The Zero Resource Speech Challenge 2015

Maarten Versteegh[1*], *Roland Thiollière*[1*], *Thomas Schatz*[1,4*], *Xuan Nga Cao*[1]
Xavier Anguera[2], *Aren Jansen*[3], *Emmanuel Dupoux*[1]

[1] Ecole Normale Supérieure / PSL Research University / EHESS / CNRS, France
[2] Telefonica Research, Barcelona, Spain
[3] HLTCOE and CLSP, Johns Hopkins University, USA
[4] SIERRA Project-team Ecole Normale Supérieure / INRIA / CNRS, France

maartenversteegh@gmail.com, rolthiolliere@gmail.com, thomas.schatz@laposte.net
ngafrance@gmail.com, xanguera@tid.es, aren@jhu.edu, emmanuel.dupoux@gmail.com

Abstract

The Interspeech 2015 Zero Resource Speech Challenge aims at discovering subword and word units from raw speech. The challenge provides the first unified and open source suite of evaluation metrics and data sets to compare and analyse the results of unsupervised linguistic unit discovery algorithms. It consists of two tracks. In the first, a psychophysically inspired evaluation task (minimal pair ABX discrimination) is used to assess how well speech feature representations discriminate between contrastive subword units. In the second, several metrics gauge the quality of discovered word-like patterns. Two data sets are provided, one for English, one for Xitsonga. Both data sets are provided without any annotation except for voice activity and talker identity. This paper introduces the evaluation metrics, presents the results of baseline systems and discusses some of the key issues in unsupervised unit discovery.

Index Terms: zero resource speech challenge, feature extraction, unsupervised term discovery, new paradigms

1. Introduction

During their first year of life, infants construct acoustic and language models for speech recognition in a robust and largely unsupervised way. Current speech technology is incapable of such a feat, and remains dominated by supervised learning paradigms that rely on massive amounts of human generated linguistic labels. It is time to address this discrepancy by setting up the rather extreme situation in which a whole language has to be learned from scratch. We expect that doing so will impact the speech and language technology field by providing adaptable algorithms that can supplement supervised systems when human annotated corpora are scarce or nonexistent, as well as aid infant language acquisition research by providing scalable quantitative models that can be compared to psycholinguistic data.

This challenge covers two levels of linguistic structure: subword units (Track 1) and word units (Track 2). These two levels have already been investigated in previous work (see [1, 2, 3, 4, 5, 6, 7], and [8, 9, 10, 11], respectively), but the performance of the different systems has not yet been compared using common evaluation metrics and data sets.

The aim of Track 1 (unsupervised subword modeling) is to construct a representation of speech sounds which can support word identification both within and across talkers. We use the ABX discriminability between phonemic minimal pairs (e.g.

---

* These authors contributed equally to this work.

"beg" and "bag") as an indicator of separability of sound categories in the representation (see [12, 13]). We aggregate a score over the entire set of phone triplet minimal pairs in the corpus and analyze separately within- and between-talker variation.

The aim of Track 2 (spoken term discovery) is the unsupervised discovery of word-like units, defined as recurring speech fragments. The systems take raw speech as input and output a list of speech fragments (timestamps indicating intervals in the original audio files) together with a discrete label for category membership. The evaluation uses the suite of metrics described in [14], which enables the detailed assessment of the different components of a spoken term discovery pipeline (matching, clustering, segmentation, parsing) and supports a direct comparison with NLP models of unsupervised word segmentation.

2. Data sets

Two data sets are provided in the challenge to test the participants' systems. The data sets are composed of selected segments from two free and open access data sets, the Buckeye Corpus [15] in American English, and the NCHLT Speech Corpus of Xitsonga [16]. The datasets are deliberately chosen to be small, keeping in mind the high computational demands zero resource systems typically pose. The Buckeye corpus consists of casual conversational speech. Twelve speakers were selected from the corpus (six male and six female, six young and six old) that had the highest common use of words. Of each speaker, between 16 and 30 minutes ($\mu = 24.16$) of speech were selected (for a total of 4h59m05s), such that they contained no speech overlap with the interviewer, no speaker noise and no pauses. Segments that contained boundary mismatches between the phone and word level annotation were similarly excluded. The section of the NCHLT corpus that was used consists of read speech recorded by 24 speakers (12 male, 12 female). Of each speaker, between 2 and 29 minutes were selected ($\mu = 13.16$) with the same criteria as for the Buckeye corpus, for a total of 2h29m07s. For both corpora, the selected segments were provided to the participants and the remaining portions of the corpus were declared non-speech. Each segment or file contained speech for only one speaker and this information as well as the speaker identity was also provided. The evaluation was based on forced aligned intervals for the phonemes in the corpora, but this information was not communicated to the participants.

3. Evaluation Metrics

The main aim of this challenge is to generate knowledge about the mechanisms that underly unsupervised language learning. As a result, for evaluation, we do not use a single applicative task, but rather multiple evaluation metrics, each one designed to characterize a particular subproblem or component of linguistic unit discovery. All the metrics use open source software libraries, some of which are developed in-house. See [17] for Track 1 and [18] for Track 2.

3.1. Track 1

Unsupervised subword modeling can be defined as the task of finding speech features that emphasize linguistically relevant properties of the speech signal (phoneme structure) and de-emphasize the linguistically irrelevant ones (speaker identity, emotion, channel, etc). Some approaches to this task use unsupervised clustering at the frame level using GMM's [2, 3]. Other approaches model phones as frame sequences, with a GMM-HMM architecture similar to those used in typical supervised systems (eg. [5, 4, 7]). Yet other approaches use Deep Neural Net (DNN) architectures with unsupervised or weakly supervised loss functions to learn phone-level embeddings [1, 6, 19]. Importantly, the output of these systems can be in many different formats: transcription in discrete categories, lattices, posteriorgrams, continuous vector embeddings, etc. This raises the problem of a fair evaluation of these models that would not be biased in favor of one format or other.

Typically, subword models are evaluated by training a classifier to decode the representation into phoneme sequences and evaluating these against a gold transcription. A major problem with this approach is that representations that are easily separable on the basis of labeled examples can be indiscriminable in the absence of those labels. This means that defects in a representation that would be fatal if it was to be used as part of a zero-resource system can be unduly corrected by an evaluation metric based on supervised classifiers. Another problem is that the final score is a compound of the quality of the representation and the quality of the decoder. Since the representations all vary in terms of number of dimensions, sparsity and other statistical properties, it is unclear how a single decoder would be fair to all of the above models. Here, we will use a minimal pair ABX task [12, 13], which does not require any training, and only requires a notion of distance between the representations of speech segments.

The ABX task is inspired by match-to-sample tasks used in human psychophysics and is a simple way to measure discriminability between two sound categories (where the sounds A and B belong to different categories x and y, respectively, and the task is to decide whether the sound X belongs to one or the other). Specifically, we define the ABX-discriminability of category x from category y as the probability that A and X are further apart than B and X when A and X are from category x and B is from category y, according to some distance d over the (model-dependent) space of featural representations for these sounds. Given two sets of featural representations that we wish to evaluate, $S(x)$ and $S(y)$ from category x and y respectively, we estimate this probability using the following formula:

$$\frac{1}{m(m-1)n} \sum_{a \in S(x)} \sum_{b \in S(y)} \sum_{x \in S(x) \setminus \{a\}}$$
$$\left(\mathbb{1}_{d(a,x)<d(b,x)} + \frac{1}{2}\mathbb{1}_{d(a,x)=d(b,x)} \right) \quad (1)$$

where m and n are the number of sounds in $S(x)$ and $S(y)$ and $\mathbb{1}$ is the indicator function. As the probability defined above is asymmetric in the two categories, we obtain a symmetric measure by taking the average of the ABX discriminability of x from y and of y from x. The default distances provided in this challenge are based on DTW divergences with the underlying frame-level distance being either cosine distance or KL-divergence. For most systems (signal processing, embeddings) the cosine distance usually gives good results, and for others (posteriorgrams) the KL distance is more appropriate. Contestants are allowed to supply their own distance function as long as it was not obtained through supervised training.

We focus on minimal pairs, the smallest difference in speech sound which makes a semantic difference (e.g. "beg" vs "bag"), as they represent the hardest problem that a speech recognizer may solve. Since there are typically not enough word minimal pairs in a small corpus to do this kind of analysis, we use phone triplet minimal pairs, i.e. sequences of 3 phonemes that differ in the central sound (e.g. "beg"-"bag", "api"-"ati", etc). Our compound measure sums over all minimal pairs of this type found in the corpus in a structured manner, that depends on the task. For the *within-speaker* task, all of the phone triplets belong to the same speaker (e.g. $A = \text{beg}_{T1}$, $B = \text{bag}_{T1}$, $X = \text{bag}'_{T1}$). The scores for a given minimal pair are first averaged across all of the speakers for which this minimal pair exists. The resulting scores are then averaged over all found contexts for a given pair of central phones (e.g. for the pair /a/-/e/, average the scores for the existing contexts such as /b_g/, /r_d/, /f_s/, etc.). Finally the scores for every pair of central phones are averaged and subtracted from 1 to yield the reported within-talker ABX error rate. For the *across-speaker* task, A and B belong to the same speaker, and X to a different one. $A = \text{beg}_{T1}$, $B = \text{bag}_{T1}$, $X = \text{bag}_{T2}$. The scores for a given minimal pair are first averaged across all of the pairs of speakers for which this contrast can be made. As above, the resulting scores are then averaged over all contexts for each possible pair of central phones and finally over all pairs of central phones before being converted to an error rate.

3.2. Track 2

The process of spoken term discovery can be broken down into a series of three operations, which can be all evaluated independently (see Figure 1). The first step consists of matching
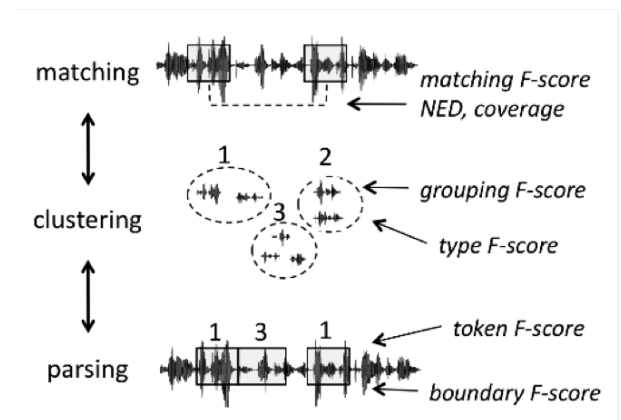

Figure 1: *Logical components of a spoken term discovery system*

pairs of stretches of speech on the basis of their similarity. The second step consists in clustering the matching pairs, thereby building a library of classes with potentially many instances. This can be seen as the equivalent of building a lexicon. In the third step, the system can use its acquired classes to parse the continuous stream into candidate tokens and boundaries. Some systems may only implement some of these steps, others may do them simultaneously rather than sequentially. The metrics below have been devised to enable comparisons between systems that may implement any or all of these steps by evaluating them separately. Each of them represents a different view of the performance of a term discovery system, highlighting the different operations in Figure 1.

All of the metrics assume a time aligned transcription, where $T_{i,j}$ is the (phoneme) transcription corresponding to the speech fragment designated by the pair of indices $[i, j]$ (i.e., the speech fragment between time points i and j). If the left or right edge of the fragment contains part of a phoneme, that phoneme is included in the transcription if it corresponds to more than more than 30ms or more than 50% of its duration.

Define C_{disc} to be the set of discovered pattern clusters (a cluster being a set of fragments). Then define the following sets, containing the *gold* supervision:

$$F_{\text{all}} = \{\langle i, j \rangle \in \mathbb{N} \times \mathbb{N} \mid$$
$$1 \leq i \leq j \leq n, 3 \leq j - i + 1 \leq 20\} \quad (2)$$

$$P_{\text{all}} = \{\langle\langle i, j \rangle, \langle k, l \rangle\rangle \in F_{\text{all}} \times F_{\text{all}} \mid$$
$$T_{i,j} = T_{k,l}, [i, j] \cap [k, l] = \varnothing\} \quad (3)$$

$$P_{\text{goldclus}} = \{\langle\langle i, j \rangle, \langle k, l \rangle\rangle \in F_{\text{all}} \times F_{\text{all}} \mid$$
$$\exists c_1, c_2 \in C_{\text{disc}} : \langle i, j \rangle \in c_1, \langle k, l \rangle \in c_2$$
$$T_{i,j} = T_{k,l}, [i, j] \cap [k, l] = \varnothing\} \quad (4)$$

$$P_{\text{goldLex}} = \{\langle\langle i, j \rangle, \langle k, l \rangle\rangle \in F_{\text{all}} \times F_{\text{all}} \mid$$
$$T_{i,j} = T_{k,l}, [i, j] \cap [k, l] = \varnothing,$$
$$\{i, j, k, l\} \subseteq \text{cover}(P_{\text{disc}})\} \quad (5)$$

Next, derive the set of discovered fragments F_{disc}, discovered fragment pairs P_{disc} and discovered boundaries B_{disc}:

$$F_{\text{disc}} = \{f \mid f \in c, c \in C_{\text{disc}}\} \quad (6)$$
$$P_{\text{disc}} = \{\langle f_1, f_2 \rangle \mid f_1 \neq f_2 \in c, c \in C_{\text{disc}}\} \quad (7)$$
$$B_{\text{disc}} = \{i \mid \exists j : \langle i, j \rangle \in F_{\text{disc}} \wedge \langle j, i \rangle \in F_{\text{disc}}\} \quad (8)$$

Lastly, define the set $P_{\text{disc*}}$ to be the pairwise substring completion of P_{disc}, i.e. for every pair in $P_{\text{disc*}}$, all alignments of substrings of at least three phones are also in $P_{\text{disc*}}$, e.g. for fragment pair $\langle abcd, efgh \rangle$, add $\langle abcd, efgh \rangle$, $\langle abcd, efg \rangle$, $\langle abcd, fgh \rangle$, $\langle abc, efgh \rangle$, $\langle abc, efg \rangle$, $\langle abc, fgh \rangle$, $\langle bcd, efgh \rangle$, $\langle bcd, efg \rangle$, and $\langle bcd, fgh \rangle$. The rationale for adding these substrings is to not unduly punish a system for partially discovered phone sequences in the metrics below.

Matching quality. Many spoken term discovery systems incorporate a step in which fragments of speech are realigned and compared. Matching quality measures the accuracy of this process. Two sets of metrics evaluate this: Normalized Edit Distance (NED) & Coverage, and Matching f-score.

NED and Coverage are quick to compute and give a qualitative estimate of the matching step. The normalized Edit Distance between a pair of fragments is equal to zero when they have exactly the same transcription, and 1 when they differ in all phonemes. Coverage is the fraction of the corpus that is covered by the discovered fragments.

$$\text{NED} = \sum_{\langle x, y \rangle \in P_{\text{disc}}} \frac{\text{ned}(x, y)}{|P_{\text{disc}}|} \quad (9)$$

$$\text{Coverage} = \frac{|\text{cover}(P_{\text{disc}})|}{|\text{cover}(P_{\text{all}})|} \quad (10)$$

where

$$\text{ned}(\langle i, j \rangle, \langle k, l \rangle) = \frac{\text{Levenshtein}(T_{i,j}, T_{k,l})}{\max(j - i + 1, k - l + 1)} \quad (11)$$

$$\text{cover}(P) = \bigcup_{\langle i, j \rangle \in \text{flat}(P)} [i, j] \quad (12)$$

The Matching metrics are more exhaustive, but require more computation. They compare $X = P_{\text{disc*}}$ the set of discovered pairs (with substring completion) to $Y = P_{\text{all}}$ the set of all possible gold pairs. The precision and recall are computed over each type of pairs, and averaged after reweighting by the frequency of the pair.

$$\text{prec} = \sum_{t \in \text{types}(P_{\text{disc*}})} w(t, P_{\text{disc*}}) \frac{|\#\text{occ}(t, P_{\text{disc*}} \cap P_{\text{all}})|}{|\#\text{occ}(t, P_{\text{disc*}})|} \quad (13)$$

$$\text{recall} = \sum_{t \in \text{types}(P_{\text{all}})} w(t, P_{\text{all}}) \frac{|\#\text{occ}(t, P_{\text{disc*}} \cap P_{\text{all}})|}{|\#\text{occ}(t, P_{\text{all}})|} \quad (14)$$

where

$$\text{types}(F) = \{T_{i,j} \mid \langle i, j \rangle \in \text{flat}(F)\} \quad (15)$$
$$\text{flat}(P) = \{p \mid \exists q : \langle p, q \rangle \in P \text{ or } \langle q, p \rangle \in P\} \quad (16)$$
$$\#\text{occ}(t, P) = \{\langle i, j \rangle \in \text{flat}(P) \mid T_{i,j} = t\} \quad (17)$$
$$w(t, P) = \frac{|\#\text{occ}(t, P)|}{|\text{flat}(P)|} \quad (18)$$

Clustering quality is evaluated using two metrics. The first set of metrics, Grouping, computes the intrinsic quality of the clusters in terms of their phonetic composition. This score is equivalent to the purity and inverse purity scores used for evaluating clustering. As the Matching score, it is computed over pairs, but contrary to the Matching scores, it focusses on the covered part of the corpus.

$$\text{prec} = \sum_{t \in \text{types}(P_{\text{clus}})} w(t, P_{\text{clus}}) \frac{|\#\text{occ}(t, P_{\text{clus}} \cap P_{\text{goldclus}})|}{|\#\text{occ}(t, P_{\text{clus}})|} \quad (19)$$

$$\text{recall} = \sum_{t \in \text{types}(P_{\text{goldclus}})} w(t, P_{\text{goldclus}}) \frac{|\#\text{occ}(t, P_{\text{clus}} \cap P_{\text{goldclus}})|}{|\#\text{occ}(t, P_{\text{goldclus}})|} \quad (20)$$

The second set of metrics, Type, takes as the gold cluster set the true lexicon and is therefore much more demanding of the systems, requiring them to find actual words and not just "word-like" patterns:

$$\text{prec} = \frac{|\text{types}(F_{\text{disc}}) \cap \text{types}(F_{\text{goldLex}})|}{|\text{types}(F_{\text{disc}})|} \quad (21)$$

$$\text{recall} = \frac{|\text{types}(F_{\text{disc}}) \cap \text{types}(F_{\text{goldLex}})|}{|\text{types}(F_{\text{goldLex}})|} \quad (22)$$

Table 1: *Baseline and topline results on all the metrics for Track 2. The baseline system is a spoken term discovery system on PLP features. The topline is a unigram adaptor grammar run on phoneme sequences from the gold transcription.*

	NED	Cov	Matching			Grouping			Type			Token			Boundary		
			P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
English																	
Baseline	0.219	0.163	0.394	0.016	0.031	0.214	0.846	0.333	0.062	0.019	0.029	0.055	0.004	0.080	0.441	0.047	0.086
Topline	0.000	1.000	0.983	0.185	0.311	0.995	1.000	0.997	0.503	0.562	0.531	0.682	0.608	0.643	0.884	0.867	0.875
Xitsonga																	
Baseline	0.120	0.162	0.691	0.003	0.005	0.521	0.774	0.622	0.032	0.014	0.020	0.026	0.005	0.008	0.223	0.056	0.089
Topline	0.000	1.000	1.000	0.068	0.127	1.000	1.000	1.000	0.151	0.181	0.165	0.341	0.497	0.404	0.666	0.919	0.772

Parsing quality is evaluated using two metrics. The first one, Token, evaluates if word tokens were correctly segmented:

$$\text{prec} = \frac{|F_{\text{disc}} \cap F_{\text{goldLex}}|}{|F_{\text{disc}}|} \quad (23)$$

$$\text{recall} = \frac{|F_{\text{disc}} \cap F_{\text{goldLex}}|}{|F_{\text{goldLex}}|} \quad (24)$$

The second metric for parsing quality, Boundary, evaluates how many of the gold word boundaries were found:

$$\text{prec} = \frac{|B_{\text{disc}} \cap B_{\text{gold}}|}{|B_{\text{disc}}|} \quad (25)$$

$$\text{recall} = \frac{|B_{\text{disc}} \cap B_{\text{gold}}|}{|B_{\text{gold}}|} \quad (26)$$

4. Baseline and Topline results

In order to provide a sense of the scale of variation of the different metrics we used, we provide baseline and topline results for the two tracks. For Track 1, the results are shown in Table 2. As baseline, we used 13 MFCC features, computed over 25ms windows with a 10ms window shift and cosine as the pairwise frame distance. As topline, we used posteriorgrams extracted from a Kaldi GMM-HMM pipeline with triphone states, speaker adaptation and a bigram word LM (for details, see [20]). The acoustic and language models were trained on the part of the corpora not used in the evaluation. The same Kaldi recipe was used for the two languages. On the evaluation sets, it gave a phone error rate (PER) of 26.4% for English, and 7.5% for Xitsonga. The ABX scores were calculated using DTW with KL-divergence on the systems' posteriors.

For the baseline, we found comparable results across the two languages, with performance within talker better than across talkers, due to the fact that MFCCs are not speaker invariant. For the topline, we found much better results for the Xitsonga than the English datasets. This is in line with the fact that the former is read speech, whereas the latter is casual conversational speech. There, the difference between within and across talkers was much reduced, reflecting the fact that after supervised training, the system's posteriors almost achieve talker invariance. The systems in the challenge are expected to fall in between the performance of these two systems, although it may be possible to reach the topline.

For Track 2, the baseline and topline results are shown in Table 1. As baseline, we used the word discovery system described in [8] run on PLP features. It performs DTW matching and uses random projections for increased efficiency, and connected component clustering as a second step. The topline is a unigram Adaptor Grammar [21] run on the gold phoneme transcription. This topline performance is probably not attainable by unsupervised systems since it uses the gold transcription.

Table 2: *Within and across talker ABX minimal pair distrimination error rate (Track 1) for the Baseline (MFCC) and Topline (supervised HMM-GMM posteriorgrams).*

	English		Xitsonga	
	within	across	within	across
Baseline	15.6	28.1	19.1	33.8
Topline	12.1	16.0	3.5	4.5

The performance metrics for the baseline system give an idea of the kind of performance that one can obtain with a spoken term discovery system run on standard speech features. For both languages, the NED was between 22% and 12%, and the coverage was only around 16% of the corpus. This baseline system is evidently weighted in favor of high precision and low coverage/recall as confirmed by the matching metrics. In terms of clustering, however, the system has a rather high recall, which corresponds to cluster collocation. In other words, the clusters contain several words, but each word tends to be in only one or very few clusters. As expected, the NLP type and token metrics are not very good for a system that does not attempt to optimize a lexicon. However, the boundary precision figures are reasonably high (between 22% and 44%). The topline gives an idea of what one could obtain with perfect speech features (oracle phoneme transcription), and a lexical optimization system. Here, the matching and grouping statistics are at ceiling (note however that the matching recall is not at 100%, since the system only finds units that are exactly one word long, and therefore fails to return repeated fragments that are larger than a word ("the dog") or straddle two words ("the do"). The type/token and boundary scores are within the expected values for NLP system with better performances for English.

5. Conclusions

We presented the Zero Speech Challenge 2015, which is the first one to unify the unsupervised discovery of linguistic units under a common methodology. In this early phase of the field, intead of focussing on a single application or metric, we preferred to keep the challenge application neutral, and therefore provided a multitude of evaluation metrics.

6. Acknowledgements

7. References

[1] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An Auto-encoder based approach to unsupervised learning of subword units," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[2] M. Huijbregts, M. McLaren, and D. van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4436–4439.

[3] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training." in *ICASSP*, 2013, pp. 8091–8095.

[4] C.-y. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 2012, pp. 40–49.

[5] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008, pp. 165–168.

[6] G. Synnaeve, T. Schatz, and E. Dupoux, "Phonetic embedding learning with side information," in *Proceedings of IEEE Spoken Language Technology*, 2014.

[7] M.-h. Siu, H. Gish, A. Chan, W. Belfield, and S. Lowe, "Unsupervized training of an HMM-based self-organizing recognizer with applications to topic classification and keyword discovery," *Computer Speech & Language*, vol. preprint, 2013.

[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*, 2011, pp. 401–406.

[9] A. Muscariello, G. Gravier, and F. Bimbot, "Unsupervised Motif Acquisition in Speech via Seeded Discovery and Template Matching Combination," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20.

[10] A. S. Park and J. R. Glass, "Unsupervised Pattern Discovery in Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[11] Y. Zhang and J. R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4366–4369.

[12] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the Minimal-Pair ABX task (I): Analysis of the classical MFC/PLP pipeline," in *INTERSPEECH*, 2013.

[13] T. Schatz, V. Peddinti, X.-N. Cao, F. Bach, H. Hermansky, and E. Dupoux, "Evaluating speech features with the Minimal-Pair ABX task (II): Resistance to noise," in *INTERSPEECH*, 2014.

[14] B. Ludusan, M. Versteegh, A. Jansen, G. Gravier, X.-N. Cao, M. Johnson, and E. Dupoux, "Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems," in *Proceedings of LREC*, 2014.

[15] M. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," www.buckeyecorpus.osu.edu, 2007.

[16] N. de Vries, M. Davel, J. Badenhorst, W. Basson, F. de Wet, E. Barnard, and A. de Waal, "A smartphone-based asr data collection tool for under-resourced languages," *Speech Communication*, vol. 56, pp. 119–131, 2014.

[17] T. Schatz, R. Thiolliere, E. Dupoux, G. Synnaeve, and E. Dunbar, "ABXpy v0.1," Mar. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.16239

[18] M. Versteegh and R. Thiolliere, "Zerospeech term discovery evaluation toolkit," Mar. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.16330

[19] H. Kamper, M. Elsner, A. Jansen, and S. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proceedings of ICASSP*, 2015.

[20] T. Schatz, X. N. Cao, G. Synnaeve, R. Thiolliere, and E. Dupoux, "abkhazia: Preliminary release," Mar. 2015, File /bin/kaldi/test_and_decode.sh. [Online]. Available: http://dx.doi.org/10.5281/zenodo.16242

[21] M. Johnson, T. Griffiths, and S. Goldwater, "Adaptor grammars: A framework for specifying compositional nonparametric bayesian models," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2007, vol. 19, pp. 641–648.