Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 233–259

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Articulatory feature based continuous speech recognition using probabilistic lexical modeling☆

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

[a] *Idiap Research Institute, Martigny, Switzerland*

[b] *Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

Received 3 July 2014; received in revised form 12 February 2015; accepted 23 April 2015

Available online 4 May 2015

Abstract

Phonological studies suggest that the typical subword units such as phones or phonemes used in automatic speech recognition systems can be decomposed into a set of features based on the articulators used to produce the sound. Most of the current approaches to integrate articulatory feature (AF) representations into an automatic speech recognition (ASR) system are based on a deterministic knowledge-based phoneme-to-AF relationship. In this paper, we propose a novel two stage approach in the framework of probabilistic lexical modeling to integrate AF representations into an ASR system. In the first stage, the relationship between acoustic feature observations and various AFs is modeled. In the second stage, a probabilistic relationship between subword units and AFs is learned using transcribed speech data. Our studies on a continuous speech recognition task show that the proposed approach effectively integrates AFs into an ASR system. Furthermore, the studies show that either phonemes or graphemes can be used as subword units. Analysis of the probabilistic relationship captured by the parameters has shown that the approach is capable of adapting the knowledge-based phoneme-to-AF representations using speech data; and allows different AFs to evolve asynchronously.

Keywords: Automatic speech recognition; Articulatory features; Probabilistic lexical modeling; Kullback–Leibler divergence based hidden Markov model; Phoneme subword units; Grapheme subword units

1. Introduction

Articulatory features describe the properties of speech production, i.e., each sound unit of a language, a phone or a phoneme, can be decomposed into a set of features based on the articulators used to produce it. The use of articulatory feature (AF) representations in an automatic speech recognition (ASR) system is motivated by their abilities such as:

- Better pronunciation modeling: AFs are hypothesized to capture acoustic variation at a finer level than the phoneme-based representation (Deng et al., 1997; Richardson et al., 2003; Livescu et al., 2008).

☆ This paper has been recommended for acceptance by Karen Livescu.

* Corresponding author at: Idiap Research Institute, Martigny, Switzerland. Tel.: +41 277217711; fax: +41 277217712.
E-mail addresses: ramya.rasipuram@idiap.ch (R. Rasipuram), mathew@idiap.ch (M. Magimai.-Doss).

http://dx.doi.org/10.1016/j.csl.2015.04.003
0885-2308/

- Robustness to noise: Different AFs may have variable noise sensitivity. The "divide and conquer" approach provides a framework to exploit the variable noise sensitivity of AFs (Kirchhoff et al., 2002).
- Multilingual and crosslingual portability: AFs can provide better sharing capabilities than phonemes across languages (Stüker et al., 2003; Lal and King, 2013; Siniscalchi et al., 2012).

To incorporate the articulatory knowledge in an ASR system, the following main concerns have to be addressed:

1. AF representations: There exist different types of articulatory representations of speech, e.g., binary features, multivalued features, and government phonological features. AFs defined by Chomsky and Halle (1968) are binary valued features, for example +voice and −voice, +sonorant and −sonorant. However, according to Ladefoged (1993), it is more natural to allow AFs to take multiple values. In government phonological feature system, speech sounds are destructed into a set of primes and can be represented by fusing them structurally (Harris, 1994). In the paper, we have used the multivalued AFs, because it has been argued that they better represent non-binary parameters such as height of vowels and place of articulation (Ladefoged, 1993).
2. Estimation of AFs from acoustic speech signal: In the literature, many approaches have been explored to extract AFs from the acoustic speech signal. For example, techniques based on acoustic-to-articulatory feature codebooks (Hogden et al., 1996; Suzuki et al., 1998), artificial neural networks (Livescu et al., 2008; Kirchhoff et al., 2002; Chang, 2002; Rasipuram and Magimai-Doss, 2011), support vector machines (Juneja and Espy-Wilson, 2004; Scharenborg et al., 2007), Gaussian mixture models (Metze and Waibel, 2002; Stüker et al., 2003), hidden Markov models (Hiroya and Honda, 2004), conditional random fields (Prabhavalkar et al., 2011), nearest neighbour (Næss et al., 2011), dynamic Bayesian networks (Frankel and King, 2005; Frankel et al., 2007) are used.
3. Integration: Integrating AFs into the conventional hidden Markov model (HMM) based ASR framework is a challenging task mainly because of the multiple AF estimators. The dynamic Bayesian network (DBN) based approaches for AF integration preserve the articulatory representation in DBN state space (Livescu and Glass, 2004; Livescu et al., 2008; King et al., 2007). These approaches have shown promising results in lexical access[1] experiments. Posterior probabilities of AFs can be transformed for use as features in tandem speech recognition systems (Cetin et al., 2007, 2007; Lal and King, 2013). Posterior probabilities of AFs are also used to enhance phoneme-based acoustic models (Kirchhoff et al., 2002; Siniscalchi et al., 2012). These approaches however lose other benefits of articulatory representation such as finer granularity and asynchronous evolution.

In this paper, we propose an approach in the framework of probabilistic lexical modeling to integrate multivalued AFs. In a probabilistic lexical model based ASR system, the relationship between subword units in the lexicon and acoustic feature observations is factored into two models using latent variables: An acoustic model which models the relationship between acoustic feature observations and the latent variables; and a lexical model which models a probabilistic relationship between the subword units in the lexicon and the latent variables. In this paper, we show that by choosing the latent variables as multiple multivalued AFs, the approach effectively integrates AFs into the HMM-based ASR framework. The lexical model parameters in the proposed approach capture a probabilistic relationship between subword units and AFs learned through transcribed speech data.

The potential of the proposed approach for AF integration is demonstrated on a continuous speech recognition task through experiments and comparisons with the tandem approach. In the proposed framework we explore the use of domain-independent data for acoustic model training; and phonemes and graphemes as subword units. Furthermore, through the analysis of the lexical model parameters we show that the approach adapts the knowledge-based phoneme-to-AF or grapheme-to-AF relationship and allows different AFs to evolve asynchronously.

The rest of the paper is organized as follows: Section 2 gives an overview of the HMM-based ASR and the framework of probabilistic lexical modeling. Section 3 presents the literature review of approaches that integrate multivalued AFs for ASR in the light of the background information given in Section 2. In Section 4, the approach for AF integration is presented and the contributions of the present paper with respect to prior work are elaborated. Sections 5 and 6 present the experimental setup and results, respectively. Section 7 presents an analysis on the subword-unit-to-AF relationship captured by the lexical model parameters. Finally, in Section 8 we provide a discussion and conclusion.

[1] The task of lexical access involves predicting a word given its phonetic or broad phonetic transcription (Huttenlocher and Zue, 1984).