Contemporary Engineering Sciences, Vol. 4, 2011, no. 5, 221 – 228

HMM-Based Recognition and Adaptation of Persian Children's Speech

Ghamarnaz Tadayon Tabrizi

Department of Computer, Science and Research Branch,
Islamic Azad University, Tehran, Iran
tadayon@mshdiau.ac.ir

Saeed Setayeshi

Amir-Kabir University of Technology, Tehran, Iran
setayesh@cic.aut.ac.ir

Mohammad Molavi Kakhki

Ferdowsi University of Mashad, Mashad, Iran
kakhkimol@ferdowsi.um.ac.ir

Abstract

There are high variability in children's speech compared to adults' which is mainly because of their shorter vocal tract length and smaller vocal fold which results in lower accuracy in speech recognition task (about 54.5% in this work). Therefore using adaptation techinques which reduce these variabilities has been suggested.
In this paper we focused on the problem of speech recognition for Persian children using adaptation techniques performed on two models: one trained on children's speech and one on adults'. We used a speaker normalization method which is combination of vocal tract length normalization and model adaptation. Experiments have shown that using adult model has low performance which inreases 37% when using adaptation techniques. It is also shown that using these

techniques will increase recognition rate by 7% when using recognizer trained on children's speech.

Keywords: Speech recognition, model adaptation, VTLN

1 Introduction

Researchs on methods for children's speech recognition have been noticed in different fields such as education, game playing, communication and even in some medical services like hearing and speaking tests[5]. Previous studies have shown that when an automatic speech recognition system trained on adult speech is used to recognize children's speech, the error rate increases significantly[10]. The acoustic parameters of speech change with age, it means that acoustic and linguistic characteristics of children's speech such as pitch and formant frequencies are higher than adults' speech[4]. These differences cause performance decrease when using recognizer which is trained with adults' speech. Wilpon and Jacobson showed that the error rate of a speech recognizer trained with data from speakers of all ages increases when testing with speech data from children which are 12 years old or younger[10]. Thus when designing a children speech recognizer system, we should collect and use adequate amount of training children data in order to train age-specific acoustic models. Howerver, even in this case recognition error rate reported for children is usually significantly higher than that for adults which decreases by increasing children's age [2].
To develop a high performance speech recognizer for children, some techniques have been offered by researchers such as vocal tract length normalization (VTLN), speaker independent linear frequency warping, MLLR and Constrained MLLR [2,3,6,7].

In this paper, recognition of Persian children's speech is concerned as a digit recognition task. Using two Persian speech databases, we designed two speech recognition systems, one trained for children and one for adults. Experiments were performed under two conditions: Matched condition when both training and testing data are from the same age group; and mismatched condition in which children's speech is used for testing adults' model.

The structure of the rest of this paper is as follows. Section 2 briefly introduces acoustic characters of children's speech. In section 3, Adaptation techniques which are used in this work, are explained. The experiments and proposed models are introduced in section 4. Section 5 reports the results of these experiments. Finally, concluding remarks are given in section 6.

2 Acoustic characteristics of children's speech

There are some systematic age-dependent variations in characteristics of children's speech which make the recognition task more complicate[4,10]:
1) The most important one is physical size. There are anatomical differences in the vocal tract length and vocal folds which make position of children's speech formants and fundamental frequency (F0) higher compared to adults'. Moreover, small size of glottis, makes the average pitch of children's voice lower.
2) Spectral and temporal variability in children's speech results in greater overlap among phonemic classes for children than for adult speakers and make classification problem more difficult.
3) Uncorrect pronunciation, disfluencies, breath noise and losing front teeth (around age six) make difficulties in recognition.
4) Different use of language and different dictionary.
5) Duration of some vowels is longer and more variable for young children than for adults.

3 Adaptation Techniques

Adaptation techniques, change the acoustical models which enhances the match between different groups of speech data. This is mainly necessary because we rarely have enough data to train on a specific speaker. In this work we used combination of two approaches: Vocal tract length Normalization (VTLN) and Maximum Likelihood Linear Regression (MLLR).

3-1 Vocal tract length normalization

As mentioned before, there are some differences between acoustic characteristics of adults and children's speech. As children have shorter vocal tract length and smeller vocal folds, their voice has higher pitch and the fundamental and formant frequencies of their speech are higher compared to adults'. One of the useful normalization techniques that can be used in feature extraction part is to control for differing vocal tract length of speakers by VTLN which is a linear or bilinear scaling of the frequency axis [6,7]. This method concernes vocal tract length of different speakers and tries to reduces this variability. In order to match children's speech with adults', VTLN warps the spectral envelope by some factor (group or speaker-specific), which can be done during training or recognition phases. During training, frequency warping can be used to create new acoustic models which are normalized based on specific speaker or group. In recognition phase, warping functions are used when computing feature vectors. There are several methods for VTLN. In this work we used a nonlinear feature transformation approach based one the one discussed in [6].
To represent the Mel-curve, following equation has been used:

$$mel(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

The method contains following steps:

1- The following equations are used to compute n sampling points which will be used to adjust the curve to the properties of the data:

$$f_i = mel^{-1}\left((i-1).\frac{f_{max}}{n-1}\right), i \in 1,\dots,n \quad (2)$$

$$p(f_i) = mel(f_i), i \in 1,\dots,n \quad (3)$$

2- The following properties are used to adjust $p(f_i)$:
(a) Interpolating cubic spline curve opt(f), are computed for sampling points.
(b) Acoustic features are obtaind using the Filterbank which is determined by opt(f),
(d) Recognition rate is computed using Gaussian classifier.
(f) Use simplex algorithm (which is used because it's not needed to compute a gradient).

3- Optimal opt(f) is returned.

3-2 Speaker adaptation

There are some well known approaches for performing adaptation: Maximum likelihood linear regression (MLLR), Maximum a posteriori (MAP) and speaker clustering/speaker space. In this work we used the first method [3].
MLLR is a linear adaptation technique which can be used even when there is just small amount of adaptation data that may be as small as 10 seconds of speech. It aims at reducing HMM's parameter variability among different speakers using some linear transformations to clusters of acoustic units, changing means and variances of the Gaussian mixtures. These transformations try to optimize the likelihood of the transformed model with adaptatin data. In this method common transformations are applied for each cluster and all models (even not seen in adaptation data) can be adapted.

4 Experimental design

4-1 Speech databases

To design the system, two different speech datasets including digit strings were used; one consisted of adults' speech and one with children's speech. There are many difficulties in recording children's speech as they cannot read from a prepared text and we cannot easily make them to repeat a word. Also, the shy and spoke in a very soft, breathy voice.
Children recordings which are used in this work, were made with 16 bits and 32 KHz via a headset michrophone in a sperated room with almost no noise. Two sets of 42 children's speech were prepared (7 children of both genders in each age group 3 to 8), each child uttering 30 digit strings. Adult speech was gathered from the existing databases in AI Lab in Azad University of Mashhad. The adults'

speech data was divided into two different train and test groups.

4-2 Features of speech

As we use mel scale frequency warping, it's better to use mel frequency cepstral coefficients (MFCC) rather than LPC to get better recognition performance [4]. In our work for each frame, the parametric representation was consisted of 12 MFCCs plus energy extracted from a 24-channel, 16 KHz bandwidth mel-scale filterbank. The first and second derivatives of the coefficients were added which produced 36-element acoustic feature vector.

4-3 Implementation

At present, the most popular approaches to design speech recognition systems use statistical methods based on Hidden Marcov Models (HMM) [9]. To use these techniques we need suitable, large speech datasets. Unfortunately there is no standard speech database for Persian children, so recordings of children were taken from our own database.
The recognizer iss based on HMMs. A standard left to right model is used. For considering pronunciation variation, a transition was added to skip each state. We used 10 models trained using Maximum Likelihood Estimation (MLE) techniques to represent Persian digits which had 12-18 states with 8 Gaussian mixtures per state. The recognizer uses modules from HTK toolkit for speech recognition processing.
Two recognizes are designed[8]:

1- Recognizer trained using children's speech.
2- Recognizer trained using adults' speech.

Experimental results are shown in following section.

5 Results

To show the performance of each model , we used two test dataset , one for

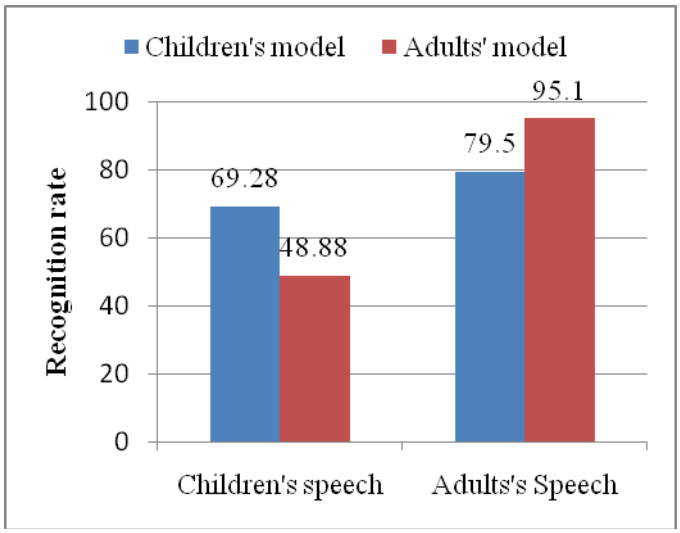

Figure 1. Recognition rate using two models, one trained with children's speech and one with adults' when testing by children and adults dataset.

children and one for adults.

The results show that there is a significant improvement in recognition accuracy when using children model to recognize children speech. More specificly, the best results are obtained when using matched model. Figure 1 shows word recognition rate obtained by four experiments.

To examine the effect of children gender on recognition rate, children dataset was devided in two categories, girls and boys. Recognition results were obtained using two models, which results are shown in figure 2.

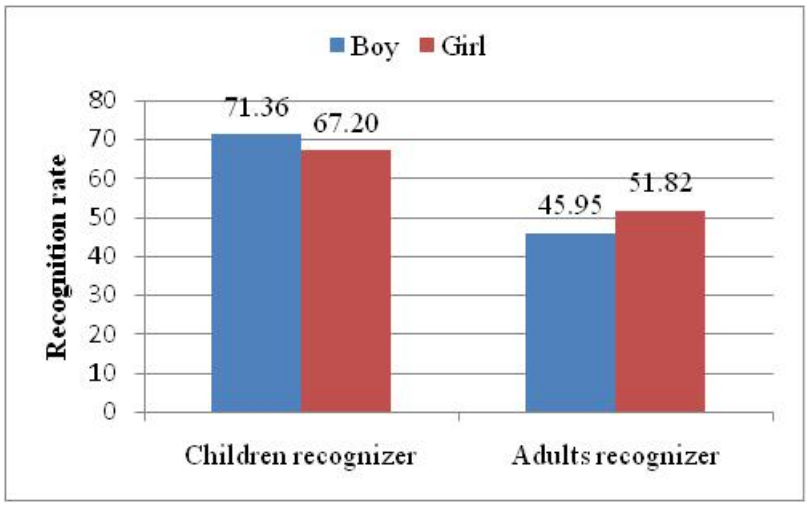

Figure 2. recognition rate using two acousic models for different genders.

Results indicate that there is no significant difference between recognition rate according to gender as the acoustic features are similar in younger children.

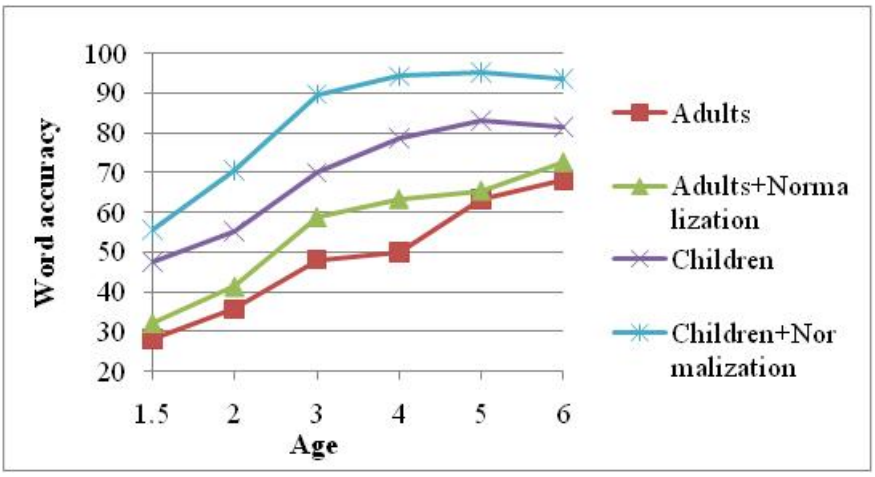

Figure 3. Children's speech recognition results achieved using four different acoustic models

As shown by previous results, the recognition performance is not acceptable, even in matched condition for young children. As disscussed before, using adaptation techniques can be usefull to improve recognition rate.
Figure 3 shows the recognition accuracy obtained for each of acoustic models in case of matched and mismatched training and testing conditions with and without normalization as a function of age. We note that the baseline system for adults is practically unsuitable to recognize speech of young children. As expected, using normalization techniques allows a better recognition of children's speech.

6 Conclusion

The performance level of persian children's speech when using adults' models is too low to be practically useful. This was improved significantly by adaptation and vocal tract lensth normalization but not to the same level as training on children. It was also observed that the variability in performance for young children was larger than for older children. We obtained better accuracy using adaptation techniques on age spesific recognizer, which is camparable to adult recognizers.

References

[1] D. Elenius, Adaptation Techniques for children's speech recognition, report, KTH/TMH, Stockholm, Sweden, 2004.

[2] M.Gerosa, D. Giuliani and F. Brugnara, Speaker adaptive acoustic modeling with mixture of adult and children's speech, In Proc. European conference on speech communication and technology [INTERSPEECH2005], Lisbon, Portugal, Sep, 2005, 2193-2196.
[3] C. Leggetter and P. C. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 1995, 171–185.
[4] A. Potamianos and S. Narayanan, Acoustics of children's speech: Developmental changes of temporal and spectral parameters, Journal of Acoust. Soc. Amer, 105 (1999), 1455–1468.
[5] L. Smith, and T. Pham, Toward voice applications for children, International Journal of Speech Technology, 5 (2002), 321-329.
[6] G. Stemmer, C. Hacker, S. Steidl and E. Noth, Acoustic Normalization of Children's Speech, Proc. EUROSPEECH, Geneva , Switzerland, Sep, 2003, 1313-1316.
[7] G. Tadayon Tabrizi, S. Setayeshi and M. Molavi Kakhki, Applying acoustic normalization to improve recognition of children's speech, Proc. 16th Iranian conference on electrical engineering, Tehran, Iran, May, 2008 (In Persian).
[8] G. Tadayon Tabrizi, S. Setayeshi, Improving Speech Recognition of Persian Children Based on Normalization and Clustering, Accepted in: Journal of Technical-Engineering, To appear in: vol 5, 2011 (In Persian).
[9] O. Watts, J. Yamagishi, K. Berkling, and S. King, HMM-based synthesis of child speech, Proc. 1st Workshop on Child, Computer and Interaction (ICMI'08 post-conference workshop), Crete, Greece, October, 2008.
[10] J. G. Wilpon, C. N. Jacobsen, A study of speech recognition for children and the elderly, Proc. ICASSP-96. IEEE international conference on acoustics, speech, and signal processing, Atlanta, USA, 1 (1996).

Received: February, 2011