CODEBOOK DEPENDENT DYNAMIC CHANNEL ESTIMATION FOR MANDARIN SPEECH RECOGNITION OVER TELEPHONE

Huayun Zhang    Zhaobing Han    Bo Xu
National Laboratory of Pattern Recognition, Chinese Academy of Sciences,
P.O.Box 2728, Beijing, China 100080
{hyzhang, zbhan, xubo}@nlpr.ia.ac.cn

ABSTRACT

Automatic speech recognition in telecommunications environment still has a lower correct rate compared to its desktop pairs. Improving the performance of telephone-quality speech recognition is an urgent problem for its application in those practical fields. Previous works have shown that the main reason for this performance degradation is the variational mismatch caused by different telephone channels between the testing and training sets. In this paper, we propose an efficient implementation to dynamically compensate this mismatch. This algorithm bases on maximum-likelihood (ML) estimation of telephone channels and dynamically follows the time-variations within the channels. It could deal with both linear channels’ (like fixed telephone lines) degradation and some noisy non-linear channels’ (like some long distance lines and wireless circuit lines, such as GSM) degradation. In our experiments on Mandarin large vocabulary continuous speech recognition (LVCSR) over telephone lines, the average character error rate (CER) decreases more than 20% when applying this algorithm. At the same time, the structural delay and computational consumptions required by this algorithm are limited. The average delay is about 300–400ms. So it could be embedded into practical telephone-based applications.

1. INTRODUCTION

It is well known that the bandwidth of switching telephone network is limited to 300–3400Hz due to historical reasons. The performance degradation of telephone speech recognition is partly due to this bandwidth’s limitation. But a more important reason is the telephone channel mismatch between training and testing sets. Inherent environmental variability introduced by the real network has the most impact on recognition accuracy.[1]

Table (1) is a summary of our series experiments on Mandarin speech recognition under 5 different channel conditions.

```
<table>
<thead>
<tr>
<th>Database</th>
<th>Speech Bandwidth</th>
<th>CER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>0-8000Hz</td>
<td>12.2</td>
</tr>
<tr>
<td>Downsampled</td>
<td>0-4000Hz</td>
<td>14.3</td>
</tr>
<tr>
<td>Linear filtered</td>
<td>300-3400Hz</td>
<td>15.4</td>
</tr>
<tr>
<td>Through PSTN</td>
<td>300-3400Hz</td>
<td>23.8</td>
</tr>
<tr>
<td>Through GSM Codec</td>
<td>300-3400Hz</td>
<td>21.5</td>
</tr>
</tbody>
</table>
```

Table 1: Comparison of Different Channel Conditions

This table lists the character error rate (CER) obtained by training and testing on (1) high quality 16KHz sampled speech recorded in studio, (2) speech resample to 8KHz through DSP toolkits, (3) speech passed through a constant linear filter designed to approximate the frequency response of a typical telephone line, (4) speech passed through the real PSTN network, (5) speech passed through a GSM codec. It could be find the character error rate (CER) increases only about 2%–3% compared to original speech when the channel is prefect and time-invariant linear. But for real PSTN and GSM channels, the CER increases about 9–12%.

It has obviously shown that the real PSTN channel and GSM channel are not perfect linear time-invariant systems. Every telephone call results in different channel response and noise condition [2]. Nonlinear distortion further complicates the problem. With the widely spreading of cellular telephones (includes GSM and CDMA), non-linearity of the whole telephone network is increasing. The speech codec algorithm (like ETSI GSM 6.10) is not a linear filter [3,4].

In recent years, some techniques have been proposed to compensate the telephone channel degradation [5,6,7,8,9,10,11]. Usually, the telephone channel is assumed to be a linear system and invariant with time. It is partly appropriate for good-quality fixed telephone line. But for other low-quality long-distance telephone lines and non-linear channels (like GSM), this assumption is quite improper and compensation effects are limited [4]. When non-stationary additive noise and non-linear distortion exists, the channel could not be simply modeled as a time-invariant linear system. In addition, it is often the case that the channel compensation is required to be done through the whole telephone call procedure universally. The delay caused by the compensation algorithms and the computational complexity has prevented them to be imbedded in the real time systems.

In this paper, we propose a short period channel estimation method that could dynamically follow the time-varying channel degradations. We name it Codebook-Dependent Dynamic channel Compensation (CDDC) algorithm. In the next part, we will describe the details of CDDC algorithm. In the 3rd part, we will describe our experiments on telephone-quality continuous Mandarin speech recognition. CDDC is compared with three other commonly used channel compensation techniques. We will present our results for fixed-telephone-line-quality speech and GSM-quality speech separately. In the 4th part, we will discuss some issues in practical fields for CDDC. Finally we will give our conclusion in the 5th part.

---

1This work was supported by National Science Foundation of China under grant number 69835003.
2. CODEBOOK DEPENDENT DYNAMIC CHANNEL ESTIMATION

Although real telephone line is not a perfect linear time-invariant system during the whole telephone call procedure, we could assume that the channel shows quasi-linearity during a short-time period.

The original clean speech at the sending of the line is \( x(t) \). The frequency response of the line is \( H_f(\omega) \). The additive noise exists in the line is \( n(t) \), the received signal on the other end is \( y(t) \). This procedure could be modeled as following:

\[
\left| Y_f(\omega) \right|^2 = \left| X_f(\omega) \right|^2 \left| H_f(\omega) \right|^2 + \left| N_f(\omega) \right|^2
\]

Then in log-spectral and cepstral domain, the channel’s effects on the speech could be modeled as:

\[
y_t = x_t + b_t
\]

Where:

\[
b_t = \log \left( \left| H_f(\omega) \right|^2 \right) + \log \left( 1 + \frac{\left| N_f(\omega) \right|^2}{\left| H_f(\omega) \right|^2 - \left| X_f(\omega) \right|^2} \right)
\]

The channel’s degradation could be considered as a fixed global constant bias when a) \( H_f(\omega) \) is invariant with time and b) \( N_f(\omega) < H_f(\omega) \cdot X_f(\omega) \). However in many practical cases, these two assumptions could not be maintained. Since additive noise exists in the telephone channel and \( SNR = \frac{X_f(\omega)}{N_f(\omega)} \), considering the channel bias as a SNR (Signal-to-Noise Ratio) dependent value is reasonable. The bias may be defined as separate values for speech and non-speech segments [7]. In some times, a phone or state dependent bias is meaningful [3]. For example, in Mandarin speech, signal power of the initial segments is usually much lower than the final segments. But even for the same phone unit, signal power varies largely for different person and in different time. So phone-dependent bias estimation is a roughly estimation for the channel degradations.

Since \( SNR \) varies with the speech, channel bias also varies through out the speech utterance. So the bias \( b_t \) is a variable in the time index. But during a period short enough (about several hundreds of milliseconds) around the time point \( t \), the short-time \( SNR \) may be considered as a constant and the frequency response \( H_f(\omega) \) could be considered as a short-time stable one, the channel bias \( b_t \) could be estimated as a fixed value.

Our formulation of the bias estimation is based on maximizing the likelihood of observation series during a short-time window on a speech codebook in which the bias is considered as the unknown parameter.

The statistics of speech could be modeled by a data-driven codebook, which could be obtained by any kind of vector quantization method or a phone or state dependent codebook. Here we use a phone dependent codebook. There are totally 24 initials and 37 Finals in Mandarin. The codebook is composed of 62 code words to represent all Mandarin phones plus a pause unit. Each codeword is described as multi-mixture components Gaussian distribution. The codebook is denoted as:

\[
\Omega = \{ \omega_n \} \quad n = 1, 2, \cdots N
\]

\( N \) is the number of code words in the codebook (\( N = 62 \) for our system). Each code word \( \omega_n \) is defined by a parameter set:

\[
\omega_n = \left\{ \alpha_{n,m}, \mu_{n,m}, \Sigma_{n,m} \right\} \quad m = 1, 2, \cdots M
\]

\( M \) is the number of mixture components in the Gaussian distribution. \( \alpha_{n,m}, \mu_{n,m}, \Sigma_{n,m} \) are the mix coefficients, mean vector and covariant matrix of the \((n,m)\)th component separately. The probabilistic model defined by codeword \( \omega_n \) is:

\[
p(x | \omega_n) = \sum_{m=1}^{M} \alpha_{n,m} p(x | \mu_{n,m}, \Sigma_{n,m})
\]

\[
\quad = \sum_{m=1}^{M} \frac{\alpha_{n,m}}{\sqrt{(2\pi)^D |\Sigma_{n,m}|}} \exp \left( -\frac{1}{2} (x - \mu_{n,m})^T \Sigma_{n,m}^{-1} (x - \mu_{n,m}) \right)
\]

Here \( D \) is the dimension of the observation \( x_t \).

Our target is to find the bias \( b_t \), which maximizes the likelihood of observations \( Y \) during a short-time window on codebook \( \Omega \).

\[
p(Y | b_t, \Omega) \quad \text{where} \quad Y = \left\{ y_{\frac{T}{2}}, \cdots, y_{\frac{T}{2}}, \cdots, y_{\frac{T}{2}} \right\}
\]

\( T \) is the length of the window and \( t \) is the center of this window.

The famous Expectation-Maximization (EM) algorithm [12] could settle this kind of ML parameter estimation problem. We define the EM auxiliary function as follows:

\[
Q(b_t | b) = E[\log p(Y | b_t, \Omega) | b_t, \Omega]
\]

EM guarantees mathematically that if \( Q(b_t | b) \geq Q(b_t | b) \), then \( p(Y | b_t, \Omega) \geq p(Y | b_t, \Omega) \).

The auxiliary function could be expressed as:

\[
Q(b_t | b) = E[\log p(Y | b_t, \Omega) | b_t, \Omega] = \sum_{i=1}^{T} E[\log p(y_{\frac{T}{2}} | b_t, \Omega)]
\]

\[
= \sum_{i=1}^{T} \sum_{n=1}^{M} p(y_{\frac{T}{2}} | n, m, b_t, \Omega) \log [p(y_{\frac{T}{2}} | n, m, b_t, \Omega)]
\]

Where:

\[
p(y_{\frac{T}{2}} | n, m, b_t, \Omega) = \frac{\alpha_{n,m}}{\sqrt{(2\pi)^D |\Sigma_{n,m}|}} \exp \left( -\frac{1}{2} (y_{\frac{T}{2}} - \mu_{n,m})^T \Sigma_{n,m}^{-1} (y_{\frac{T}{2}} - \mu_{n,m}) \right)
\]

If we assume that there are non-correlations between dimensions of the cepstral vector, then \( \Sigma_{n,m} \) becomes a diagonal matrix. We denote the non-zero diagonal elements as:

\[
\sigma_{n,m}^2 = \left( \sigma_{n,m,1}^2, \sigma_{n,m,2}^2, \cdots, \sigma_{n,m,i}^2, \cdots, \sigma_{n,m,D}^2 \right)
\]

The auxiliary function could be decomposed to dimension \( i \) as:

\[
Q_i(b_t | b) = \sum_{i=1}^{T} \sum_{n=1}^{M} p(y_{\frac{T}{2}} | n, m, b_t, \Omega) \left( K(n, m) - \frac{(y_{\frac{T}{2}} - \mu_{n,m})^T}{2\sigma_{n,m}^2} \right)
\]

Where \( K(n, m) = \log \left( \frac{\alpha_{n,m}}{\sqrt{2\pi\sigma_{n,m}^2}} \right) \).
In order to find the maximum value of $Q$, we could differentiate it with respect to $b'$ and solve for its zeros.

$$\frac{\partial Q}{\partial b'} = 0 \Rightarrow$$

$$\sum_{i} \sum_{n} \sum_{m} p(y_{i}, n, m | b, \Omega) \frac{y_{i, j} - b_{i} - \mu_{n, n}}{\sigma_{n, n}} = 0$$

(11)

The final resolution is:

$$b_{i}' = \frac{\sum_{i} \sum_{n} \sum_{m} p(y_{i}, n, m | b, \Omega) \frac{y_{i, j} - \mu_{n, m}}{\sigma_{n, m}}}{\sum_{i} \sum_{n} \sum_{m} p(y_{i}, n, m | b, \Omega) \frac{1}{\sigma_{n, m}}}$$

(12)

This resolution has similar form with the one in [7]. But there are many differences between them. Here, we use a phone dependent codebook instead of complicated HMM to represent speech. This means the Viterbi algorithm used in [7] to find the optimal state sequence is not needed in our CDDC. This alteration could save large mounts of computation and make the CDDC more easily to be implemented in real time systems.

The EM resolution is an iterative process. First set $b = 0$, use formula (12) to do iteration until the increase in the likelihood defined in (7) is less than some predetermined threshold.

Shifting the short time window along the observations $Y$, we could obtain a dynamic trajectory of the channel bias $b$.

Obviously, repeating the EM iterations within every short time window is an exhausting work. Fortunately, we found the EM iteration converges very fast on the phone dependent codebook. Figure (1) shows the probability’s increase of a typical Mandarin sentence on the codebook according to the iterations.

We also investigate the accuracy rate variation along with the iteration times in experiments. It shows no obvious improvements when the EM iteration times are set beyond 2.

Since the algorithm converges so rapidly, it is acceptable that we do the EM iteration only once in CDDC. For the incoming observation $Y$, we only need to estimate its probability through out the codebook once a time.

![Figure 1: Convergence of the CDDC.](image)

The only structural delay of CDDC is the half-length of the short time window. This makes our algorithm very easily to be adopted in real-time applications, such as telephone systems.

Once the channel bias has been estimated, the channel’s frequency response could be reconstructed. The following figure (2) is the reconstructed channel frequency response from a segment of telephone speech data about 2.4s.

![Figure 2: The Channel’s Time-Frequency Response Reconstructed by the CDDC.](image)

Figure (2) shows that the telephone channel has a higher attenuation at lower frequencies and higher frequencies. However, at the middle frequencies, the response is not invariant with time, which it is often assumed to be.

### 3. EXPERIMENTS

We investigated CDDC in telephone quality speaker-independent Mandarin speech recognition. It’s obvious that in order to maximize speech recognition performance, the training and testing sets should be as similar as possible. This means that we should use real telephone quality speech to train acoustic model for recognition. However there are still smaller amounts of real telephone speech available for training at present.

In order to obtain telephone quality speech material for the acoustical model training and CDDC codebook training, we utilize the Mandarin 863 speech database. This database was developed by Chinese national 863 program for LVCSR. It contains about 70 hours’ speech. The speech data are 16000Hz sampled and 16bit linear quantized. We disposed this database with three methods: 1) resample the database to 8000Hz and $\mu$-law quantization; 2) pass the database through the real PSTN network by Dialogic telephone cards plugged in PCs; 3) pass the database through GSM full rate (GSM FR 06.10) coder and decoder. All these three transcoded databases are used as training data for the acoustical model and the CDDC codebook.

The vocabulary of this task consists of more than 40K words. N-gram statistics are used for language modeling.

There are two test sets in our experiments named as TELTEST and GSMTEST separately. TELTEST is gathered through the PSTN network and GSMTEST is gathered through GSM-FR codec. Each contains about 240 continuous Mandarin sentences from 4 different speakers.

We compare our CDDC algorithm with other three widely used channel normalization techniques:

1. Long term CMS (Cepstral Mean Subtraction over the duration of a whole Mandarin sentence)
2. RASTA filtering [5]
3. SBR (Signal Bias Removal) [6]. The codebook in SBR used in our experiments is the same used in CDDC.

The acoustic features consist of energy, pitch, 12 mel-cepstral with delta and delta-delta features. Estimated Bias is subtracted from mel-cepstral directly and affects delta and delta-delta mel-cepstral indirectly.

The characters in error are of three types, namely substitution errors, deletion errors, and insertion errors. Identification of these errors results from the process of aligning the characters...
in the reference transcription with the characters in the system output. This alignment is performed using NIST's SCLITE. The results are showed in the following figure (3).

![Figure 3: Experimental results using different channel compensation algorithm](image)

In our experiment, CDDC is the most effective one in these four compensation techniques. Compared to the CMS, CDDC decreases the CER about 22.3% for TELTEST and 25% for GSMTEST. Especially for GSMTEST, where GSM channel is more non-linear compared to fixed lines, CDDC is the only one that could improve the system performance obviously.

4. DISCUSSION

There are two important parameters for applying CDDC in real time systems. One is the length of the shifting window, which is the structural delay of CDDC. The other is the size of the codebook, which directly related to the computational delay of CDDC.

We alter the shifting window length from 100ms to 1.5s. The results are recorded in figure (4). The shorter the window is, the better adaptability CDDC has. But when the window is too short, there are no enough observations for estimation. When the window is too long, CDDC could not follow the channel’s variation. We could find an appropriate value empirically. In our experiments, the delay of 400ms is an appropriate selection.

![Figure 4: CER with different window length.](image)

The number of mixtures in codebook Gaussian distribution affects the amount of computation largely. We give the system performance for different mixture number in figure (5).

![Figure 5: CER with different mixture number](image)

When the number of mixture components changes between 2 to 16, it has no obvious relation with the system performance.

5. CONCLUSION

We have proposed a codebook dependent dynamic compensation (CDDC) algorithm. There are two advantages for this algorithm: 1). A median sized phone dependent codebook is used to represent the statistics of speech in the maximum likelihood estimation for the telephone channel, which saves large amounts of computation compared to the HMM used in [7]. At the same time, CDDC obviously improves the system performance 2). This algorithm could dynamically follow the variations inherent with the real telephone network. It could tackle with degradations introduced by low-quality, non-linear channel distortion and non-stationary noise contamination. These two improvements make CDDC more suitable for fields’ applications. We have tested this algorithm on two kinds of telephone quality speech sets: the PSTN quality speech and the GSM quality speech. Compared to other channel compensation methods, CDDC shows more adaptability to the variations introduced by the channel and obtains about 23% character error rate decreasing in Mandarin telephone speech recognition.

6. REFERENCES