THE ZERO RESOURCE SPEECH CHALLENGE 2015: PROPOSED APPROACHES AND RESULTS

Maarten Versteegh1  Xavier Anguera2  Aren Jansen3  Emmanuel Dupoux1

1École Normale Supérieure / PSL Research University / EHESS / CNRS, France  
2Telefonica Research, Barcelona, Spain  
3HLTCOE 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 website1, 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

1www.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 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

²System available at github.com/arenjansen/ZRtools
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 successful 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-

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

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*.
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.
5. REFERENCES


