Speech Communication 86 (2017) 107–120

Contents lists available at ScienceDirect

Speech Communication

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

An automated technique to generate phone-to-articulatory label mapping

Basil Abraham*, S. Umesh

Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, Tamil Nadu 600036, India

ARTICLE INFO

Article history:
Received 1 February 2016
Revised 16 November 2016
Accepted 23 November 2016
Available online 28 November 2016

Keywords:
Articulatory features
Mapping
Phone-CAT
Under-resourced languages
Cross-lingual techniques
Multilayer perceptrons

ABSTRACT

Recent studies have shown that in the case of under-resourced languages, use of articulatory features (AF) emerging from an articulatory model results in improved automatic speech recognition (ASR) compared to conventional mel frequency cepstral coefficient (MFCC) features. Articulatory features are more robust to noise and pronunciation variability compared to conventional acoustic features. To extract articulatory features, one method is to take conventional acoustic features like MFCC and build an articulatory classifier that would output articulatory features (known as pseudo-AF). However, these classifiers require a mapping from phone to different articulatory labels (AL) (e.g., place of articulation and manner of articulation), which is not readily available for many of the under-resourced languages. In this article, we have proposed an automated technique to generate phone-to-articulatory label (phone-to-AL) mapping for a new target language based on the knowledge of phone-to-AL mapping of a well-resourced language. The proposed mapping technique is based on the center-phone capturing property of interpolation vectors emerging from the recently proposed phone cluster adaptive training (Phone-CAT) method. Phone-CAT is an acoustic modeling technique that belongs to the broad category of canonical state models (CSM) that includes subspace Gaussian mixture model (SGMM). In Phone-CAT, the interpolation vector belonging to a particular context-dependent state has maximum weight for the center-phone in case of monophone clusters or by the AL of the center-phone in case of AL clusters. These relationships from the various context-dependent states are used to generate a phone-to-AL mapping. The Phone-CAT technique makes use of all the speech data belonging to a particular context-dependent state. Therefore, multiple segments of speech are used to generate the mapping, which makes it more robust to noise and other variations. In this study, we have obtained a phone-to-AL mapping for three under-resourced Indian languages namely Assamese, Hindi and Tamil based on the phone-to-AL mapping available for English. With the generated mappings, articulatory features are extracted for these languages using varying amounts of data in order to build an articulatory classifier. Experiments were also performed in a cross-lingual scenario assuming a small training data set ($\approx$ 2 h) from each of the Indian languages with articulatory classifiers built using a lot of training data ($\approx$ 22 h) from other languages including English (Switchboard task). Interestingly, cross-lingual performance is comparable to that of an articulatory classifier built with large amounts of native training data. Using articulatory features, more than 30% relative improvement was observed over the conventional MFCC features for all the three languages in a DNN framework.

1. Introduction

The performance of automatic speech recognition (ASR) systems has significantly improved with the application of better acoustic modeling techniques based on deep neural network (DNN). However, many factors such as speaker variability and noisy environment affect the performance of ASR. In the literature, the use of articulatory features has shown to improve the performance of an ASR system in noisy environment in Kirchhoff et al. (2002) and conversational speech in Frankel et al. (2007). Articulatory features are features that represent a speech signal in terms of the underlying articulatory attributes of speech production. These features were used by Schmidbauer (1989a), Schmidbauer (1989b), Kirchhoff et al. (2002), King et al. (2007), Cetin et al. (2007) and Frankel et al. (2007) in their work on automatic speech recognition. Articulatory features can be obtained by the following three ways (Kirchhoff et al., 2002):

* Corresponding author.
E-mail addresses: ee11d032@ee.iitm.ac.in (B. Abraham), umeshs@ee.iitm.ac.in (S. Umesh).

http://dx.doi.org/10.1016/j.specom.2016.11.010
0167-6393/

1. Direct measurement of articulatory parameters, for example, those obtained by cine-radiography (Papcun et al., 1992)
2. Articulatory parameters recovered from acoustic signal by inverse filtering (Schroeter and Sondhi, 1994)
3. Posterior probabilities extracted from conventional acoustic features by means of a statistical classifier (Kirchhoff et al., 2002). This approach generates the so-called *pseudo-articulatory features*.

Only pseudo-articulatory features are considered in this article. To extract pseudo-articulatory features in any language, articulatory classifiers are constructed for different AL groups, such as a group of labels referring to the manner of articulation, and those referring to the place of articulation, etc. To build a robust articulatory classifier, large amounts of data transcribed in terms of AL (e.g., alveolar and dental) in that language are required. It is very difficult to manually transcribe data at frame level in terms of AL. Hence, the usual practice is to obtain a phone-level alignment and convert it into AL using phone-to-AL mapping. The two broad approaches for building articulatory classifiers are as follows:

1. Phone-to-AL mapping, which is *available* in that language: This is a direct way of building articulatory classifiers. Using the phone-to-AL mapping, articulatory classifiers are built in that language and are used to extract articulatory features in that language.
2. Phone-to-AL mapping, which is *not available* in that language: Here, the articulatory classifiers need to be constructed from another language where the phone-to-AL mapping is available. Then the articulatory features are extracted using these articulatory classifiers. This approach was used in Sivadas and Hermansk (2004), Tóth et al. (2008), Thomas et al. (2010), Lal and King (2013) and Çetin et al. (2007).

The articulatory features extracted by the first approach showed consistent improvements compared to conventional acoustic features for ASR (Kirchhoff et al., 2002). However, a failure was noted in some cases when the second approach was used. For example, in Çetin et al. (2007), articulatory classifiers trained using English continuous telephone speech data were used to extract articulatory features for Mandarin broadcast news task. However, they failed to improve the recognition performance of Mandarin task. Similarly, in Tóth et al. (2008), articulatory classifiers trained on English speech were used to generate articulatory features for Hungarian telephone speech. It was reported that improved recognition performance was obtained, but they failed to perform at par with features generated by articulatory classifiers trained only with Hungarian data. The performance degradation observed while using different language data was due to differences in domain and channel variations between databases. A similar effect was also noted in Thomas et al. (2010).

In Lal and King (2013), German, Portuguese and Spanish data were combined and used to build the articulatory classifiers, which were then used to extract articulatory features for each of these languages. The extracted articulatory features showed improved recognition performance for all three languages, but not as much as the articulatory features extracted from articulatory classifiers built with the corresponding language data alone. Hence, conventional wisdom is to train articulatory classifiers with the same language data or other language data collected in similar conditions.

To build articulatory classifiers in any language, a phone-to-AL mapping is required along with the data. For many languages, phone-to-AL mapping is not readily available. In most of the previous works that used AL-based features, a manually generated phone-to-AL mapping was used. For many languages, manually obtaining a phone-to-AL mapping is often difficult, since it requires the assistance of a phonetic expert and transcription may often require phone labels that are not listed in the International Phonetic Alphabet (IPA) symbols. In case of rare languages, getting a phonetic expert is often difficult and hence the situation becomes worse. Even if phonetic experts are available, they may not have a consensus agreement on AL. In order to overcome these problems, we propose an automated way to generate phone-to-AL mapping in a particular language based on the knowledge of phone-to-AL mapping in some well-resourced language. To our knowledge, there are no previous reports on techniques to automatically generate phone-to-AL mapping in any language. The proposed technique uses the interpolation vectors of recently proposed phone cluster adaptive training (Phone-CAT) acoustic modeling technique (Manohar et al., 2013) to generate phone-to-AL mapping.

This paper is organized as follows: In Section 2, a brief review of ASR using articulatory features is given. In Section 3, the proposed phone-to-AL mapping technique is described. In Section 4, the proposed technique is compared with alternate techniques that we have used to generate phone-to-AL mapping. In Section 5, a detailed description of the experimental setup is given. In Section 6, the results and analysis of various experiments performed with articulatory features extracted using the proposed technique are given. In Section 7, application of the proposed technique in language with limited training data is described. Finally, Section 8 deals with conclusion.

2. Review of articulatory features for ASR

Initial attempts to use articulatory features in ASR are reported in Schmidbauer (1989a, 1989b), Elenius and Takács (1991), Eide et al. (1993), Deng and Sun (1994) and Erler and Freeman (1996). Recent studies reported that articulatory features extracted from neural networks fed into tandem HMM (Ellis et al., 2001) showed improvements in *recognition* performance (Cetin et al., 2007; Frankel et al., 2007; Kirchhoff et al., 2002). Kirchhoff et al. (2002) also showed that these features were robust to noise. Articulatory features in tandem HMM framework were further explored in Johns Hopkins 2006 summer workshop (Cetin et al., 2007; Frankel et al., 2007). In this paper, we have followed the articulatory feature set and feature extraction methodology outlined in the workshop.

2.1. Articulatory label set

In Frankel et al. (2007), a discrete multilevel label set with eight AL groups each having specific AL was introduced. This label set is given in Table 1. Each AL group has a 'none'; for example, in the case of Degree & Manner this class covers the non-speech sounds (silences), whereas in the case of Place (associated with consonants), it covers all vowels and non-speech. For diphthongs (e.g., /aw/), the begin state is denoted by /aw1/ and end state by /aw2/.

2.2. Review of articulatory feature extraction technique

Articulatory features are extracted from articulatory classifiers built for each of the eight AL groups listed in Table 1. Articulatory feature extraction is performed as per Algorithm 1 (Cetin et al., 2007). Consider a specific example of building an articulatory classifier for the AL group "Degree & Manner" as shown in Fig. 1. Given an acoustic feature as input, a multilayer perceptrons (MLP) is trained with six AL in "Degree & Manner" group as output targets. This requires the input acoustic features to be aligned at frame level with the six AL. Frankel et al. (2007) showed that manual transcription of data at frame level in terms of AL is laborious. Hence, the usual practice is to obtain a phone-level alignment (from an efficient acoustic model built in terms of phones) and convert it into AL using a phone-to-AL mapping. Articulatory