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

Assessment of pitch-adaptive front-end signal processing for children's speech recognition☆

Q1 Rohit Sinha, S. Shahnawazuddin*

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

Received 1 May 2016; received in revised form 13 October 2017; accepted 21 October 2017
Available online xxx

Abstract

On account of large acoustic mismatch, automatic speech recognition (ASR) systems trained using adults' speech data yield poor recognition performance when evaluated on children's speech data. Despite the use of common speaker normalization techniques like feature-space maximum likelihood regression (fMLLR) and vocal tract length normalization (VTLN), a significant gap remains between the recognition rates for matched and mismatched testing. Our earlier works have already highlighted the sensitivity of salient front-end features including the popular Mel-frequency cepstral coefficient (MFCC) to gross pitch variation across adult and child speakers. Motivated by that, in this work, we explore pitch-adaptive front-end signal processing in deriving the MFCC features to reduce the sensitivity to pitch variation. For this purpose, first an existing vocoder approach known as STRAIGHT spectral analysis is employed for obtaining the smoothed spectrum devoid of pitch harmonics. Secondly, a much simpler spectrum smoothing approach exploiting pitch adaptive-liferting is also presented. The proposed approach is noted to be less sensitive to errors in the pitch estimation than the STRAIGHT-based approach. Both these approaches result in significant improvements for children's mismatch ASR. The effectiveness of the proposed adaptive-liftering-based approach is also demonstrated in the context of acoustic modeling paradigms based on the subspace Gaussian mixture model (SGMM) and the deep neural network (DNN). Further, it has been shown that the effectiveness of existing speaker normalization techniques remain intact even with the use of proposed pitch-adaptive MFCCs, thus leading to additional gains.

Keywords: Children's ASR; Acoustic mismatch; Pitch-adaptive features; STRAIGHT MFCC; SGMM; DNN

1. Introduction

Automatic speech recognition (ASR) is a long standing area of research. Over the years, many applications involving human-machine interactions have evolved. A few example of these are reading tutors, language learning tools, entertainment and voice-based search (Hagen et al., 2003; Nisimura et al, 2004; Bell and Gustafson, 2007; Hagen et al., 2007; Gray et al., 2014; Schalkwyk et al., 2010). In these applications, the ASR systems are usually accessed by both adult and child speakers. The acoustic properties of adults' and children's speech differ

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.
* Corresponding author.
E-mail address: rsinha@iitg.ernet.in (R. Sinha), s.syed@iitg.ernet.in (S. Shahnawazuddin).

http://dx.doi.org/10.1016/j.csl.2017.10.007

substantially due to morphological and physiological differences. Consequently, achieving high recognition performance for both adult and child speakers on an ASR system becomes quite challenging. Traditionally, ASR systems employ Gaussian mixture model-based hidden Markov model (GMM-HMM) for acoustic modeling (Rabiner and Juang, 1993). The model parameters are typically learned on Mel-frequency cepstral coefficient (MFCC) (Davis and Mermelstein, 1980) based front-end parameterization of speech signals. The MFCC is a popular feature in ASR and is derived following static (non-signal dependent) processing techniques.

In the past decade, a number of advancements have taken place in the ASR domain. In particular, these include the normalization of the acoustic features prior to modeling by applying a set of linear transformations (Digalakis et al., 1995; Gales, 1999; Rath et al., 2013) and, the use of subspace Gaussian mixture model (SGMM) (Povey et al., 2011a) and deep neural networks (DNN) (Hinton et al., 2012; Dahl et al., 2012) based alternate paradigms towards achieving more robustness in the acoustic modeling. Despite the recent advances in acoustic modeling, a large gap in the recognition performance can still be noticed when children's speech is recognized on adults' speech trained ASR system or vice versa. Commonly separate ASR systems, one trained using adults' speech and the other using children's speech, are employed to address this large gap in performances. As the collection of speech data from child speakers is tedious, the data available for system development is scarce. Thus, due to paucity of the data, the complex modeling techniques are usually not well supported while creating children's ASR system. To overcome these challenges, more research is needed for effective addressal of different sources of acoustic mismatch between adults' and children's speech. The work presented in this paper forms one step in that direction.

In the following, we briefly review existing works in the area of children's speech recognition and highlight the challenges involved. For the size of the vocal organs in children being much smaller compared to adults, children's speech is characterized by higher fundamental and formant frequencies (Eguchi and Hirsh, 1969; Kent, 1976). During the growing phase, a child's speech undergoes considerable variation along with improvement in his/her ability to correctly pronounce the complex words (Russell and D'Arcy, 2007). In addition to that, the overall speaking rate is slower in the case of children and they have more variability in the speaking rate as well (Potaminaos and Narayanan, 2003). Children's speech is reported to have greater values for the mean and the variance for the acoustic correlates of speech than those for adults' speech (Ghai and Sinha, 2010b). For example, as observed in Eguchi and Hirsh (1969), for most of the vowels, the area of the F1-F2 formant ellipses is larger in the case of children than for adults. Consequently, children's speech suffers from a higher degree of inter- and intra-speaker acoustic variability in contrast to adults' speech (Potaminaos and Narayanan, 2003; Gerosa et al., 2007). All these factors contribute towards making automatic recognition of children's speech a much tougher problem (Lee et al., 1999; Narayanan and Potamianos, 2002; Potaminaos and Narayanan, 2003; Gerosa et al., 2009; Shivakumar et al., 2014). From the linguistic perspective, children are more likely to use imaginative words, ungrammatical phrases and incorrect pronunciations as discussed in Gray et al. (2014). Recently, some works have explored DNN-based acoustic modeling as well as the inclusion of vocal tract length normalization (VTLN) for children's ASR (Serizel and Giuliani, 2014; Metallinou and Cheng, 2014; Liao et al., 2015; Serizel and Giuliani, 2016). Differences in the acoustic and linguistic correlates of speech from adult and child speakers have been observed to affect the performances of speaker recognition tasks as well (Safavi et al., 2012; 2014). Both of these works employed MFCC features for front-end speech parameterization while classification was performed using GMM-UBM and GMM-SVM paradigms. The speaker verification equal error rate for children was reported to be nearly four times worse than that for adults. Furthermore, several works on the analysis of child language environment and child verbal communication have been reported recently (Najafian et al., 2016a; Najafian and Hansen, 2016; Najafian et al., 2016b). These works have highlighted the fact that whenever a system is supposed to work with children's speech, there is a high level of background noise from other children or tools in the ambient environment. For the sensitivity of ASR systems to background noise, the recognition performance of children's ASR system gets severely affected. As suggested in the referenced works, one way to address this is to use a fused speech activity detector with a diarization unit that can separate the target child speech from other children/adults or other non-speech events. The relevance of such studies cannot be understated for children's ASR, but fall beyond the scope of this work.

It's evident from the above discussions that a major source of acoustic mismatch between adult and child speakers lies in gross differences in the length of their vocal tracts. This explains why a large improvement is noted with the inclusion of VTLN (Lee and Rose, 1998) in the case of children's mismatched ASR (Ghai and Sinha, 2010b; Serizel and Giuliani, 2016). In addition, the acoustic mismatch is also caused by differences in fundamental frequency or pitch across these two group of speakers. A number studies have already been reported for addressing the pitch

Please cite this article as: R. Sinha, S. Shahnawazuddin, Assessment of pitch-adaptive front-end signal processing for children's speech recognition, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.10.007