POWER SPECTRAL DENSITY BASED CHANNEL EQUALIZATION OF LARGE SPEECH DATABASE FOR CONCATENATIVE TTS SYSTEM

Yu Shi, Eric Chang, Hu Peng, and Min Chu

Microsoft Research Asia
5F, Beijing Sigma Center, No. 49, Zhichun Road, Haidian District
Beijing 100080, P.R.C.
{i-yshi,echang,hupeng,minchu}@microsoft.com

ABSTRACT
This paper proposes a channel equalization algorithm for a large speech database with application in concatenative TTS systems. The convolutional channel distortion is equalized by comparing the power spectral densities (PSDs) of utterances from different recording sessions. Autoregressive linear filters are designed on a corpus level and are used offline to filter the corresponding sentences to compensate for the relative distortions caused by the channel effects. Two experiments are carried out to evaluate the benefit of the channel equalization approach. First, this method is used to reduce the distance of their PSDs between two recording sessions to verify the effectiveness of the method. Secondly, it is applied practically in the TTS system. The whole TTS speech database is processed to reduce the PSDs variance over all sessions. Moreover, a subjective listening test is carried out to obtain human evaluation of the new TTS system. Almost all listeners prefer the synthetic speech generated by the new TTS system. Furthermore, an analysis of variance (ANOVA) on this subjective listening test demonstrates that the channel equalization process has significant effect on increasing the perceived voice-quality consistency of the TTS system.

1. INTRODUCTION
Channel equalization techniques have various usages in speech processing. In concatenative text-to-speech (TTS) systems, channel equalization can be used to compensate for the voice-quality inconsistency among units in large speech databases. In recent years, naturalness of TTS systems has vastly improved. This is mostly due to a wider acceptance of corpus-based synthesis paradigms and improvements in unit selection. For concatenative speech synthesis, an optimum set of basic units that matches the desired prosodic features of the sentence will be selected from a large speech database by using a unit-selection approach. The availability and consistency of the selected units seem to be the key for natural sounding TTS systems.

However, recording the necessary large speech databases for speech synthesis is a long process lasting from many days to months. The time between recording sessions can be weeks. The duration of each recording session can be as long as many hours including some breaks. Different recording sessions or segments may have differences in transmission channels caused by the changes of the distance between mouth and microphone, the status of the recording equipments, and the recording environment. Therefore, it is very possible to have variations in voice quality from one recording session to another as well as during the same recording session. The effect is that sentences from different sessions or from the same session but having long intervals may be perceived as being said by different speakers even when they are spoken by a single person. The inconsistencies in voice quality within the same unit speech database may degrade the overall perceived quality of TTS systems.

To solve this problem, one can increase the number of instances of the basic units used by a TTS system by increasing the size of database and then use a powerful discriminative unit selection strategy to reject the concatenation units of the database with mismatch in voice quality. Unfortunately, this method will result in the TTS system only using part of the database while time and money was spent to record the complete database. And this way, increasing the size of the database will not necessarily increase proportionally the number of units available for synthesis. Alternatively, channel equalization of the units in the existing large speech database is possible. This technique can be implemented offline and will save time and money in recording new databases.

Several researchers have concentrated on modifying speech units to remove formant discontinuities based on either inverse filtering, usually by an LPC filter [1] or sinusoidal modeling [2]. Another type of channel equalization is applied on a sentence or corpus level, such as normalization of acoustic differences between recording sessions [3].

In this paper we design autoregressive filters via the PSDs of signals recorded through different channels. These filters are directly used on utterances in the time domain. In the next section we will describe the convolutional effects of the channels on voice quality and give the content independent assumption of the PSDs. Based on this, Section 3 will present a channel equalization algorithm and a filter design method. The implementation of the channel equalization algorithm and two experiments are described in Section 4. Results show the effectiveness of the algorithm for reducing the voice-quality variance, which is confirmed by subjective listening tests.

2. CHANNEL EFFECTS ON VOICE QUALITY
Figure 1 shows a simple recording system with s being the orig-
nal speech waveform before passing through a channel, \( h \) a transmission channel, and \( x \) the recorded signal. While there is only one microphone in Figure 1 that represents the channel, channels in practice are very complex. Here we only consider the channels that affect speech as convolutional filters. Multiple convolutional filters can be combined into a single one. It is assumed that the recording environment is very clean, so there isn’t any additive noise in the corpus.

\[ x = s \ast h + n \]

*Figure 1: A simple speech recording system.*

Based on above assumptions, we can write the input-output relationship of the channel in recording session \( i \) as:

\[ x_i(n) = s_i(n) \ast h_i(n) \]  

(1)

Let \( X_i(m), S_i(m), \) and \( H_i(m) \) denote the Fourier transforms and \( P_i(m), P_s(m), \) and \( P_h(m) \) denote the PSDs of \( x_i(n), s_i(n), \) and \( h_i(n) \), respectively. With this, equation (1) can be expressed in frequency domain as:

\[ X_i(m) = S_i(m)H_i(m) \]  

(2)

The PSDs, can be expressed similarly:

\[ P_i(m) = |X_i(m)|^2 = |S_i(m)|^2 |H_i(m)|^2 = P_s(m)P_h(m) \]  

(3)

If an utterance from session \( i \) is long enough, its Fourier transform and PSD can be assumed to be content independent. So equations (2) and (3) can be rewritten approximately as:

\[ X_i(m) = S(m)H_i(m) \]  

(4)

\[ P_i(m) = P_s(m)P_h(m) \]  

(5)

3. CHANNEL EQUALIZATION ALGORITHM

According to equation (4), the recording system can be described by the block diagram shown in Figure 2. Suppose we have a sentence \( s_i(m) \) from session \( i \) as a reference (target), and another sentence \( s_j(n) \) from session \( j \) that needs to be equalized. The goal is to design a filter, by which after \( s_j(n) \) is filtered, the Fourier transform or PSD of \( s_j(n) \) should equal to that of \( s_i(n) \). This can be implemented by using a channel equalization system shown in Figure 3. The channel equalizer \( H \) in the right part is the filter to be designed.

From Figure 2 and Figure 3 one can easily obtain an optimal channel filter as:

\[ H(m) = H_i(m)/H_j(m) \]  

(6)

Substituting equation (4) into equation (6), the channel equalizer can be rewritten as:

\[ H(m) = X_i(m)/X_j(m) \]  

(7)

Apparently, PSDs have the similar relationship as (7):

\[ P_k(m) = P_s(m)/P_j(m) \]  

(8)

Therefore the channel equalization algorithm proceeds as follows (see the block diagram in Figure 4).

**Step 1**: Estimate the long-term average PSDs [4] for both \( s_i(n) \) and \( s_j(n) \) as:

\[ P_s(m) = \frac{1}{L_k} \sum_{n=0}^{L_k} P_s(m) \text{, } k = i, j \]  

(9)

with \( L_k \) being the number of speech frames extracted from waveform \( s_k(n) \) and \( P_s(m) \) being the short-term PSD estimate of the \( l \)-th speech frame given by

\[ P_s(m) = \frac{1}{N} \sum_{n=0}^{N-1} x_k(n)w(n) \exp(-j2\pi mn/N) \]  

(10)

where \( x_k(n) \) is the \( n \)-th sample of the \( l \)-th speech frame of \( s_k(n) \), \( N \) is the length of speech frames and \( w(n) \) is a Hamming window with

\[ \sum_{n=0}^{N-1} |w(n)|^2 = 1 \]  

(11)

Note that not all of the frames in \( x_k(n) \) contribute to the PSD in equation (9). Only voiced speech frames are used in this paper. A simple way to judge whether a frame is voiced or unvoiced is comparing its energy with a threshold:

\[ E_k = \sum_{n=0}^{N-1} |x_k(n)|^2 \begin{cases} \text{voiced} & \text{if } E_k \geq E_v \\ \text{unvoiced} & \text{if } E_k < E_u \end{cases} \]  

(12)

Because the corpus is very clean, this simple method is sufficient to reliably extract voiced frames.

**Step 2**: Estimate the PSD, \( P_l(m) \) of the channel equalizer \( h(n) \) according to equation (8).
that after the channel equalization process, the channel effects on the voice quality have been reduced substantially.

The filter coefficients are ordered in descending powers of $z$: $$H(z) = \frac{B(z)}{A(z)} = \frac{b}{1 + a(1)z^{-1} + \cdots + a(p)z^{-p}}$$ (15)

where $b$ is the sum of the filter coefficients.

We have a database consisting of 20816 clean sentences uttered by one speaker over 23 days during a period of several months. Each day has at least one session caused by changing tapes, having breaks, and so on. The speaker is capable of control her glottal very well, so the voice qualities mainly depend on the channel status. There are 59 sessions in the database. Because the original sentences are not long enough to satisfy equation (4) or (5), sentences recorded consecutively are concatenated into a single group first. The amount of the speech in each group will affect the quality of channel equalization. If too short, the distribution of the phones covered is not uniform, and the PSD reflects the characteristics of the phones it covered as well as the channel effect (the PSD is content dependent). For this reason, the channel equalizer not only compensates for the channel difference but also destroys the phones’ characteristics, which can induce some distortions. On the other hand, if the speech is of a very long time, then the channel may vary substantially from the beginning to the end. In this paper, we choose every group to be about 5 minutes long. Experiments show that designed filters based on speech of that length will almost never induce distortions into the filtered sentences. After the concatenation process, there are 200 groups in the database, and 200 PSDs are calculated.

The sampling frequency is 16 kHz. In the experiments, both the size of analysis window and the frame rate are 10 msec. The voiced/unvoiced threshold $E_a$ is set to 0.03, the length of FFT for estimating the PSDs is 256, and the order of the IIR filters is 30.

To test the performance of the channel equalization method, two experiments were performed. The first one is carried out on speech waveforms of two different sessions extracted randomly from the database, while in the second experiment, all of the speech signals in the database are equalized and then a new database is created.

## 4. EXPERIMENTAL RESULTS

### 4.1. Channel Equalization Between Two Sessions

In the first experiment, two sessions are randomly selected from the database, session one for reference and session two for channel equalization. The sentences in session two are equalized by using the algorithm described in Section 3. Figure 5 shows the PSD curves of the reference session and both the original and equalized versions of the test session. It can be seen that session two has very strong power at lower frequencies before equalization. After equalization, its PSD becomes closer to the reference session. Therefore, the voice quality of the signals in the equalized test session becomes more similar to that of the references.

### 4.2. Channel Equalization for TTS System

In the second experiment, we process the whole database group by group. The reference PSD is set to the average of all 200 PSDs. All groups of speech waveforms are equalized using the channel equalization algorithm described in Section 3, after which a new database is created. The mean value of the pair wise distances of the PSDs of the old database is 0.2158, while after channel equalization, this value reduces to 0.0518. The PSDs of all groups have become closer to each other greatly, which confirms that after the channel equalization process, the channel effects on the voice quality have been reduced substantially.
In order to further evaluate the benefit for the TTS system from using the channel equalization approach, a human evaluation experiment is performed. In this preference experiment, two TTS systems are built based on the databases before and after channel equalization, respectively. Afterwards, two classes of utterances are synthesized by the TTS systems via the same unit selection algorithm [5].

When designing the preference experiment, we found that if the optimal units of a synthesized sentence are selected from the same recording session or from the different sessions with similar voice quality, then it may sound quite consistent no matter by which TTS system it is synthesized. That is to say, the sentence synthesized by the non-equalized TTS system already has an equally good consistency to the equalized one so that human could almost not perceive the difference between them. To avoid inducing the unwanted confusion into the subjective listening experiment, we choose the test sentences by using the following steps:

- **Step 1**: Synthesize 1000 sentences by selecting units from the non-equalized database and find out the original session of each selected unit.
- **Step 2**: Calculate the PSD distances between every two sessions, from which two consecutive units of a sentence are selected.
- **Step 3**: Average the PSD distances over each sentence to estimate how large the channel inconsistency might be.
- **Step 4**: Choose 30 sentences with the largest average PSD distances as the test sentences.
- **Step 5**: Synthesize those test sentences also by using the TTS system built on the channel equalized database. Then together with the sentences chosen in step 4, there are totally 60 sentences in the test set.

Later, the synthesized test sentences are played to 20 listeners pair by pair in random order. Each listener, with no prior instruction to compare voice quality inconsistency, is asked to indicate the more natural sentence from each pair.

An analysis of variance (ANOVA) is performed on the preference experiment to demonstrate that channel equalization has a significant effect on increasing the voice quality consistency of TTS systems. A single-factor ANOVA comparing the non-equalized TTS system and equalized TTS system shown in Table 1 is significant ($F_{0.05}(1, 58) = 5.3$).

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
<th>Sum</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>no equalization</td>
<td>30</td>
<td>10.7</td>
<td>0.36</td>
<td>0.0148</td>
</tr>
<tr>
<td>with equalization</td>
<td>30</td>
<td>19.3</td>
<td>0.64</td>
<td>0.0148</td>
</tr>
</tbody>
</table>

**Table 1**: Single factor ANOVA comparing the non-equalized and equalized TTS systems.

**5. CONCLUSION**

In this paper, we propose a channel equalization algorithm to compensate for the channel mismatch among the utterances of different recording sessions in a large speech database with application in concatenative TTS systems. This is a corpus level approach and can be accomplished offline. Given a speech sentence having a reference (target) channel quality, the implementation of this method is to design an IIR filter for a speech sentence corrupted by another channel by comparing the PSDs of them. The IIR filter has the ability to compensate for the channel difference and make the latter sentence sound like the reference one.

This method is introduced in the TTS system. It normalizes all of the sentences in the speech database used by the TTS system, to make them have similar voice quality no matter which recording session they belong to. Sentences synthesized by the equalized database have a more consistent voice quality than the sentences synthesized by the non-equalized one. Furthermore, in order to confirm the effectiveness of the channel equalization algorithm, a subjective listening test is carried out on both new and old TTS systems. In this experiment, almost all listeners prefer the synthetic speech generated by the new TTS system. After that, an ANOVA is done on the human evaluation results, which also confirms that the channel equalization process has significant effect on increasing the voice quality consistency of the TTS system.

**6. REFERENCES**


