms. This is indeed what we
observe here: for a small COV or number of pairs, longer terms yield much better results that short ones. However, for larger COV and number of pairs, short terms yield a much better NED than long ones. This is because the frequency of a long term is usually low, thus, when a high number of pairs is required (by setting δ to a permissive value for instance), our structure pairs long words with other long words and the chances of these being different occurrences of the same word get lower. On the contrary, small words’ frequency is much higher (words lasting between 100ms and 200ms are almost syllables) and it is still plausible to find theoretically good pairs even when δ is very permissive.

As expected, results for terms of length ranging between 200ms and 500ms approximately behave like terms of length 300-500ms for low COV and number of pairs, and roughly follow terms of length 200-300ms for high COV and number of pairs.

Wider ranges allows for better results but only up to a point. A range between 200ms to 1000ms, for instance, (not shown in the Figure) yields results which are almost comparable to 200-500ms for all COV values, with much longer computation time. This is why in the following, we stick to the 200-500ms range.

5. EXPERIMENT II

The aim of this experiment is to compare our KNN pipeline to the Zero Ressource 2015 DTW-based baseline. To do this, we picked our best system as evaluated in the Mandarin dataset in Experiment 1 and applied it without changing any parameter to the two datasets of the ZR15 Challenge: English and Xitsonga.

We first run the same evaluation as in Experiment 1 on the duration range of 200-500ms, systematically varying the threshold parameters of both algorithms in order to compare their quality/quantity tradeoffs. Next, we pick a value for the threshold parameter δ of the KNN system so that the NED is comparable to that of the DTW-baseline and points with higher NED. We then evaluate the results using the standard metrics of the challenge and compare them to state-of-the-art implementations.

5.1. Datasets

The English dataset comes from the Buckeye corpus and includes 12 speakers speaking between 16 and 30 minutes each, for a total duration of 5 hours. The Xitsonga corpus includes 24 speakers speaking between 2 and 30 minutes each, for a total duration of 2 hours and 30 minutes.

5.2. Evaluation

We use the full set of evaluation metrics of the second Track of the Zero Resource Challenge 2017. This includes Ned and Cov (already described above), and a set of 4 precision/recall/F-score metrics each of them dedicated to a particular aspect of spoken term discovery: Grouping, Type, Token and Boundary. The Grouping precision and recall metrics are similar to cluster purity and collocation scores, respectively, except they are computed on a pair by pair basis and do not require a majority decision (each pair is labelled as belonging to the ‘same cluster’ or ‘different cluster’ set). The Type metrics compare the discovered clusters to the gold lexicon (also pair-wise). The token metrics compare the set of discovered terms with the set of gold tokens, and the boundary metrics compare the discovered and gold boundaries. The Fscores are the harmonic means of precision and recall.

5.3. Results and discussion

Figure 3 shows the NED / COV and NED / Number of pairs graphics for the DTW baseline system and the KNN system on the two test languages of the Zero Ressource Challenge
We present a new pipeline for spoken term discovery which bypasses the need of performing DTW on the speech data, and rather uses a performant KNN library on fixed-words embeddings. The results, optimized on Mandarin, transfer to two new languages, and beat a DTW-based baseline that had been optimized for these two languages.

The datasets that were used here were small enough that we could use the exact search mode of the KNN library. Using 10 CPUs, our searches run approximately in real time (2 to 3 hours for Mandarin - 5 hours for the Buckeye) and necessitate the storage of an index of 1 to 2 GB depending on the dataset’s size. For larger datasets, we would have to resort to the approximate mode and/or the parallelized GPU mode.
of the KNN library. What’s important, though, is that because we have separated the problem of representing the speech motifs from the problem of search, the scaling up problems can be addressed by generic systems without having to make particular tweaks about speech.

The pre-segmentation and representation steps are of course critical for the quality and efficiency of the algorithm. Further work would need to explore other pre-segmentation ideas, in particular the idea of using syllabic boundaries as in Rasinen et al.[28]. If this turns out to be reliable, we could expect a speed up factor of about 10-20. As for the representation, down-sampling is not the most efficient way to represent spoken terms. Other work has shown that recurrent NNs can achieve even better results, on some occasions overtaking DTW methods [26] and would need to be tested in this framework.

7. ACKNOWLEDGMENTS

This work was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL* ), Almerys (industrial chair Data Science and Security), Microsoft Research (joint MSR-INRIA center), Amazon Web Service (AWS Research Credits) and a Google Faculty Award.

8. REFERENCES


[6] Aren Jansen, Daniel Garcia-Romero, Pascal Clark, and
Jaime Hernandez-Cordero, “Unsupervised idiolect discovery for speaker recognition.”


