SOME ISSUES ON THE STUDY OF VOCAL TRACT NORMALIZATION

Zhuo WANG, Peng DING, Bo XU

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing 100080
{zwang,pding,xubo}@nlpr.ia.ac.cn

ABSTRACT
Vocal tract normalization (VTN) is an effective way to reduce inter-speaker variability mainly caused by variation of vocal tract shape among speakers of different genders and age groups. In this paper, some practical implementation issues of VTN are discussed. We adopted a method to train model and selected the proper normalization scales of different speakers. The acoustic model is estimated from the unnormalized acoustic vectors of large speakers by maximum likelihood training. Then we use the gender-independent model to select the proper normalization scales of different speaker. The above steps are repeated. For VTN in training, we discussed with the drift effect of the warp parameter with the increasing of the number of iterations and the number of mixtures of the acoustic model. We studied the distribution of the warp parameter of different genders and age groups. To facilitate the fast warp parameter selection process, we proposed a hierarchical method and compared with the traditional methods.

I. INTRODUCTION
The vocal tract length can vary from approximately 12 cm for adult females to 20 cm for adult males. For child, the vocal tract length will be shorter. The variation of vocal tract length results in a significant degradation from speak dependent to speak independent speech recognition performance. Because the variation of vocal tract length can be thought as a simple linear warping in the frequency domain of the speech signal, some so called vocal tract normalization (VTN) methods are proposed to eliminate the affection of distortion of vocal tract length. The motivation of VTN is that the positions of spectral formants are inversely proportional to the length of the vocal tract.

In the usually adopted maximum likelihood based speaker normalization method [1], a parameter is introduced to warp the speaker’s spectrum to map on that of a “standard” speaker. Selection of the warp parameter is a tentative process in which the warp parameter varies from 0.88 to 1.12 with the step 0.02, if take children into consideration, the floor level may be lower. Of course, the “standard” speaker depends on the parameters of the acoustic model. In this paper, we studied the procedure for selecting the proper normalization scales of different speakers, discussed the convergence of normalization scales and the affection of number of mixtures of the acoustic model to the drift of the warp parameter, proposed a method to select the proper normalization scales quickly.

The paper is organized as follows: The next section presents the description of the large database including the number of speakers of different ages and genders. Section III we will talk about the training of a normalization acoustic model using VTN. In this section, we present the detail description of procedures for selecting the proper normalization scales of different speakers, the method of selecting the proper normalization scales of different speakers, the training of model based on normalized vectors, the distribution of normalization scales, the relation between the drift of normalization scales and the number of mixture with the increasing number of iterations. In section IV, we present a new hierarchical method to fast the selection the proper normalization scales of different speakers. At the last, we compare this method to the traditional method.

II. EXPERIMENTAL CONDITIONS
We use the following database: 863-database (83 male speakers, 83 female speakers, 520 sentences each speaker, total 88920 sentences), Beijing database, the data is recorded in Beijing city. (150 male speakers, 150 female speakers, 250 sentences each speaker, total 75000 sentences). Science museum Female 62960 491 277 19

Figure 1: description of database
museum (513 male speakers, 787 female speakers, 80 sentences each speaker, total 104000 sentences of different ages). Just as Figure 1 shows:

III A FREQUENCY WARPING FACTOR SELECTION APPROACH

In this section, we will describe the procedures that used to get the normalization scales of different speakers. The method attempts to reduce the affection of variation of vocal tract length and warp the speaker’s spectrum to map on that of a standard speaker. Of course the “standard” depend on the parameters of gender- independent HMM trained by all training speakers.

This section is organized as follows: In section III-A, we will present the process of estimating the normalization scales. In section III-B, we will present the distribution of normalization scales based on large population of speakers. In section III-C, we will discuss the relation between drifts of normalization scales and the number of mixtures of acoustic model when the number of iterations increased. The convergence of training will be tested by experiment.

A. Process of estimating the normalization scales

The process of estimating the normalization scales is divided into four steps:

1. We should get the first gender-independent acoustic model trained by all the speakers listed in figure 1. We use a male acoustic model (the number of output is 2733, the number of triphone is 13307, per output is represented by 16 Gaussian densities) to extract the 42 dimensions MFCC (including energy and pitch) of all male speakers. We use a female acoustic model (the number of output is 2923, the number of triphone is 13702, per output is represented by 16 Gaussian densities) to extract that of all female speakers. Then we established the decision tree and produced the first gender-independent acoustic model based on these unnormalized acoustic vectors by maximum likelihood training. The total data used to train the model is about 267920 sentences.

2. For each speaker, we should find the proper normalization scales. We select 15 sentences from each speaker and let the normalization scales $\alpha$ changed from 0.80 to 1.14 with the step size 0.02. For each sentence, the selection of $\alpha$ made the likelihood excluding silence maximum. The $\alpha$ that showed most often is used to represent the ratio between the speaker’s vocal tract length and the “standard”:

$$\alpha_{\tau} = \arg \max_{\alpha} P_{\tau}(X^\alpha_{\tau} / W, M)$$  \hspace{1cm} (1)

$W$ is the word sequence of sentence. $\alpha_{\tau}$ is the proper normalization scales of this sentence that make the likelihood maximum. $M$ is the first gender-independent acoustic model. $X^\alpha_{\tau}$ is the sequence of the warped feature vectors of this sentence.

3. We will train the normalized gender-independent acoustic model based on normalized vectors in step 2 by maximum likelihood training. We control the threshold of the decision tree to let the number of outputs about 2500 and per output are represented by 16 Gaussian densities.

4. We change the number of Gaussian densities and repeat the step 1, 2, 3. We analyzed the relation between the number of Gaussian densities and the drift of normalization scales.

B. Distribution of normalization scales

Figure 2: distribution of male normalization scales

Figure 3: distribution of female normalization scales
We divided all the speakers into three groups according to their age. The age of the first group are from 0 to 18. The age of the second group is between 19 and 40. The others belong to the third groups. From the Figure 1 and the Figure 2, the distribution of every group is approximately a Gaussian function. The centers of the male speakers of different groups are about 0.88, 1.02, and 1.08. The centers of female speakers of different groups are about 0.84, 0.90, and 0.94. We can know that the means of normalization scales of the male speakers is always higher than those of the female speakers when the speakers are adult. The result can explain that the vocal tract length of female speakers is shorter than that of the male speakers, that means the formant of the female speakers is higher. The distribution of normalization scales of the female speakers is similar to that of the male speakers when the age of speakers below 18. According to the commonplace of physics, this is reasonable.

c. Drift of normalization scales and convergence

The training of vocal tract normalization is, essentially, a transform in feature space. While the iterative normalized HMM training procedure is not guaranteed to converge mathematically [1], and sometimes the warp parameters will drift [2], a phenomena caused by feature space shrinking [4]. In this section, we first study the convergence of vocal tract normalization by observing percentage of speakers that can get the proper normalization scales in first three sentences as the number of training iterations increased. It is highly important in the practical implementation of vocal tract normalization. Then, we will discuss with the drift effect of the warp parameter with the increasing of the number of iterations and the number of mixtures of acoustic model.

In Figure 4, three important observations can be achieved. The first is: with the increasing of the number of iterations, the number of speakers that can get the proper normalization scales in first three sentences increased. The second is: when the number of training is above three, the result is approximately the same. The third is: at the third times of training, the percentage that can get the proper normalization scales in first three sentences is about 83.7%, if consider four sentences, it is about 91.3%. So in recognition process, we can use three sentences to select the proper normalization scales of each speaker.

In the large database listed in Figure 1, we selected 80 speakers (the male speakers is 40) from Beijing database and do the same thing from 863 database. We selected the speakers whose age is below 15 or above 31 from science museum database because the most speakers of Beijing and 863 are adult. We use these speakers to study the drift of normalization scales. We resumed that the normalization scales of each speaker made by the first gender-independent model based on unnormalized acoustic vectors by maximum likelihood training is “standard” and the threshold of changing is 0.04. That means if the warping factor changed from 1.02 to 1.04, we thought the drift didn’t occur.

The Figure 5 shows that with the increasing of the number of mixtures of acoustic model, the number of speakers whose normalization scale drifted decreased. If we let the threshold of changing is 0.02, the number of speakers whose normalization scale drifted will increase but the trend is the same.

IV HIERARCHICAL METHOD FOR FAST VTN

In [2], the author present a fast method to select the normalization scales in recognition. You should calculate the score of every normalization scales from 0.80 to 1.14 with the step size 0.02. The number of calculations is about 18. Now, we present a new fast method for the selection of proper normalization scales based on some Gaussian mixture model that represent the distribution of the unnormalized feature vectors of different groups of age. By employing the LBG algorithm and the maximum likelihood criterion, we produced the Gaussian mixture model based on the
unnormalized feature vectors that produced the first gender-independent acoustic model. We totally established 18 Gaussian mixture models, each model represent the distribution of the unnormalized feature vectors that belong to a normalization scale. In Figure 2 and Figure 3, we know the distribution of every group is approximately a Gaussian function. So we consider whether we can establish hierarchical model. In the first level, there are three Gaussian mixture models based on unnormalized feature vectors belong to different groups of normalization scales. In the second level, there are some Gaussian mixture models that belong to normalization scales of this group.

For this method, 64 component densities compose the Gaussian mixture models. For each speaker, we select 4 sentences to find the proper normalization scale of this speaker. For each sentence, we first calculate the score of the feature vectors excluding the silence in the first level of the method. Then we use the second level models of this method according to result of the first level to select the proper normalization scales. The warping factor that shows most often is the last result of this speaker.

We still resume that the normalization scale of each speaker that made by the first gender-independent model based on unnormalized acoustic vectors by maximum likelihood training is “standard” and do vocal tract normalization with this method for the same speakers, the error ratio is 4.82%, consider the improvement of calculation speed, the method is effective.

V. SUMMARY

In this paper, we made some research on vocal tract normalization and get some contribution.

First we have large population database of different ages. Based on it, we can study the distribution of normalization scales of different age groups.

Second we study the convergence of the training and the relation between the drift of normalization scales and the number of mixtures of the acoustic model. From the experiment result, we know the convergence of the training can be guaranteed and with the increasing of number of iterations, the effect of smaller number of mixtures of the acoustic model is better.

Third based on the method of reference [3], we presented a new hierarchical fast selection of warping factor, the experiment shows this method has low error ratio and improve the calculation speed.

All above studies put emphasis on practical issues of the implementation of vocal tract normalization techniques. Since there is hardly any mathematical deduction for them, studies are performed based on various experiments.

REFERENCES