THE ZERO RESOURCE SPEECH CHALLENGE 2015: PROPOSED APPROACHES AND RESULTS

Maarten Versteegh[1] *Xavier Anguera*[2] *Aren Jansen*[3] *Emmanuel Dupoux*[1]

[1]École Normale Supérieure / PSL Research University / EHESS / CNRS, France
[2]Telefonica Research, Barcelona, Spain
[3]HLTCOE and CLSP, Johns Hopkins University, USA

ABSTRACT

This paper reports on the results of the Zero Resource Speech Challenge 2015, the first unified benchmark for zero resource speech technology, which aims at the unsupervised discovery of subword and word units from raw speech. This paper discusses the motivation for the challenge, its data sets, tasks and baseline systems. We outline the ideas behind the systems that were submitted for the two challenge tracks: unsupervised subword unit modeling and spoken term discovery, and summarize their results. The results obtained by participating teams show great promise; many systems beat the provided baselines and some even perform better than comparable supervised systems.

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

1. INTRODUCTION

Current speech technology relies on larger and larger amounts of labeled data to train acoustic and language models. This is not compatible with the development of speech technologies in under-resourced languages, where there is a long tail of diverse languages used by small communities with limited access to expert knowledge or labelled data. In addition, infants learn acoustic and language models appropriate to their mother tongue during their first year of life in a largely unsupervised manner, providing a proof of principle that one could bootstrap a speech recognition system from raw speech only.

The so-called "zero resource setting" (zero labelled data) is attracting a growing number of research teams [1], but progress has been hampered so far by the absence of common evaluation tools and datasets. To a very large extent, each published paper uses its own datasets, metrics, and (sometimes proprietary) code, resulting in great difficulties to replicate results, compare systems and measure progress.

In 2015, the first Zero Resource Speech Challenge [2] was organized with the aim to address this issue by inviting participating teams to compare their systems within a common open source evaluation scheme. The challenge consisted of two tracks. The aim of Track 1 (subword modeling) was to produce a feature representation from unlabeled speech which maximizes phoneme discriminability. In the unsupervised spirit of the challenge, this track was evaluated without any classifier training, but solely based on the discriminability of phonemes within the feature space. The goal of Track 2 (spoken term discovery) was the unsupervised discovery of word-like units in the speech signal. The systems participating in this track took as input raw speech files and output classes of recurring speech fragments.

The Zero Resource Speech Challenge attracted participants from several groups, who presented their submitted systems in a Special Session at Interspeech 2015. Details of the systems as well as an introductory paper by the challenge organizers can be found in the conference proceedings [2, 3, 4, 5, 6, 7, 8, 9, 10]. Here, we summarize the challenge design decisions and present and discuss the main results and lessons of the submitted systems, providing the first comparative overview of zero resource speech technology.

2. CHALLENGE DESIGN AND BASELINES

The goal of the Zero Resource Speech challenge was to produce a replicable benchmark on which researchers can compare approaches, with both evaluation code and data sets available openly and freely. To this end, two data sets were constructed from the publicly available Buckeye corpus of conversational English [11] and the Xitsonga section of the NCHLT corpus of South Africa's languages [12]. For the English part, 6 male and 6 female speakers were selected for a total of 4h59m05s of speech was selected; for the Xitsonga part, 12 male and 12 female for 2h29m07s. Instructions for reproducing the data sets are available through the challenge website[1], so that researchers not initially involved in the challenge can test their systems under the same conditions.

The evaluation tools used in the challenge are also publicly available, including source code that can be easily adapted to data sets outside the two datasets provided, see [13, 14] for details. In the challenge, participants were responsible for evaluating their own systems, using source code

[1]www.zerospeech.com

provided by the organizers. To aid comparison and interpretation of the participants' results, the challenge provided scores for baseline systems run on the provided databases.

2.1. Track 1: Subword Unit Modeling

The task of unsupervised subword modeling is defined as finding speech features that emphasize linguistically relevant properties of speech, i.e. the phoneme structure, and de-emphasize aspects that are not linguistically relevant, e.g. speaker identity, emotion or channel. Participants received the raw speech of the provided corpora and are tasked with returning a feature representation that maximizes the discriminability between phonemes.

The typical evaluation of feature representations usually proceeds through training a phone classifier and evaluating its classification accuracy. This implies making decisions regarding the choice of the classifier, the optimizing technique, and the measures to limit overfitting that may limit the comparability of the results across systems. For this reason, in the present challenge, we took a different approach and evaluated phoneme discriminability directly on the feature representation using the Minimal-Pair ABX (MP-ABX) task [15, 16]. MP-ABX provides an unsupervised and non-parametric way of evaluating speech representations that has previously proven useful in analysing existing feature pipelines. It measures the ABX-discriminability between phoneme triples that differ only in their center phoneme (the minimal pairs). For phoneme triples a and x from category A and b from category B, the ABX-discriminability in the challenge is defined as the probability that the Dynamic Time Warping (DTW) divergence between a and x is smaller than that between b and x.

2.2. Track 2: Spoken Term Discovery

Spoken term discovery is the task of finding recurring speech fragments, ideally corresponding to the words or word-like units of a language. The challenge provided a total of 17 different metrics for studying each of these steps. Full details on these are available in the introductory paper [2], but here we present a smaller set of metrics that highlight the performance of the submitted systems. Spoken term discovery systems typically consist of a sequence of three steps, each of which can be evaluated independently against a gold annotation at the phoneme level [17]. The first step is pairwise fragment discovery. In this step, pairs of speech fragments in an audio stream are matched if their acoustic similarity is high. At this level the normalized edit distance ("NED") of the phoneme sequences corresponding to the paired speech intervals and the coverage ("COV"), which is the fraction of the audio stream that is covered by the discovered fragments are evaluated. In the second step of term discovery, the previously discovered fragments are clustered into classes. We can evaluate the discovered clusters against the gold lexicon (the "Type" score). In the third step of term discovery, the discovered speech fragments are used to "parse" the audio stream. At this stage, the challenge metrics calculated how many word tokens were correctly segmented (the "Token" score) as well as how many of the gold word boundaries were found (the "Boundary" score).

2.3. Baselines and toplines

The challenge provided baselines for all evaluation measures. We also provided *toplines*, i.e. scores derived from labeled data, which give an indication of the best attainable scores.

For Track 1, on subword unit modeling, the baseline feature representation is MFCC's, a common representation in automatic speech recognition. The topline consists of posteriorgrams derived from a Kaldi GMM-HMM system with triphone states, speaker adaptation and a bigram word model (details in [2]). Table 1 gives the resulting scores.

For Track 2, on spoken term discovery, we provided the following baseline and topline scores. As the baseline, we evaluated a previously existing spoken term discovery system [18]. This systems performs all three steps of spoken term discovery outlined above, so it is a suitable candidate for comparison.[2] For the topline we evaluated the patterns discovered by a an adaptor grammar [19] system based on the phoneme annotation. Table 2 gives the resulting scores.

3. SUBMITTED SYSTEMS

The organizers received 28 registrations and a total of 7 papers (5 for Track 1 and 2 for Track 2) from 14 institutions were accepted for Interspeech publication.

3.1. Track 1: Subword Unit Modeling

The scores of the systems submitted to Track 1 are shown in Table 1. Three main ideas are present in the systems; using top-down information, using articulatory information, and modeling the distribution of the features.

3.1.1. Exploiting top-down information

Renshaw et al. [9] and Thiollière et al. [10] approach the task by exploiting top-down information. They generate word-like pairs using an unsupervised term discovery system (similar to the ones used in Track 2 of this challenge), then use the found matches as input to a neural network, in an effort to find a representation that brings the matches close together in the feature space. The results of this approach seem to consistently beat the baseline, in one case producing the best score in the benchmark. Renshaw et al. input the discovered patterns into a correspondence auto-encoder (CAE), and report on several variants, two of which are shown in Table 1, using word

[2] System available at `github.com/arenjansen/ZRtools`

Authors	System	English		Xitsonga	
		across	within	across	within
baseline	MFCC	28.1	15.6	33.8	19.1
topline	posteriorgrams	*16.0*	*12.1*	*4.5*	*3.5*
Thiollière et al. [10]	STD+ABnet	17.9	12.0	**16.6**	11.7
Renshaw et al. [9]	STD+CAE (English)	21.1	13.5	19.3	11.9
	STD+CAE (Xitsonga)			18.5	11.6
Chen et al. [6]	DPGMM	**16.3**	**10.8**	17.2	**9.6**
Badino et al. [4]	AE6	26.3	17.3	23.6	14.1
	AE12	26.8	16.7	27.4	16.0
	AE12-Bin1 (soft)	28.7	19.7	26.4	17.1
Baljekar et al. [5]	articulatory	*29.8*	*18.4*	*29.7*	*18.1*
	inferred phonemes			*46.0*	*42.8*

Table 1: Results for Track 1 - Subword Unit Modeling. The table shows the ABX scores for the within and across speaker tasks on the two languages in the challenge. Best unsupervised system per condition in **bold**. Scores for systems that make use of supervision in some form are in *italic*. Scores for systems that produce binary features are underlined.

matches from either English or Xitsonga. Thiollière et al. use the discovered segment pairs to train a siamese network [20], to find an embedding in which matching fragments are close together and mismatching fragments are distant. It is possible that the ability to use negative evidence gives the siamese architecture an edge on the correspondance auto-encoder.

3.1.2. Articulatory information

Baljekar et al. [5] use features that are derived from a previously trained speech synthesis system for languages without a writing system. They compare features that are based on a cross-lingual phonetic system with features from segment-based inferred phones, using articulatory features derived directly from the acoustics. While this system uses side-information gleaned from a partially supervised system, it provides an intriguing insight into what is possible with articulatory features, which have been proven to be useful in supervised settings [21].

3.1.3. Modeling the feature space

Badino et al. [4] propose two auto-encoder variants (binarized auto-encoders and hidden-markov-model encoders) to learn very compact representations of the input features. This results in representations that perform better than MFCC's with only 6 features. Interesting variants of the system produce binary features and can learn distinctive phonological features, such as nasality and frication, from raw data. The approach by Chen et al. [6] consist of a pipeline of talker-normalized MFCC's followed by a Dirichlet process Gaussian mixture model (DPGMM). The DPGMM posteriors for each of its inferred components are used directly as features in the task. This approach proves surprisingly succesful in capturing phoneme discriminability, in one case outperforming the topline. This result is an indication that in the zero resource setting, traditional wisdom has to be revisited. In the spirit of the challenge, although not part of it, Agenbag & Niesler [3] propose a system that employs dictionary learning to model the acoustic space, leading to good results on TIMIT.

3.2. Spoken Term Discovery

The scores of the systems participating in Track 2 are shown in Table 2. Two groups participated in this part of the challenge providing 6 systems in total. It is interesting to see that parts of the baselines seem hard to beat. For example the baseline NED is not beaten by any system. But for other measures there is plenty of progress made. For example the type, token and boundary scores show significant improvements.

Räsänen et al. [8] proposes to start spoken term discovery based on the segmentation of the input signal into syllables. They compare three different systems, one existing (Vseg) and two novel (EnvMin and Osc) for segmenting the signal into syllable-length units. The aim in the study is to produce quality candidate word onset and offset locations that are subsequently clustered into longer recurring segments. This approach is highly original and effectively exploits a priori knowledge about the structure of the signal. The results of the procedure tend to be between the baseline and the topline. The systems are especially effective in recovering speech segments that correspond to lexical words. It would be worth exploring how the ideas proposed by Räsänen et al. can be combined with other spoken term discovery systems.

Lyzinski et al. [7] offer a comprehensive exploration of the second step in the spoken term discovery process: cluster-

	System	NED	Cov	Type			Token			Boundary		
English				P	R	F	P	R	F	P	R	F
baseline		**21.9**	16.3	6.2	1.9	2.9	5.5	0.4	8.0	44.1	4.7	8.6
topline		*0.0*	*100.0*	*50.3*	*56.2*	*53.1*	*68.2*	*60.8*	*64.3*	*88.4*	*86.7*	*87.5*
Räsänen et al. [8]	Vseg	89.6	40.6	13.5	11.3	12.3	21.6	4.8	7.9	**76.1**	28.5	41.4
	EnvMin	88.0	42.2	12.7	10.8	11.6	21.6	4.7	7.8	75.7	27.4	40.3
	Osc	70.8	42.4	**14.1**	**12.9**	**13.5**	**22.6**	**6.1**	**9.6**	75.7	33.7	**46.7**
Lyzinski et al. [7]	CC-PLP	77.3	25.5	4.7	2.5	3.3	4.2	0.6	1.0	39.6	7.5	12.7
	CC-FDLPS	61.2	**80.2**	3.1	9.2	4.6	2.4	3.5	2.8	18.8	**64.0**	29.0
	FG-BNF	*36.4*	*46.7*	*2.3*	*2.9*	*2.6*	*1.9*	*0.7*	*1.0*	*31.7*	*14.2*	*19.6*
Xitsonga												
baseline		**12.0**	16.2	3.2	1.4	2.0	**2.6**	0.5	0.8	22.3	5.6	8.9
topline		0.0	100.0	15.1	18.1	16.5	34.1	49.7	40.4	66.6	91.9	77.2
Räsänen et al. [8]	Vseg	78.4	77.7	1.7	4.1	2.4	1.8	1.8	1.8	26.2	26.3	26.3
	EnvMin	61.2	**95.0**	1.1	3.3	1.7	0.8	1.3	1.0	16.3	24.4	19.5
	Osc	63.1	94.7	2.2	6.2	3.3	2.3	3.4	**2.7**	**29.2**	39.4	**33.5**
Lyzinski et al. [7]	FG-PLP	36.1	30.2	3.0	2.7	2.8	2.0	0.9	1.2	19.4	11.2	14.2
	CC-FDLPS	43.2	89.4	**4.9**	**18.8**	**7.8**	2.2	**12.6**	0.8	18.8	**64.0**	29.0
	Louvain-BNF	*34.1*	*67.6*	*2.6*	*6.0*	*3.6*	*1.5*	*2.3*	*2.0*	*14.8*	*29.5*	*19.7*

Table 2: Results for Track 2 - Spoken Term Discovery. The table shows the Normalized Edit Distance (NED) and coverage (Cov) scores, in addition to the precision, recall and f1-scores for Types, Tokens and Boundaries. Best results for fully unsupervised systems are in **bold**. Scores for systems using some form of supervision are in *italic*.

ing discovered pairwise matches into larger classes of word-like units. The study proceeds from the first stage output, i.e. pairwise matching segments, produced by an existing spoken term discovery similar to the baseline of the challenge [18], and evaluates the performance of a set of graph clustering algorithms, of which simple Connected Components (CC), and two modularity based algorithms, FG and Louvain, give the best results. The clustering algorithms are evaluated given different feature representations in the input to the STD algorithm (PLP, FDLPS and supervised bottleneck features trained on a corpus of English speech). This investigation provides crucial insight and shows that the choice of the clustering algorithm can have a large impact on the attainable performance of a spoken term discovery system.

4. CONCLUSIONS

The aim of the Zero Resource Speech Challenge was to provide an open and unified benchmark for evaluating and comparing zero resource speech systems. The challenge introduced two tracks on which to evaluate zero resource systems, subword unit modeling and spoken term discovery, each highlighting an aspect of speech technology in which there was a scarcity of unsupervised systems. The challenge resulted in the comparison of an unprecedented number of systems on the same data sets and using the same evaluation metrics.

The submitted systems show a wealth of novel ideas. For Track 1, The systems employing unsupervised top-down information at the word level [9, 10] have introduced a completely new way of exploiting supervision in acoustic modeling. The systems based around modeling the acoustic space in an unsupervised manner [3, 4, 6] provide insight into the effectiveness of these methods that could be transferred to supervised methods. Lastly, the systems using articulatory information [5] show a promising and intriguing way of extracting and exploiting this type of information.

For Track 2, the systems introducing unsupervised syllable segmentation as a stepping stone in spoken term discovery [8] point to a hitherto unexploited source of information about the location of word-like units. The exploration of clustering algorithms in [7] provides a much-needed underpinning for this important step in spoken term discovery as well as highlighting the need for a good input feature representation.

In future directions, we hope to see the combination of the many ideas showcased in this challenge, which is still open to new participants. The interaction between the two tracks is for now under-explored. For example, the high quality features from Track 1 could be used to improve the term discovery in Track 2. Or the feedback from Track 2 into Track 1 as was shown in several systems [9, 10] could be applied to other feature extraction systems. In general, we expect that the techniques developed in the challenge will supply the speech and language technology fields with powerful, flexible algorithms that can aid supervised speech technology systems in cases in which manually annotated data is scarce or nonexistent.

5. REFERENCES

[1] Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, et al., "A summary of the 2012 JH CLSP Workshop on zero resource speech technologies and models of early language acquisition," in *Proceedings of ICASSP 2013*.

[2] Maarten Versteegh, Roland Thiollière, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux, "The zero resource speech challenge 2015," in *Proceedings of Interspeech*, 2015.

[3] Wiehan Agenbag and Thomas Niesler, "Automatic segmentation and clustering of speech using sparse coding and metaheuristic search," in *Proceedings of Interspeech*, 2015.

[4] Leonardo Badino, Alessio Mereta, and Lorenzo Rosasco, "Discovering discrete subword units with binarized autoencoders and hidden-markov-model encoders," in *Proceedings of Interspeech*, 2015.

[5] Pallavi Baljekar, Sunayana Sitaram, Prasanna Kumar Muthukumar, and Alan Black, "Using articulatory features and inferred phonological segments in zero resource speech processing," in *Proceedings of Interspeech*, 2015.

[6] Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li, "Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study," in *Proceedings of Interspeech*, 2015.

[7] Vince Lyzinski, Gregory Sell, and Aren Jansen, "An evaluation of graph clustering methods for unsupervised term discovery," in *Proceedings of Interspeech*, 2015.

[8] Okko Räsänen, Gabriel Doyle, and Michael C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Proceedings of Interspeech*, 2015.

[9] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater, "A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge," in *Proceedings of Interspeech*, 2015.

[10] Roland Thiollière, Ewan Dunbar, Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux, "A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling," in *Proceedings of Interspeech*, 2015.

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

[12] 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.

[13] Thomas Schatz, Roland Thiolliere, Emmanuel Dupoux, Gabriel Synnaeve, and Ewan Dunbar, "ABXpy v0.1," http://dx.doi.org/10.5281/zenodo.16239, Mar. 2015.

[14] Maarten Versteegh and Roland Thiolliere, "Zerospeech term discovery evaluation toolkit," http://dx.doi.org/10.5281/zenodo.16330, Mar. 2015.

[15] T. Schatz, V. Peddinti, F. Back, 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 *Proceedings of Interspeech*, 2013.

[16] 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 *Proceedings of Interspeech*, 2014.

[17] Bogdan Ludusan, Maarten Versteegh, Aren Jansen, Guillaume Gravier, Xuan-Nga Cao, Mark Johnson, and Emmanuel Dupoux, "Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems," in *Proceedings of LREC*, 2014.

[18] 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.

[19] 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., vol. 19, pp. 641–648. MIT Press, 2007.

[20] Gabriel Synnaeve, Thomas Schatz, and Emmanuel Dupoux, "Phonetics embedding learning with side information," in *IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 106–111.

[21] Vikramjit Mitra, Ganesh Sivaraman, Hosung Nam, Carol Espy-Wilson, and Elliot Saltzman, "Articulatory features from deep neural networks and their role in speech recognition," in *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014, pp. 3041–3045.