Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

A segmental framework for fully-unsupervised large-vocabulary speech recognition

Herman Kamper*[a], Aren Jansen[b], Sharon Goldwater[a]

[a] School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK
[b] Google, Inc., Mountain View 94043, CA , USA

Received 23 June 2016; received in revised form 16 January 2017; accepted 29 April 2017

Abstract

Zero-resource speech technology is a growing research area that aims to develop methods for speech processing in the absence of transcriptions, lexicons, or language modelling text. Early term discovery systems focused on identifying isolated recurring patterns in a corpus, while more recent full-coverage systems attempt to completely segment and cluster the audio into word-like units—effectively performing unsupervised speech recognition. This article presents the first attempt we are aware of to apply such a system to large-vocabulary multi-speaker data. Our system uses a Bayesian modelling framework with segmental word representations: each word segment is represented as a fixed-dimensional acoustic embedding obtained by mapping the sequence of feature frames to a single embedding vector. We compare our system on English and Xitsonga datasets to state-of-the-art baselines, using a variety of measures including word error rate (obtained by mapping the unsupervised output to ground truth transcriptions). Very high word error rates are reported—in the order of 70−80% for speaker-dependent and 80−95% for speaker-independent systems—highlighting the difficulty of this task. Nevertheless, in terms of cluster quality and word segmentation metrics, we show that by imposing a consistent top-down segmentation while also using bottom-up knowledge from detected syllable boundaries, both single-speaker and multi-speaker versions of our system outperform a purely bottom-up single-speaker syllable-based approach. We also show that the discovered clusters can be made less speaker- and gender-specific by using an unsupervised autoencoder-like feature extractor to learn better frame-level features (prior to embedding). Our system's discovered clusters are still less pure than those of unsupervised term discovery systems, but provide far greater coverage.

Keywords: Unsupervised speech processing; Representation learning; Segmentation; Clustering; Language acquisition

1. Introduction

Despite major advances in supervised speech recognition over the last few years, current methods still rely on huge amounts of transcribed speech audio, pronunciation dictionaries, and texts for language modelling. The collection of these pose a major obstacle for speech technology in under-resourced languages. In some extreme cases, unlabelled speech data might be the only available resource. In this *zero-resource* scenario, unsupervised methods are required to learn representations and linguistic structure directly from the speech signal. Such methods can, for

* Corresponding author.
E-mail address: kamperh@gmail.com (H. Kamper), arenjansen@google.com (A. Jansen), sgwater@inf.ed.ac.uk (S. Goldwater).

http://dx.doi.org/10.1016/j.csl.2017.04.008
10885-2308/

instance, make it possible to search through a corpus of unlabelled speech using voice queries (Park and Glass, 2008), allow topics within speech utterances to be identified without supervision (Siu et al., 2014), or can be used to automatically cluster related spoken documents (Dredze et al., 2010). Similar techniques are required to model how human infants acquire language from speech input (Räsänen, 2012), and for developing robotic applications that can learn a new language in an unknown environment (Sun and Van hamme, 2013; Taniguchi et al., 2015).

Interest in zero-resource speech processing has grown considerably in the last few years, with two central research areas emerging (Jansen et al., 2013a; Versteegh et al., 2015). The first deals with unsupervised representation learning, where the task is to find speech features (often at the frame level) that make it easier to discriminate between meaningful linguistic units (phones or words). This task has been described as 'phonetic discovery', 'unsupervised acoustic modelling' and 'unsupervised subword modelling', depending on the type of feature representations that are produced. Approaches include those using bottom-up trained Gaussian mixture models (GMMs) to produce frame-level posteriorgrams (Zhang and Glass, 2010; Chen et al., 2015), using unsupervised hidden Markov models (HMMs) to obtain discrete categorical output in terms of discovered subword units (Varadarajan et al., 2008; Lee and Glass, 2012; Siu et al., 2014), and using unsupervised neural networks (NNs) to obtain frame-level continuous vector representations (Synnaeve et al., 2014; Renshaw et al., 2015; Zeghidour et al., 2016b).

The second area of zero-resource research deals with unsupervised segmentation and clustering of speech into meaningful units. This is important in tasks such as query-by-example search (Zhang et al., 2012; Levin et al., 2015), where a system needs to find all the utterances in a corpus containing a spoken query, or in unsupervised term discovery (UTD), where a system needs to automatically find repeated word- or phrase-like patterns in a speech collection (Park and Glass, 2008; Jansen and Van Durme, 2011; Lyzinski et al., 2015). UTD systems typically find and cluster only isolated acoustic segments, leaving the rest of the data as background. We are interested in full-coverage segmentation and clustering, where word boundaries and lexical categories are predicted for the entire input. Several recent studies share this goal (Sun and Van hamme, 2013; Chung et al., 2013; Walter et al., 2013; Lee et al., 2015; Räsänen et al., 2015). Successful full-coverage segmentation systems would perform a type of unsupervised speech recognition. This would allow downstream applications, such as query-by-example search and speech indexing (grouping together related utterances in a corpus), to be developed in a manner similar to when supervised systems are available. Unsupervised segmentation and clustering, however, is a daunting task, and current performance lags behind that of even minimally-supervised systems. Nevertheless, previous work has shown that high-error rate unsupervised systems can still be used effectively for a wide range of tasks including topic identification and clustering of spoken documents (Gish et al., 2009; Dredze et al., 2010; Siu et al., 2014), speech-to-speech translation of low-resource languages (Martin et al., 2015; Wilkinson et al., 2016), language recognition (Shum et al., 2016), and in improving purely supervised keyword search systems (Jansen et al., 2013a).

In previous work (Kamper et al., 2016a), we introduced a novel unsupervised segmental Bayesian model for full-coverage segmentation and clustering of small-vocabulary speech. Other approaches mostly perform frame-by-frame modelling using subword discovery with subsequent or joint word discovery. In contrast, our approach models whole-word units directly using a fixed-dimensional embedding representation; any potential word segment (of arbitrary length) is mapped to a fixed-length vector, its *acoustic word embedding*, and the model builds a whole-word acoustic model in the embedding space while jointly performing segmentation. In Kamper et al. (2016a) we evaluated the model in an unsupervised digit recognition task using the TIDigits corpus. Although it was able to accurately segment and cluster the small number of word types (lexical items) in the data, the same system could not be applied directly to multi-speaker data with larger vocabularies. This was due to the large number of embeddings that had to be computed, and the efficiency of the embedding method itself.

In this paper, we present a new system that uses the same overall framework as our previous small-vocabulary system, but with several changes designed to improve efficiency and speaker independence, allowing us to scale up to large-vocabulary multi-speaker data. We believe this is the first full-coverage unsupervised speech recognition system to be applied in this regime; previous systems have either focused on identifying isolated terms (Park and Glass, 2008; Jansen and Van Durme, 2011; Lyzinski et al., 2015), were speaker-dependent (Lee et al., 2015; Räsänen et al., 2015), or used only a small vocabulary (Walter et al., 2013; Kamper et al., 2016a). Given this is the first attempt we are aware of, the results reported here will serve as a useful baseline for future work on unsupervised speech recognition of multi-speaker data with realistic vocabularies.

For our efficiency improvements, we use a bottom-up unsupervised syllable boundary detection method (Räsänen et al., 2015) to eliminate unlikely word boundaries, reducing the number of potential word segments that need to be

Please cite this article as: H. Kamper et al., A segmental framework for fully-unsupervised large-vocabulary speech recognition, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.04.008