ROBUST SPEECH RECOGNITION BASED ON SPECTRAL ADJUSTING AND WARPING

Rui Zhao and Zuoying Wang

Department of Electronics Engineering, Tsinghua University, Beijing 100084, P.R.China
zr@thsp.ee.tsinghua.edu.cn

ABSTRACT

In this paper, we first propose a new channel adaptation method named spectral adjusting (SA) which adjusts the amplitude spectrum of the channel distorted speech with an adjusting function to reduce the channel distortion. Then, we combine Vocal Tract Length Normalization (VTLN), which warps the frequency scale of the speech spectrum to do speaker normalization, with SA to adjust and warp the speech spectrum. So the channel and speaker variations can be compensated for together. We call the combined method spectral adjusting and warping (SAW). In SA method, the adjusting function is approximated by a piece-wise linear function, and the parameters of the piece-wise linear function are estimated by Gradient Projection algorithm with short adaptation utterances based on ML rule. The evaluating experiments were carried out on telephone speech recognition in Duration Distribution Based HMM (DDBHMM) system. Experimental results showed that SA yielded a relative error rate reduction of 10.44% over the baseline, and SAW led to a greater reduction of 14.6%.

1. INTRODUCTION

One major source of degradation in speech recognition performance is channel mismatch. A communication channel or a transducer is a multiplying factor to the speech signal in the spectral domain. In the log spectral and the cepstral domain, it acts as additive bias.

There have been considerable interests in handling channel or convoluted noise, and recent approaches are focused on removal of cepstral bias. In this kind of method，the channel is modeled as an additive bias vector in cepstral domain. The bias is estimated and subtracted from the distorted speech cepstral. In the commonly-used cepstral mean subtraction (CMS) method, the bias is simply the mean of the cepstral of the utterance. In signal bias removal (SBR) method [2, 3], the bias is estimated based on ML (SBR_ML) or MAP (SBR_MAP) rule. Since the cepstral features are commonly used in recent speech recognition research, the cepstral bias removal (CBR) method can be integrated to the speech recognition process without much additional calculation. However, for some cepstral features, e.g. MFCC (Mel-Frequency Cepstrum Coefficients), the channel mismatch can't be simply expressed as additive distortion in feature domain. CBR method may be not proper for this kind of features.

Besides, Zhao has assumed that the distortion channel can be modeled by an FIR filter in [4]. The parameters of the filter are estimated by EM algorithm in spectral domain. We called this method FIR_EM. Although modeling the channel more precisely than CBR, FIR_EM method brings complex calculation in spectral domain.

The proposed spectral adjusting (SA) method is different from them. In this method, a channel adjusting function is defined in spectral domain, which adjusts the amplitude spectrum of the channel distorted speech to the "standard" spectrum when multiplied by the former. The adjusting function is approximated by a piece-wise linear function. Since the spectral characteristic of any channel or transducer in the real world varies smoothly along the frequency scale, it's a good approximation just with small number of pieces, which means a few parameters to be estimated.

For MFCC feature, the proposed method is supposed to be more effective than CBR because it deals with the channel distortion in spectral domain directly. On the other hand, the parameters of the adjusting function are estimated by Gradient Projection algorithm in MFCC feature domain but not in spectral domain, which, compared with FIR_EM method, reduces the calculation.

Besides channel mismatch, the variance in speakers is another source of performance degradation in speech recognition. VTLN is a widely used speaker normalization method which warps the frequency scale of the speech spectrum. Since both SA and VTLN deal with the speech spectrum, SA adjusts the spectrum amplitude, VTLN warps the frequency scale. We combine SA and VTLN to adjust and warp the speech spectrum, aiming to compensate for the channel and speaker variations together. We called this combination method as spectral adjusting and warping (SAW).

The proposed methods were evaluated on the experiments of telephone speech recognition in DDBHMM [1] system. In the experiments, SA and SAW achieved 10.44% and 14.6% relative syllable error rate reduction over the baseline, while CMS, SBR_ML and VTLN led to the reduction of 8.48% ,6.16% and 8.38%. The results showed that SA performed better than CMS and SBR_ML, and SAW is more effective than either SA or VTLN alone.

This paper is organized as follows. Section 2 gives the principle of SA in detail. In Section 3, the method of SAW as well as a brief introduction of VTLN is presented. In Section 4, the experiments and results are provided. Finally, we conclude the paper in Section 5.

2. SPECTRAL ADJUSTING

2.1 Theory analysis

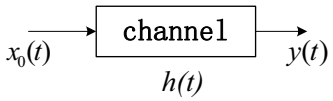

Figure 1 Speech transmission

Figure 1 illustrates the effect of a transmission channel on the speech.

$x_0(t)$ denotes the original speech, $h(t)$ is the impulse response of the transmission channel and $y(t)$ the output speech. $X_0(\omega)$, $H(\omega)$ and $Y(\omega)$ are their spectrum respectively.

In time domain: $$y(t) = x_0(t) * h(t) \quad (1)$$

In spectral domain: $$Y(\omega) = X_0(\omega)H(\omega) \quad (2)$$

Equation (2) shows that channel acts as a multiplying factor to speech in spectral domain. Spectral adjusting is based on this fact. In the following paragraphs, we specify the principle.

For a special channel S with frequency response $H_s(\omega)$, let $x_0(t)$ be the original speech, $y_s(t)$ the channel S's output and $X_0(\omega), Y_s(\omega)$ their spectrum, then we have

$$|X_0(\omega)||H_s(\omega)| = |Y_s(\omega)| \quad (3)$$

and

$$|X_0(\omega)| = |Y_s(\omega)| * 1/|H_s(\omega)| \quad (4)$$

That is, the distorted speech spectrum can be adjusted to the original speech spectrum when multiplied by the factor $1/|H_s(\omega)|$. Notice that the purpose of channel adaptation is to reduce the channel mismatch between the training and the testing data. If we adjust both the training and the testing speech amplitude spectrum to the "standard" spectrum, the mismatch can be reduced. Therefore, for the given channel S, an adjusting function $F_s(\omega)$ is defined to satisfy:

$$F_s(\omega)|Y_s(\omega)| = Y_N(\omega) \quad (5)$$

Where $Y_N(\omega)$ is the "standard" spectrum.

If the adjusting function is found, the channel distorted spectrum can be adjusted to the "standard" spectrum by (5). If both the training and testing speech is adjusted in this way, the mismatch caused by channel variation can be reduced. Furthermore, the feature extracted from the adjusted spectrum is robust against channel distortions.

Distinctly, the key point of this method is the mathematic form of the adjusting function $F_s(\omega)$. In our work, the function is approximated by a piece-wise linear function as:

$$F_s(\omega) = \sum_{l=1}^{L} a_l^s h_l(\omega) \quad a_l^s \geq 0 \ (l = 1,2,...,L) \quad (6)$$

$h_l(\omega)$ $(l = 1,2,...,L)$ are triangle shaped overlapping filters spaced uniformly on frequency scale. The overlap percent is 50 and each filter's height is 1. $\mathbf{a^s} = (a_1^s, a_2^s, ..., a_L^s)$ is nonnegative adjusting factor due to the property of amplitude spectrum (shown in equation (5)). Figure 2 shows the approximation of $F_s(\omega)$.

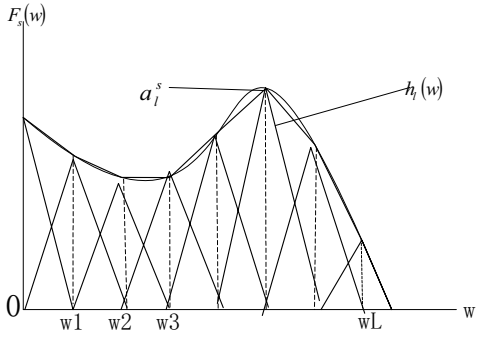

Figure 2 Piece-wise linear interpolation of $F_s(\omega)$

As shown in equation (6), once a a^s is obtained, we could get a certain $F_s(\omega)$. Here a^s is estimated by minimizing objective function $J(a^s)$.

$$\min \ J(a^s) \quad s.t. \ a_l^s \geq 0 \ (l = 1,2,...,L) \quad (7)$$

$J(a^s)$ can be defined as the difference between $F_s(\omega)Y_s(\omega)$ and $Y_N(\omega)$.

$$J(\alpha^s) = diff(F_s(\omega)Y_s(\omega), Y_N(\omega))$$

For speech recognition, equation (7) is rewritten based on ML rule as:

$$\alpha^* = \arg\max_a P(O_{a^s}(t) | S(t), \Lambda, W) \quad a_l \geq 0 \ (l = 1,2,...,L) \quad (8)$$

Where $O_{a^s}(t)$ is the speech feature vector sequence of the output utterance from channel S which is the function of the adjusting factor a^s. W is the transcription, Λ the parameter set of acoustic models and $S(t)$ the state segment. Gradient projection method is used to estimate a^s from equation (8).

It should be noted that for a given channel, the adjusting function can be estimated with short utterance and then be applied to other utterances output from the same channel.

2.2. Training and testing procedure

The goal of the training procedure is to appropriately adjust the amplitude spectrum of the utterances for each channel in the training set. An iterative procedure is used to choose the best adjusting factor for each channel and build the acoustic model using the adjusted utterances. The procedure is described as follows:

1 Train the acoustic model Λ_0 using unadjusted utterances, set j = 0.

2 Estimate the best adjusting factor a^* for each channel via equation (8) with model Λ_j, and then adjust the utterances with a^*.

3 Train the new model Λ_{j+1} using adjusted utterances.

4 If not convergence, set j=j+1, go to step 2; otherwise, stop.

The convergence is judged by whether there is significant difference in the adjusting factors between two iterations. Typically, this procedure converges after about 3 iterations. The final model is defined as Λ_N.

The goal of the testing process is to adjust the spectrum of the testing utterance to match Λ_N. Note that the transcription of adaptation data is not given, the recognition must be done firstly to get the transcription before the adjusting. The following process is used.

1 Get the transcription of the adaptation utterance W and the state segment $S(t)$ by recognizing the utterance with Λ_N.

2 get the best adjusting factor a^* with W and $S(t)$ via equation (8).

3 Adjust all the utterances from the same channel with adjusting factor a^* and recognize them.

3. SPECTRAL ADJUSTING AND WARPING

3.1 VTLN

Vocal Tract Length Normalization (VTLN) [5-9] is an approach aiming to reduce the variance among speakers. Previous research has shown that the positions of spectral formant peaks for utterances of a given sound are inversely proportional to the vocal tract length. Therefore, vocal tract length normalization should yield formants which have less variability.

An intuitive method of VTLN is to warp the spectrum in frequency axis.

$$\omega' = g(\omega)$$

This kind of VTLN is also referred as frequency warping (FWP) in [7]. The common choices of warping functions are linear, piecewise linear and bilinear functions.

Linear function:

$$g(\omega) = a^{-1}\omega \tag{9}$$

Piecewise linear function

$$g(\omega) = \begin{cases} a^{-1}\omega & if\ \omega < \omega_0 \\ b\omega + c & if\ \omega \geq \omega_0 \end{cases} \tag{10}$$

Bilinear function

$$g(\omega) = \omega + 2\tan^{-1}\left(\frac{(1-a)\sin(\omega)}{1-(1-a)\cos(\omega)}\right) \tag{11}$$

The warping factor a can be estimated based on two methods: formant-based [6,8,9] and ML-based [5,7]. In formant-based method, a is chosen to normalize the formant positions in the spectrum of testing utterances to those in the "standard" spectrum. The main drawback of this method lies in the difficulty of estimating the correct formant positions. In ML-based method, a is searched from a discrete set of possible values to maximize the likelihood of the warped utterance with regard to the given acoustic model and the transcription. This approach is inefficient due to the exhaustive grid search.

3.2 Spectral adjusting and warping

In spectral adjusting and warping (SAW), we combine SA and VTLN to adjust and warp the speech spectrum, SA is used for adjusting the amplitude spectrum to reduce the channel mismatch, VTLN is used for warping the frequency scale to do speaker normalization.

In the combination, SA is carried out in the way as described in Section 2. As for VTLN, Zhan and Westphal have proved the ML-based method performed better than formant-based one in their work [7]. Their experiments have also shown that the linear warping function, besides the appeal of simplicity, costs less but provides the best benefit. So we choose the linear warping function and ML approach to implement it when combined with SA.

Since the estimation methods of adaptation factors are quite different between SA and VTLN, we did them one after another, and then iterated the procedure till reach convergence.

4. EXPERIMENTS

To evaluate the proposed methods, we first tested the performance of SA method on adjusting the spectrum with artificial communication channel distortion. Then, the real telephone speech recognition experiments were carried out to investigate the recognition performances of SA and SAW. We also compared them with CMS SBR_ML and VTLN. The following sub-sections outline the experiments.

4.1. Spectrum adjusting testing

To test the spectrum adjusting performance of SA method, the following experiment was carried out: First, the clean speech was passed through an artificial distorting communication channel. Then, the adjusting function was estimated for the distorted speech with one sentence as the adaptation data. In the end, the distorted speech was adjusted by the function and then recognized. We also did the adaptation on the distorted speech using SBR_ML method for comparison.

The testing clean speech was from 863 speech database, totally 60 sentences uttered by three male.

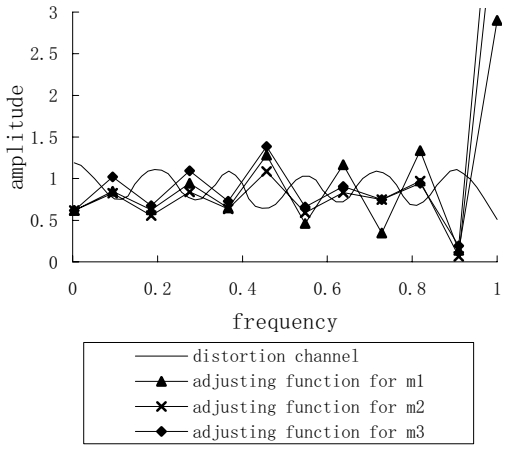

Figure 3 Distorting communication channel and the adjusting functions

	Clean	Distorted	SA	SBR_ML
M1	28.86%	29.97%	28.71%	31.39%
M2	35.38%	37.91%	35.92%	36.64%
M3	24.77%	27.09%	24.92%	26.63%
Average	29.67%	31.66%	29.85%	31.55%

Table 1 syllable error rate of the clean speech, distorted speech and SA, CBR method

Figure 3 shows the communication channel amplitude spectrum and piece-wise linear adjusting functions with 12 pieces (e.g. L=12) estimated for each speaker. It is shown that the adjusting function compensates for the communication channel spectrum very well for each testing person. The recognition results listed in table 1 further demonstrate this conclusion. As for SBR_ML, its performance is worse than SA.

4.2 Speech recognition on telephone speech

4.2.1 Baseline and database

In our baseline recognition system, DDBHMM with Gaussian Mixture Distribution (GMD) acoustic model was used. The feature of utterance was 45-dimension vector: 14 Mel-frequency cepstral coefficients (MFCC), and their first and second derivatives, along with the power, its first and second derivatives.

The telephone speech database was partitioned into training and testing data. The training data included totally 44,000 sentences uttered by 880 speakers, 410 male and 470 female. The testing portion consisted of nearly 2000 sentences uttered by 40 speakers, out of which 20 were male and 20 were female.

We used syllable error rate (SER) as the performance measure.

4.2.2 Experiments using SA

In our experiments, we selected L=10 by testing different values from 5 to 30. In the training procedure, the process stopped after 3 iterations. It should be pointed out that the channel characteristic (including record device and transmitting channel etc.) didn't change utterance by utterance for the same speaker, thus the adjusting factors for each speaker could be estimated from a small number of utterances by the same person. In this test, the adaptation data was the first 5 sentences, about 20 syllables. The short adaptation utterance assured the fast adaptation speed.

We also compared our method with CMS and SBR_ML, In CMS, for each sentence, the bias was the cesptral mean of this sentence. As for SBR, the first 5 sentences were used to estimate the bias, then, it was subtracted from all the sentences of the same speaker, just as in SA. Table 2 shows the average syllable error rates. SA performed better than CMS and SBR_ML. CMS outperformed SBR_ML due to the sent by sent adaptation method.

	SER	Error Reduction
Baseline	46.56%	--
SA	41.70%	10.44%
SBR_ML	43.69%	6.16%
CMS	42.61%	8.48%

Table 2 syllable error rate for SA, SBR_ML and CMS

4.2.3Eexperiments using SAW

In this test, we first adjusted the amplitude spectrum by SA. Then, VTLN was done to warp the frequency scale of the adjusted spectrum. The procedure was iterated till reach convergence. In SA, we also set L=10. For each speaker, the first 5 sentences were used to estimate the adaptation factors for both SA and VTLN.

The results presented in Table 2 shows that SAW performed much better than SA or VTLN alone.

	SER	Error Reduction
Baseline	46.56%	--
SA	41.70%	10.44%
VTLN	42.66%	8.38%
SAW	39.76%	14.60%

Table 3 syllable error rate for SA, VTLN and SAW

5. CONCLUSIONS

Firstly, this paper described Spectral adjusting (SA) channel adaptation method which adjusts the speech spectrum with piece-wise linear adjusting function to reduce the channel mismatch. Then, the spectral adjusting and warping (SAW) method was introduced. In this method, we adjusted and warped the speech spectrum by the combination of SA and VTLN to compensate for the channel and speaker variation together.

The telephone speech recognition results showed SA performed better than CMS or SBR_ML, SAW performed better than SA or VTLN alone, which indicated the idea of adapting the spectrum of speech in scale and amplitude for channel and speaker adaptation was helpful to get better performance in speech recognition.

6 REFERENCES

[1] Z.Y. Wang, "An inhomogeneous HMM speech recognition algorithm", *Chinese Journal of Electronics*, vol. 7, no. 1, pp 73-74, 1998.

[2] Y. Zhao, "An Acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 380-394, 1994.

[3] M.G. Rahim and B.J. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp 19-30, 1998.

[4] Y. Zhao, "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Trans. Speech Audio Processing,* , vol. 8, pp. 255-266, 2000.

[5] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on speech and audio processing,* Vol. 6, No. 1, pp. 49-60, 1998.

[6] E. Eide and H. Gish, "A parametric approach to vocal length normalization," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 346-348, 1996.

[7] P. Zhan, M. Westphal, "Speaker Normalization based on frequency warping," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1039-1042, 1997.

[8] E.B. Gouvea and R.M. Stern, "Speaker normalization through formant-based warping of the frequency Scale," *5th European Conference on Speech Communication and Technology,* vol. 3, pp. 1139-1142, 1997.

[9] M. Lincoln, S. Cox and S. Ringland, "A fast method of speaker normalization using formant estimation," *5th European Conference on Speech Communication and Technology,* vol. 4, pp. 2095-2098, 1997.