UNSUPERVISED IDIOLECT DISCOVERY FOR SPEAKER RECOGNITION

Aren Jansen[1], *Daniel Garcia-Romero*[1], *Pascal Clark*[1], *Jaime Hernandez-Cordero*[2]

[1]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, MD USA
[2]U.S. Department of Defense

ABSTRACT

Short-time spectral characterizations of the human voice have proven to be the most dependable features available to modern speaker recognition systems. However, it is well-known that high-level linguistic information such as word usage and pronunciation patterns can provide complementary discriminative power. In an automatic setting, the availability of these idiolectal cues is dependent on access to a word or phonetic tokenizer, ideally in the given language and domain. In this paper, we propose a novel approach to speaker recognition that leverages recently developed zero-resource term discovery algorithms to identify speaker-characteristic lexical and phrasal acoustic patterns without the need for any supervised speech recognition tools. We use the enrollment audio itself to score each trial and perform no model training (supervised or unsupervised) at any stage of the processing, allowing immediate application to any language or domain. We evaluate our approach on the extended 8-conversation core condition of the 2010 NIST SRE and demonstrate a 16% relative (0.06 absolute) reduction in minDCF when combined with a state-of-the-art unsupervised i-vector cosine system.

Index Terms— Zero resource, unsupervised term discovery, speaker recognition, idiolect

1. INTRODUCTION

Current state-of-the-art speaker recognition systems rely on characterizations of the short-time spectral content of the speech signal, where clustering and averaging at the frame level limits the interference of conversation specific linguistic content. These methods, the most successful being recent i-vector techniques [1], have dominated evaluations in large part due to their ability to function with very limited enrollment and test audio. However, robustness to noise and channel effects remains a major challenge [2], leaving an opening for alternative high-level features to play a continued role.

In addition to a characteristic spectral distribution, speakers exhibit characteristic language usage patterns. Commonly known as *idiolect*, these patterns include those parts of an individual's vocabulary, grammar, and pronunciation not commonly shared by other speakers of their language and dialect. While the prominence of these high-level linguistic cues varies by individual, they can become distinctive for some when sufficient enrollment data is provided. This notion was first introduced to the speaker recognition community by Doddington [3] in an effort to capture characteristic word usage and grammatical patterns using n-gram features derived from manual transcriptions of the Switchboard corpus. This was later extended to automatic transcriptions and was demonstrated to provide competent standalone performance (when several minutes of enrollment audio is available) and complementarity with short-time spectral systems [4, 5, 6]. However, the dependence of these methods on in-domain and in-language speech recognition tools greatly limits the versatility and scalability of the technology.

Even when available, speaker-independent speech recognizers are trained to discard speaker-specific pronunciation patterns of the words and phrases present. To capture these aspects of one's idiolect, past efforts have considered prosodic features [7], subword unit n-grams generated by ensembles of phonetic recognizers (known as parallel phonetic recognition followed by language modeling, or PPRLM) [8, 9], explicit pronunciation modeling [4, 10], and hidden Markov models (HMM) of words shared between the enrollment and test audio [11]. Each of these approaches have been demonstrated to provide useful signal for speaker recognition, but they too have weaknesses. Prosodic features are used to augment traditional acoustic feature spaces and are not designed to support a system in their own right. While the PPRLM and HMM approaches can perform well on their own, like the word n-gram approach both require access to supervised speech recognition tools.

In this paper, we propose a scalable, fully unsupervised approach to exploiting idiolect for speaker recognition that removes any dependency on supervised speech recognition tools. In their place, we apply efficient spoken term discovery algorithms [12], which identify repeating acoustic feature trajectories that typically correspond to words and phrases. These automatically discovered units, which we call *pseudoterms*, have been demonstrated to be a suitable proxy for words in vector space document models [13]. By choosing a *speaker-dependent* acoustic front-end, we can discriminate based on both the words and phrases speakers tend to say as well as *how* they say them. Trials are scored using normalized inner products of the pseudoterm frequency distributions of the enrollment and test audio. Unlike similarly motivated past approaches [14, 15, 16], we do not require predefined word segmentations or subword/word models. Using the 2010 NIST speaker recognition evaluation data, we measure standalone performance comparable to the best idiolect-based approaches reported in the literature and demonstrate significant complementarity with state-of-the-art i-vector technology.

2. SYSTEM ARCHITECTURE

In traditional word-based idiolect approaches for speaker recognition, a spoken document is characterized by its n-gram frequency distribution relative to the background distribution estimated from a large multispeaker corpus. In our proposed approach, we replace word n-grams with the notion of pseudoterms, defined in [13] to be any acoustic pattern of at least approximately 0.5 s in duration that is repeated in a given set of speech documents. Pseudoterms most often correspond to words and phrases, but may also be silence regions, audio anomalies, or filled pauses. By using speaker dependent acoustic representations, the "vocabulary" of pseudoterms is roughly equivalent to the Cartesian product of the set of word n-grams and

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

the set of speakers. Our goal is to develop a measure that characterizes pseudoterm frequency distributional similarity between cuts.

In the traditional speaker recognition evaluation paradigm, the goal is to provide confidence scores for a set of trials. Each trial involves scoring a single test cut against a given known enrollment speaker for which we have been provided one or more enrollment cuts. In addition, for purposes of score normalization and additional non-speech filtering, we assume we have access to two sets of unlabeled development cuts that do not involve the enrollment speakers. Our proposed approach to generating normalized trial confidence scores consists of four steps: (i) use the unsupervised term discovery system to compare the enrollment cuts against a development set to identify non-speaker-specific pseudoterms to be filtered from the enrollment data; (ii) use the unsupervised term discovery system to identify pseudoterms shared between the filtered enrollment cuts and the second development set to extract score normalization statistics; (iii) use the unsupervised term discovery system to identify pseudoterms shared between the filtered enrollment cuts and the test cuts and use them to generate trial scores; and (iv) apply per enrollment speaker score normalization for pooled evaluation metrics. Each of the involved subcomponents are described below.

2.1. Efficient Unsupervised Term Discovery

We use the scalable term discovery system presented in [12] to drive all stages of system processing. Given its central role in our speaker recognition system, we include an abbreviated summary of the discovery system architecture. We begin with two audio cuts (e.g. an enrollment cut and a test cut) for which we extract two acoustic feature vector times series (e.g. PLP or MFCC) denoted $X = x_1 x_2 \dots x_n$ and $Y = y_1 y_2 \dots y_m$, where each $x_i \in \mathbb{R}^d$ and $y_j \in \mathbb{R}^d$. The goal is to identify all high similarity segments shared between X and Y without having to resort to brute force $O(nm)$ search presented in [17]. The scalable approach uses a collection of approximation techniques to accomplish the search in nearly linear time, as described below.

Before we can identify repeated segments, we must efficiently identify individual nearby frames both X and Y. To do this we apply an approximate nearest neighbor search method based on two randomized algorithms: locality sensitive hashing (LSH) and point location in equal balls (PLEB) [18]. LSH is a randomized hashing technique that maps points $x \in \mathbb{R}^d$ to bit signatures $h(x) \in \{0,1\}^b$ that preserve the ability to approximate distances in the original vector space. We consider the variant of LSH that preserves cosine distance, which is accomplished by projecting x onto b random d-vectors and thresholding at zero, each producing a bit that encodes membership in a randomly oriented halfspace. Denoting Hamming distance between bit signatures by $H(\cdot,\cdot)$, then the cosine distance between $x_i, y_j \in \mathbb{R}^d$ can be approximated by $1 - \cos\left(H(h(x_i), h(y_j))\pi/b\right)$ with an approximation error that approaches zero as $b \to \infty$.

The PLEB algorithm constrains nearest neighbor search space by using lexicographic sorting (i.e. alphabetical order if treated as strings) of LSH signatures. This sort ordering has the interesting property that adjacent signatures share a common prefix of bits, which implies a bound on the Hamming distance and, in turn, the cosine distance. Thus, we need only check a constant beam width of B nearby points in the list against a prescribed distance threshold δ for inclusion in the sparse approximate distance matrix. Since the lexicographic sort gives preference to the initial bits, we must permute the signatures P times, each time resorting and rescanning. In the present case, for each permutation we sort the LSH signatures of X and Y separately. Then, for each signature h in the sorted list for X, we consider B points centered around the would-be insertion position of h in the sorted list for Y. While exhaustive cosine distance matrix computation requires $O(nm)$ time, PLEB for P permutations requires only $O(Pm \log m + Pn \log n)$ sorting operations and $O(PBn)$ neighbor comparisons. For the problem sizes involved in the present system, the linear-time neighbor comparisons dominate.

Given the sparse (approximate) neighbor distance matrix, M, computed using LSH and PLEB above, the next step is to efficiently search for runs of frame-level matches between X and Y. We employ a two-pass coarse-to-fine strategy. The first pass employs a collection of image processing techniques to M. First, we transform the matrix to binary form M', where $M'_{ij} = 1$ if $M_{ij} < \delta$ and 0 otherwise. Second, we apply and diagonal median filter of width 50 frames to M', which imposes a minimum match duration and removes noise introduced by LSH. Finally, we using a Hough transform to identify remaining diagonal line segments. The second pass involves using the center points of these diagonal line segments as starting points for local segmental dynamic time warping (DTW) search. Complete details are provided in [19].

The main challenge for the term discovery technology on its original intended application has been identifying word and phrase repetitions *across speaker*. However, when applied to speaker recognition, we can turn this difficulty into a computational advantage. First, we are now only interested in very similar within-speaker word repetitions, meaning we can safely reduce PLEB beam width B without substantial loss in discovery recall (in the experiments below, two 5 minute cuts can be compared in less than 1 second). Second, we can reduce the cosine distance threshold δ for inclusion in the distance matrix, which reduces the computational burden on the downstream two-pass segment search.

2.2. Filtering Non-discriminative Audio

Every acoustic pattern from an enrollment cut that matches across speaker can potentially increase non-target trial scores. The primary suspects for this failure mode are silence regions missed by the speech activity system, anomalous audio events (e.g. tones), and filled pauses. If provided an unlabeled collection of out-of-set development cuts, we can easily identify these error-inducing segments of the enrollment audio using the above-described term discovery algorithm. We consider any discovered segments as additional non-speech regions and remove them from the speech activity marks for all subsequent processing stages. While this process is completely unsupervised, it could be easily swapped for supervised variants that attempt to cluster pseudoterms and discriminatively identify those that appear in a speaker's enrollment speech and nowhere else. We leave exploration of supervised techniques for future work.

2.3. Confidence Scoring and Normalization

After filtering the non-discriminative portions of the enrollment cuts according to the above procedure, we can proceed to score each trial using the term discovery system. All enrollment cuts for a given speaker are implicitly merged and taken as X, while the test cut is taken as Y. The above-described term discovery system applied to X and Y produces a collection of K enrollment-test repetitions, each characterized by a pair of intervals and their DTW alignment cost. We first apply a DTW threshold $\mathcal{T}$ with the goal of limiting the pseudoterms considered to those arising from high confidence matches likely to involve the same speaker.

A desirable score should capture the similarity between the pseudoterm frequency distributions $F_X \in \mathbb{N}^w$ and $F_Y \in \mathbb{N}^w$ for the enrollment speech and the test cut, respectively, where w is the cardinality of the speaker-specific pseudoterm vocabulary. However, to explicitly estimate these distributions for a given audio file, we would have to proceed to cluster the individual repetitions into discrete pseudoterm categories as done in [13]. To avoid this, we notice that if a given pseudoterm occurs n_1 times in document X and n_2 times in document Y, the discovery system will generate $n_1 \cdot n_2$ matches. It follows that the total match count K is itself an estimate of the inner product between F_X and F_Y that we define as our raw trial score. In lieu of L^1 or L^2 normalization common in bags-of-words applications, which would again require explicit pseudoterm clustering, we instead divide each trial score by the logarithms of the enrollment and test speech durations.

The proposed count-based trial scores above approximately follow a zero-inflated exponential distribution with substantially shorter decay length for non-target trials than for target trials. To enable measurement of pooled performance metrics, we perform a Z-norm analogue for each enrollment speaker as follows. First, we generate a set of count-based scores for the speaker's enrollment cuts against a moderately-sized collection of unlabeled out-of-set development cuts. Next, we fit a shifted exponential distribution of form $p(s) = \exp(-(s - s_0)/\tau)$ to the non-zero development set scores only, where τ is the decay length and $s_0 > 0$ is an offset of the distribution away from zero. Each test trial score s is then normalized according to $(s - s_0)/\tau$.

3. EXPERIMENTS

We evaluated our pseudoterm-based speaker recognition system on the 8 conversation training condition and extended core test condition of the 2010 NIST Speaker Recognition Evaluation.[1] This involves over 687K trials (442 target) involving 267 enrollment speakers. We used PLP features as input to the term discovery system (13 cepstral coefficients, plus velocity and acceleration, with 25 ms window size and 100 Hz frame rate). From past success on a separate speaker recognition evaluation, we used $b = 64$ bit LSH signatures, a beam width of $B = 5$, $P = 8$ bit ordering permutations, a cosine distance threshold of $\delta = 0.5$, and DTW alignment cost threshold of $\mathcal{T} = 0.15$. We used two unlabeled development sets of size 2000 and 1860 conversation sides from previous SREs to perform the filtering described in Section 2.2 and to estimate the score normalization statistics as described in Section 2.3, respectively. Before applying term discovery, we filtered non-speech audio using the speech activity system described below in Section 3.1. We compared against a state-of-the-art i-vector system, using both unsupervised cosine scoring and supervised probabilistic linear discriminative analysis (PLDA) scoring, as described below in Section 3.2.

3.1. Speech Activity System

As with any speaker verification system, our term discovery approach benefits from an initial screening that removes segments of non-speech that would otherwise produce deluge of uninformative acoustic matches. Toward this end, we designed a new, robust speech detection system based on modulation spectral analysis. Briefly, our method estimates the power spectrum of individual cepstral coefficients over a sliding window 500 milliseconds long. The spectral centroid [20] and the proportion of energy in the 2-8 Hz modulation frequency band [21] represent two reductive yet discriminative features for identifying the presence of syllabic dynamics associated with speech. For adaptability to variable acoustics, we estimated a speech/non-speech decision threshold file-by-file using k-means clustering in the centroid-syllabic features, with k equal to 2.

[1] Our original effort was on NIST SRE 2012. Due to our system's impeccable ability to identify duplicate cuts as gigantic pseudoterms, we discovered that LDC constructed test cuts by splicing segments of enrollment data. As a result, we decided to revert the SRE 2010 evaluation.

Table 1. 2010 NIST SRE performance of the proposed and baseline systems for the 8 conversation extended trial set. The subscript of minDCF indicates the prior on target trials used in the cost function.

System	EER (%)	minDCF$_{0.01}$	minDCF$_{0.001}$
Pseudoterms	7.24	0.577	0.752
i-vector cosine	1.55	0.202	0.378
+ Pseudoterms	1.58	0.162	0.317
i-vector PLDA	0.45	0.088	0.191
+ Pseudoterms	0.48	0.079	0.177

3.2. Baseline i-vector Systems

The two i-vector based systems used 40-dimensional MFCCs (20 base + deltas) with short-time mean and variance normalization. They were configured in a completely gender-independent way. The i-vector extractor [1] used a 2048 mixture universal background model and produced 600 dimensional i-vectors. All i-vectors were centered and whitened based on first and second order statistics derived from an in-domain unlabeled set. Additionally, the PLDA system [22] used a 400 dimensional speaker subspace, and it was trained using 3790 speakers with a total of 36470 conversation sides from SRE04, 05, 06, and 08. Score fusion of the proposed pseudoterm and i-vector systems was performed using 5-fold cross-validation. Here, the fusion scores for each fold were generated using a logistic regression trained on the labeled trials in the remaining four folds. The fusion scores for all five folds were then combined into a single ranked list for scoring.

3.3. Results

Table 1 lists the equal error rate (EER) and minimum decision cost function (minDCF) at two target trial priors (0.01 and 0.001, with equal miss and false alarm costs) for the proposed pseudoterm and i-vector systems, as well as the fusion of the two. We measured pseudoterm system performance that is comparable to standalone performance of best high-level idiolectal features considered in past [5]. Since the pseudoterm system makes no use of labeled out-of-set data, its fair baseline is the unsupervised i-vector system using cosine scoring. The minDCF in the low false alarm region (target prior of 0.001) is 2X that of the i-vector cosine system, but fusion of the two produces a 16% relative (0.06 absolute) reduction. Additional supervision available to the PLDA system functions to erase most of the complementarity of the two systems, though future supervised versions of our pseudoterm approach may restore fusion gains.

Figure 1 shows detection error trade-off (DET) curves for the pseudoterm system, baseline i-vector systems, and the fusion of the two. The pseudoterm system has a more horizontal slope than the i-vector systems that translates to improvements in minDCF and not EER in the fusion results. We see clear separation introduced by the pseudoterm system when fused with the unsupervised i-vector cosine system which is reflected in the substantial reduction in minDCF

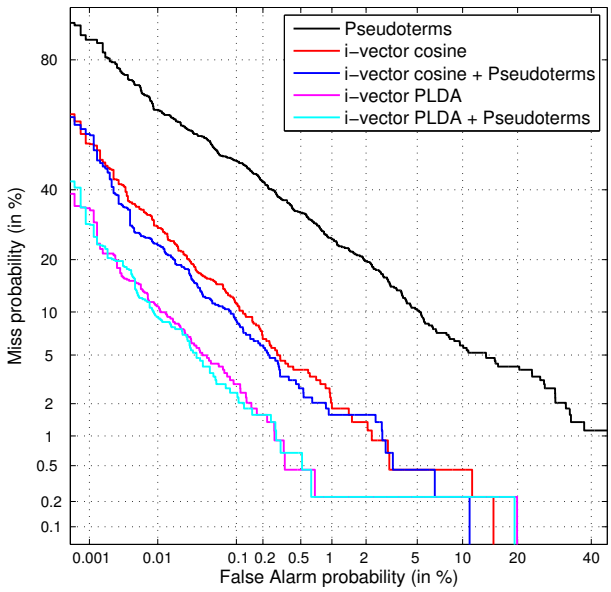

Fig. 1. DET curves for the various systems and combinations.

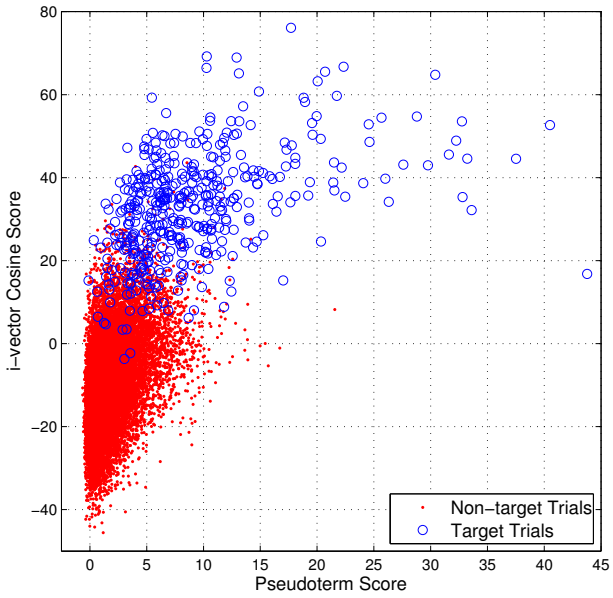

Fig. 2. Comparison of i-vector cosine and pseudoterm scores.

listed in Table 1. The lack of clear separation between the PLDA system with and without fusion indicate that the minDCF improvements listed in Table 1 are not significant.

4. DISCUSSION

Short-time spectral speaker recognition methods are capable of representing a speech cut of arbitrarily short duration, accumulating information at the frame rate (100 Hz). Our pseudoterm system, on the other hand, can at best accumulate speaker identifying information at the rate of occurrence of words (1 Hz). When given a finite amount of enrollment and test cut speech, this low accumulation rate can combine with limited idiolect variation from some speakers to produce zero (or near-zero) scoring target trials. In these cases, target trials are equivalent to empty cuts and are indistinguishable from non-target trials. This phenomena produces a miss rate floor that would remain insurmountable even if we had perfect ranking of non-zero scoring trials, leading to the substantial gap in performance between the pseudoterm and i-vector systems.

To further illustrate this behavior, Figure 2 shows a scatterplot of two pseudoterm-only and baseline i-vector cosine scores for target (open circles) and non-target (dots) trials. We see that the i-vector system scores for each trial subset are normally distributed, while those for the pseudoterm system are closer to an exponential form. The non-target trial scores for the pseudoterm system are concentrated under a score of 10, with a high degree of overlap with a large proportion of the target trials. However, when the pseudoterm score is greater than 10, the bias towards target trials is drastic and the tail is long. This implies that when the idiolect signal is present it is a reliable indicator. However, its presence for only a fraction of speakers leads to substantially higher error rates than the i-vector baseline.

We can clearly observe from Figure 2 the imperfect correlation between the two system scores that results in the fusion gains listed in Table 1. The 0.06 absolute reduction in $\text{minDCF}_{0.001}$ from fusing the pseudoterm and the i-vector cosine system can be broken down into a 6% reduction in misses and a much larger 46% reduction in false alarms. The reduction in miss rate arises from boosting target trials where channel mismatch limits short-time spectral similarity, but that contain an abnormally high frequency of speaker characteristic pseudoterms (e.g. the speaker consistently utters "right" or "okay" in the back-channel). The reduction in false alarms arises from suppressing trials where spectrally similar speakers have relatively little idiolectal overlap.

Finally, we consider the important issue of run-time. The computational efficiency of i-vector systems is well documented, with feature extraction (acoustic features plus i-vectors) running at approximately 50X faster than real-time, with essentially negligible runtime for both cosine and PLDA scoring. For the pseudoterm system, acoustic feature extraction plus LSH runs about 500X faster than real-time. However, the computational bottleneck transfers onto our scoring procedure, which requires us to perform term discovery between enrollment speech and the test cut for each trial. The runtime for this step depends on both the amount of enrollment speech and test cut duration. We found that for a speaker with the average amount of enrollment speech, trials can be scored 50X faster than real-time (in the duration of the test cut). However, this scoring procedure must be repeated for each enrollment speaker.

5. CONCLUSIONS

We have presented a fully unsupervised approach to discovering a speaker's idiosyncratic pronunciation and word usage for application to speaker recognition. Unlike previous idiolectal approaches, we do not require any in-language speech recognition tools, allowing us to match the domain versatility of short-time spectral features. However, like other idiolectal approaches, we require a substantial amount of enrollment speech and even when provided, the strength of the idiolectal cues can vary widely by speaker. Still, we measure substantial minDCF reductions when fused with a state-of-the-art unsupervised i-vector baseline system. Future work on our pseudoterm-based system will include exploring alternative acoustic features as input and introducing supervision (comparable to that used for PLDA) to the scoring procedure.

6. REFERENCES

[1] N. Dehak et al., "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 19, no. 4, pp. 788–798, 2011.

[2] D. Garcia-Romero, X. Zhou, and C. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. of ICASSP*, 2012.

[3] G. R. Doddington, "Speaker recognition based on idiolectal differences between speakers.," in *Proc. of Interspeech*, 2001.

[4] D. Reynolds et al., "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. of ICASSP*, 2003.

[5] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I*, pp. 241–259. Springer, 2007.

[6] D. Gillick, S. Stafford, and B. Peskin, "Speaker detection without models," in *Proc. of ICASSP*, 2005.

[7] C.-C. Leung, M. Ferras, C. Barras, and J.-L. Gauvain, "Comparing prosodic models for speaker recognition.," in *Proc. of Interspeech*, 2008.

[8] Walter D Andrews, Mary A Kohler, Joseph P Campbell, and John J Godfrey, "Phonetic, idiolectal and acoustic speaker recognition," in *2001: A Speaker Odyssey-The Speaker Recognition Workshop*, 2001.

[9] W. Andrews et al., "Gender-dependent phonetic refraction for speaker recognition," in *Proc. of ICASSP*, 2002.

[10] D. Klusácek, J. Navrátil, D. A. Reynolds, and J. P. Campbell, "Conditional pronunciation modeling in speaker detection," in *Proc. of ICASSP*, 2003.

[11] K. Boakye and B. Peskin, "Text-constrained speaker recognition on a text-independent task," in *ODYSSEY04*, 2004.

[12] A. Jansen and B. Van Durme, "Efficient spoken term discovery using randomized algorithms," in *Proc. ASRU*, 2011.

[13] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. of EMNLP*, 2010.

[14] H. Aronowitz, D. Burshtein, and A. Amir, "Text independent speaker recognition using speaker dependent word spotting," in *Proc. of Interspeech*, 2004.

[15] M. Gerber and B. Pfister, "Fast search for common segments in speech signals for speaker verification," in *Proc. of Interspeech*, 2008.

[16] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Speaker verification using text-constrained gaussian mixture models," in *Proc. of ICASSP*, 2002.

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

[18] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality.," in *STOC*, 1998.

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

[20] D.C. Smith, J. Townsend, D.J. Nelson, and D. Richman, "A multivariate speech activity detector based on the syllable rate," in *Proc. of ICASSP*, 1999.

[21] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of ICASSP*, 1997.

[22] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. of Interspeech*, 2011.