Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 67 – 72

Procedia
Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

The Zero Resource Speech Challenge 2015: Proposed Approaches and Results

Maarten Versteegh[a,*], Xavier Anguera[b], Aren Jansen[c], Emmanuel Dupoux[a]

[a]*École Normale Supérieure / PSL Research University / EHESS / CNRS, France*
[b]*Telefonica Research, Spain*
[c]*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.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords: 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.

[*] Corresponding author.
E-mail address: maartenversteegh@gmail.com (Maarten Versteegh).

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.031

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 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.

[1] `www.zerospeech.com`