Speech Communication 96 (2018) 168–183

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Towards weakly supervised acoustic subword unit discovery and lexicon development using hidden Markov models

Marzieh Razavi[*,a,b], Ramya Rasipuram[c], Mathew Magimai.-Doss[a]

[a] Idiap Research Institute, Martigny CH-1920, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne CH-1015, Switzerland
[c] Apple Inc., Cupertino, CA, USA

ARTICLE INFO

Keywords:
Automatic subword unit derivation
Pronunciation generation
Hidden Markov model
Kullback–Leibler divergence based hidden Markov model
Under-resourced language
Automatic speech recognition

ABSTRACT

State-of-the-art automatic speech recognition and text-to-speech systems are based on subword units, typically phonemes. This necessitates a lexicon that maps each word to a sequence of subword units. Development of a phonetic lexicon for a language requires linguistic knowledge as well as human effort, which may not be always readily available, particularly for under-resourced languages. In such scenarios, an alternative approach is to use a lexicon based on units such as, graphemes or subword units automatically derived from the acoustic data. This article focuses on automatic subword unit based lexicon development using methods that are employed for development of grapheme-based systems. Specifically, we present a novel hidden Markov model (HMM) based formalism for automatic derivation of subword units and pronunciation generation using only transcribed speech data. In this approach, the subword units are derived from the clustered context-dependent units in a grapheme based system using the maximum-likelihood criterion. The subword unit based pronunciations are then generated by learning either a deterministic or a probabilistic relationship between the graphemes and the acoustic subword units (ASWUs). In this article, we first establish the proposed framework on a well-resourced language by comparing it against related approaches in the literature and investigating the transferability of the derived subword units to other domains. We then show the scalability of the proposed approach on real under-resourced scenarios by conducting studies on Scottish Gaelic, a genuinely under-resourced language, and comparing the approach against state-of-the-art grapheme-based ASR approaches. Our experimental studies on English show that the derived subword units can not only lead to better ASR systems compared to graphemes, but can also be transferred across domains. The experimental studies on Scottish Gaelic show that the proposed ASWU-based lexicon development approach scales without any language specific considerations and leads to better ASR systems compared to a grapheme-based lexicon, including the case where ASR system performance is boosted through the use of acoustic models built with multilingual resources from resource-rich languages.

1. Introduction

Speech technologies such as automatic speech recognition (ASR) systems and text-to-speech (TTS) systems typically model subword units as they are 1) more trainable compared to words and, 2) more generalizable toward unseen contexts or words. Subword modeling entails development of a pronunciation lexicon that represents each word as a sequence of subword units. Typically in the literature, the subword units are the phonemes or phones. Phonetic lexicon development requires linguistic expert knowledge about the phone set of the language and the relationship between the written form, i.e., graphemes and phonemes. Therefore, it is a time consuming and tedious task. To reduce the amount of human effort, grapheme-to-phoneme (G2P) conversion approaches have been proposed (Pagel et al., 1998; Sejnowski and Rosenberg, 1987; Taylor, 2005; Bisani and Ney, 2008). The G2P conversion approaches still require an initial phonetic lexicon in the target language to learn the relation between graphemes and phonemes through data-driven approaches. While majority languages such as English and French have well-developed phonetic lexicons, there are many other languages such as Scottish Gaelic and Vietnamese that lack proper phonetic resources.

In the absence of a phonetic lexicon, alternatively grapheme subword units based on the writing system have been explored in the literature (Kanthak and Ney, 2002a; Killer et al., 2003; Dines and Magimai.-Doss, 2007; Magimai-Doss et al., 2011; Ko and Mak, 2014; Rasipuram and Magimai.-Doss, 2015; Gales et al., 2015). The main

[*] Corresponding author at: Idiap Research Institute, CH-1920 Martigny, Switzerland.
E-mail addresses: marzieh.razavi@idiap.ch (M. Razavi), ramya.murali@gmail.com (R. Rasipuram), mathew@idiap.ch (M. Magimai.-Doss).

https://doi.org/10.1016/j.specom.2017.11.011
Received 17 March 2017; Received in revised form 13 October 2017; Accepted 30 November 2017
Available online 07 December 2017

advantage of using graphemes as subword units is that they make development of lexicons easy. However, the success of grapheme-based ASR systems depends on the G2P relationship of the language. For languages with a regular or shallow G2P relationship such as Spanish, the performance of grapheme-based and phoneme-based ASR systems is typically comparable, whereas for languages with an irregular or deep G2P relationship such as English, the performance of a grapheme-based ASR system is relatively poor when compared to a phoneme-based system (Kanthak and Ney, 2002a; Killer et al., 2003).

Yet another way to handle lack of phonetic lexicon is to derive subword units automatically from the speech signal and build a lexicon based on that. In the literature, interest in acoustic subword unit (ASWU) based lexicon development emerged from the pronunciation variation modeling perspective, specifically with the idea of overcoming the limitations of linguistically motivated subword units, i.e., phones (Lee et al., 1988; Svendsen et al., 1989; Paliwal, 1990; Lee et al., 1988; Bacchiani and Ostendorf, 1998; Holter and Svendsen, 1997). However, recently, there has been a renewed interest from the perspective of handling lexical resource constraints (Singh et al., 2000; Lee et al., 2013; Hartmann et al., 2013). A limitation of most of the existing methods for acoustic subword units based lexicon development is that they are not able to handle unseen words.

In this article, building upon the recent developments in grapheme-based ASR, we propose an approach to derive "phone-like" subword units and develop a pronunciation lexicon given limited amount of transcribed speech data. In this approach, first a set of ASWUs is derived by modeling the relationship between the graphemes and the acoustic speech signal in a hidden Markov model (HMM) framework based on two well-known aspects,

1. alphabetic writing systems carry information regarding the spoken system. Alternatively, a written text embeds information about how it should be spoken. Though this embedding can be deep or shallow depending on the language; and
2. the envelope of the short-term spectrum tends to carry information related to phones.

The ASWU-based pronunciation lexicon is then developed by learning the grapheme-to-ASWU (G2ASWU) relationship through the acoustic signal, and inferring pronunciations using G2ASWU conversion (analogous to G2P conversion). The G2ASWU conversion process inherently brings in the capability to generate pronunciation for unseen words. The viability of the proposed approach has been demonstrated through preliminary studies on English (Razavi and Magimai-Doss, 2015) and Scottish Gaelic (Razavi et al., 2015), where a probabilistic G2ASWU relationship was learned and pronunciation lexicon was developed.

This article builds on the preliminary works to first extend the approach to the case where a deterministic G2ASWU relationship is learned. We then study and contrast the two G2ASWU relationship learning methods and investigate the following aspects:

1. *Domain-independency of the ASWUs*: Subword units such as phones and graphemes are by default domain-independent. This enables using a lexicon based on either of them across different domains. ASWUs are derived from a limited amount of acoustic speech signal from a domain. Furthermore, the limited data can have undesirable variabilities based on the hardware used and the conditions under which the data is collected. Therefore a question that arises is whether the derived ASWUs are domain independent. Through a cross-domain study on English, we show that our approach indeed yields ASWUs that are domain independent. Furthermore, the proposed approach inherently enables transferring ASWU based lexicon developed on one domain to another.
2. *Potential of ASWUs in improving mulitilingual ASR*: It has been shown that both acoustic resource and lexical resource constraints can be effectively addressed by learning a probabilistic relationship between graphemes of the target languages and a multilingual phone set obtained from lexical resources of auxiliary languages using acoustic data (Rasipuram and Magimai.-Doss, 2015). Success of such approaches lies on the fact that there exists a systematic relationship between linguistically motivated grapheme units and phonemes. Therefore a question that arises is: Does the ASWU-based lexicon based on the proposed approach hold the advantage over grapheme-based lexicon in such a case? Alternately, do the ASWUs exhibit similar systematic relationship to multilingual phones and can it be exploited to further improve the under-resourced language ASR? Through a study on Scottish Gaelic, a genuinely under-resourced language, we show that there exists a systematic relationship between the ASWUs and multilingual phones, which can not only be exploited to yield systems better than grapheme-based lexicons, but also to gain insight into the derived units.

It is worth mentioning that, to the best of our knowledge, this is the first work that aims to establish these aspects in the context of ASWU-based lexicon development. Consequently, it paves the path for adopting ASWU-based lexicon development and its use for ASR technology development, especially for under-resourced languages.

The remainder of the article is organized as follows. Section 2 provides a background about the grapheme-based ASR and related approaches in the literature for subword unit derivation and pronunciation generation. Section 3 describes the proposed approach. Section 4 presents investigations on the well-resourced majority language English and Section 5 presents the investigations on the under-resourced minority language Scottish Gaelic. Section 6 provides a brief analysis of the derived ASWUs and the generated pronunciations. Finally, Section 7 concludes the article.

2. Background

This section provides the relevant background for understanding the proposed approach for ASWU based lexicon development. Sections 2.1 and 2.2 first present a background on HMM-based ASR and grapheme-based ASR approaches, which form the basis for our proposed approach for automatic subword unit derivation and pronunciation generation. Section 2.3 then presents a survey on the existing approaches for derivation of ASWUs and lexicon development.

2.1. HMM-based ASR

In statistical automatic speech recognition, given the acoustic observation sequence $X = [\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_T]$ with T denoting the total number of frames, the goal is to find the most probable sequence of words W^*,

$$W^* = \arg\max_{W \in \mathscr{W}} P(W|X, \Theta), \tag{1}$$

$$= \arg\max_{W \in \mathscr{W}} p(W, X|\Theta), \tag{2}$$

where $\mathscr{W}$ denotes the set of hypotheses and Θ denotes the set of parameters. Eq. (2) is obtained result of applying Bayes' rule and assuming $p(X)$ to be constant w.r.t all word hypotheses. Hereafter for simplicity, we drop Θ from the equations.

The HMM-based ASR approach achieves that goal by finding the most probable sequence of states Q^* representing W^* by incorporating lexical and syntactic knowledge:

$$Q^* = \arg\max_{Q \in \mathscr{Q}} p(Q, X), \tag{3}$$

$$= \arg\max_{Q \in \mathscr{Q}} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t = l^i) \cdot P(q_t = l^i|q_{t-1} = l^j), \tag{4}$$