Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 80 (2016) 1–21

www.elsevier.com/locate/specom

Acoustic data-driven grapheme-to-phoneme conversion in the probabilistic lexical modeling framework

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

[a] Idiap Research Institute, Martigny CH-1920, Switzerland
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne CH-1015, Switzerland

Received 13 May 2015; received in revised form 14 March 2016; accepted 21 March 2016
Available online 25 March 2016

Abstract

One of the primary steps in building automatic speech recognition (ASR) and text-to-speech systems is the development of a phonemic lexicon that provides a mapping between each word and its pronunciation as a sequence of phonemes. Phoneme lexicons can be developed by humans through use of linguistic knowledge, however, this would be a costly and time-consuming task. To facilitate this process, grapheme-to-phoneme conversion (G2P) techniques are used in which, given an initial phoneme lexicon, the relationship between graphemes and phonemes is learned through data-driven methods. This article presents a novel G2P formalism which learns the grapheme-to-phoneme relationship through acoustic data and potentially relaxes the need for an initial phonemic lexicon in the target language. The formalism involves a training part followed by an inference part. In the training part, the grapheme-to-phoneme relationship is captured in a probabilistic lexical modeling framework. In this framework, a hidden Markov model (HMM) is trained in which each HMM state representing a grapheme is parameterized by a categorical distribution of phonemes. Then in the inference part, given the orthographic transcription of the word and the learned HMM, the most probable sequence of phonemes is inferred. In this article, we show that the recently proposed acoustic G2P approach in the Kullback–Leibler divergence-based HMM (KL-HMM) framework is a particular case of this formalism. We then benchmark the approach against two popular G2P approaches, namely joint multigram approach and decision tree-based approach. Our experimental studies on English and French show that despite relatively poor performance at the pronunciation level, the performance of the proposed approach is not significantly different than the state-of-the-art G2P methods at the ASR level.

Keywords: Grapheme-to-phoneme conversion; Probabilistic lexical modeling framework; Kullback–Leibler divergence-based hidden Markov model; Automatic speech recognition; Lexicon development.

1. Introduction

Speech technologies such as automatic speech recognition (ASR) and text-to-speech (TTS) systems aim to link two modes of communication, namely the spoken form (speech) and the written form (text). In order to model the relation between the two forms, a shared unit is commonly used. The shared units can typically be the whole words or subword units. However, subword units are preferred to words especially in large vocabulary tasks for two main reasons: (1) they are easily trainable compared to the whole words as the frequency of words in a text follows Zipf's law[1](Powers, 1998), and (2) they are generalizable for unseen words.

The most widely used subword units in current speech processing systems are phones or phonemes[2]. Phonemes can be related to the spoken form (i.e., speech signal). More

* Corresponding author at: Idiap Research Institute, CH-1920 Martigny, Switzerland. Tel.: +41789750084.

E-mail addresses: marzieh.razavi@idiap.ch, marziehrazavi@gmail.com (M. Razavi), ramya.rasipuram@idiap.ch (R. Rasipuram), mathew@idiap.ch (M. Magimai.-Doss).

http://dx.doi.org/10.1016/j.specom.2016.03.003
0167-6393/© 2016 Elsevier B.V. All rights reserved.

[1] According to Zipf's law, the frequency of a word is inversely proportional to its rank in the frequency table.

[2] Phones are units of the speech sounds which can be designed to cover the set of sounds in all languages, while phonemes are "the smallest contrastive linguistic units which may bring about a change of meaning" (Chomsky and Halle, 1968) in a specific language. For the sake of clarity, throughout this article we use the term phoneme as in the literature the grapheme-to-phoneme terminology is dominantly used.

precisely, the envelope of magnitude spectrum of short-term speech signals typically depicts the characteristics of phonemes. They can also be related to the alphabetic written symbols (i.e., graphemes). The link between phonemes and graphemes originates from the alphabetic orthographies which aim to present the phonetic structure of the spoken words in a graphic form (Frost, 1989). The alphabetic orthographies can be deep or shallow depending on the language[3].

Typically, the development of phoneme-based speech technology systems consists of two steps: development of a lexicon consisting of a mapping between each word and its phoneme-based pronunciation followed by system training. The focus of this article is mainly on the phonemic lexicon development. A phonemic lexicon can be developed manually through use of linguistic knowledge. However, manual development of lexicons can be costly in terms of time and money (Davel and Barnard, 2003). In addition, the developed lexicons are required to be constantly augmented with evolution of languages and emergence of new words. Therefore, it is necessary to develop automatic pronunciation generation methods to reduce the amount of human effort. Towards that goal, grapheme-to-phoneme conversion (G2P) methods are applied in which given an initial phonemic lexicon called a *seed lexicon*, typically data-driven and machine learning techniques such as decision trees (Black et al., 1998) or conditional random fields (Wang and King, 2011) are used to learn the grapheme-to-phoneme relationship. The learned grapheme-to-phoneme relationship is then used to infer pronunciations for the unseen words. Most of the G2P approaches rely solely on the seed lexicon for learning the grapheme-to-phoneme relationship while no acoustic information is incorporated within the G2P process.

This article presents a novel G2P formalism in which the grapheme-to-phoneme relationship is learned through speech data along with word level transcriptions. The formalism consists of two phases: a training phase and an inference phase. In the training phase, as the first step, the relationship between acoustic feature observations and phonemes is learned through an acoustic model, such as an artificial neural network (ANN). Then as the second step, the relationship between the graphemes and phonemes is learned in a hidden Markov model (HMM) framework in which the outputs of the acoustic model are used as feature observations. In this HMM framework, each state represents a grapheme and is parameterized by a categorical distribution of phonemes. In the inference phase, given the orthographic transcription of the word, the grapheme-based HMM acts as a generative model and emits a sequence of phoneme posterior probabilities. The sequence of phoneme posterior probabilities is then decoded using an HMM in which each state represents a phoneme to infer the most probable pronunciation for each word.

In this article, we show that the recently proposed acoustic data-driven G2P approach in the framework of Kullback–Leibler divergence-based HMM (KL-HMM) (Rasipuram and Magimai.-Doss, 2012a) is a particular case of this G2P formalism. We then build upon the previous studies on the acoustic G2P approach and study possible ways to refine the method by incorporating recent trends in ANNs including using ANNs with more layers and output units. Furthermore, we benchmark the approach against two popular conventional G2P approaches, namely the joint multigram and the decision tree-based methods. We evaluate the proposed G2P approach at both pronunciation and application (ASR) levels. For the evaluation at the ASR level, we study different facets including combining the proposed G2P approach with conventional G2P approaches.

Our experimental studies on English and Swiss French show that the performance of the proposed approach is not significantly different than the state-of-the-art G2P approaches at the ASR level. In addition, through combining the acoustic G2P approach with conventional G2P approaches, improvements in the ASR performance can be achieved, in particular when a limited amount of data (for G2P model and acoustic model training) is available.

This article is organized as follows. Section 2 provides a background about the existing approaches for pronunciation generation in the literature. Section 3 proposes the novel G2P formalism for learning the grapheme-to-phoneme relationship through acoustic data. Section 4 describes the databases along with the evaluation setups in this study. Section 5 presents the pronunciation level setup, results and analysis. Section 6 provides the experimental setup and results at the ASR level. Finally Section 7 brings the conclusion.

2. Relevant literature

The first step towards building phoneme-based speech technology systems is the development of a phonemic lexicon. Phonemic pronunciations are typically hand-crafted by exploiting the linguistic knowledge. During the preparation of the pronunciation lexicon by linguists, care is taken to minimize word level confusions and consistency is ensured across the lexicon. The hand-crafted phoneme pronunciation lexicon could possibly provide an optimum performance for ASR or TTS. However, design of the phonemic pronunciation lexicon of significant size by linguistic experts is a tedious and costly task. Furthermore, a finite lexicon will always have limited coverage for ASR and TTS systems. For this reason, ASR and TTS systems use G2P methods when hand crafted pronunciations fail to cover the vocabulary of a particular domain. In this section, we first elucidate two classes of G2P methods, namely knowledge-based and data-driven approaches, which have been explored in the literature.

[3] In shallow orthographies, the grapheme-to-phoneme correspondence is one-to-one (e.g., Finnish). In deep orthographies, however, the correspondence between the graphemes and phonemes is not direct. More precisely, the grapheme-to-phoneme relationship may be irregular (e.g., English) in which some prior knowledge about the word is required to accurately predict the relationship (i.e., different rules can be applied to various words). The grapheme-to-phoneme relationship may also be regular, i.e., predictable given a set of linguistic rules. However, accurate prediction of the grapheme-to-phoneme relationship in deep orthographies requires complex linguistic rules (e.g., French).