Telephone Speech Recognition Using Simulated Data from Clean Database

Guoyu Zuo††, Wenju Liu† and Xiaogang Ruan††

†National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, P. O. Box 2728, Beijing 100080, P. R. China
††School of Electronics Information and Control Engineering, Beijing University of Technology, Beijing 100022, P. R. China
Email: {gyzuo, lwj}@nlpr.ia.ac.cn, adrxg@bjpu.edu.cn

Abstract
Speech recognition over lines forms an integral part of various applications of large vocabulary continuous speech recognition (LVCSR). This paper describes an implementation system completely in software form to produce simulated telephone data starting from clean databases. Filters adopted in this system are well-designed to simulate the frequency properties of analogue transmission equipments in telephone connection. A speech recognizer was trained from speech data extracted from clean corpus piped through a hardware simulator. The recognition performances are evaluated on a real telephone speech set and several test sets simulated from the clean database for test use. The experiments verified the effectiveness and feasibility of software simulation from a recognition testing point of view, and the results showed using simulated data derived from clean corpus could achieve the same recognition performance as real telephone speech.

1 Introduction
Telephone speech technology are getting more involved in many new applications of spoken language processing such as voice service center in hotels and restaurants and voice navigation in traffic and transportation system. However, there exists a great problem obtaining real telephone speech, and good recognition accuracy challenges the researchers of speech field if the speech came through a telephone channel. To build a simulation method is therefore a good idea of getting simulated speech from clean speech, when there are clean speech data in profusion but meager real telephone speech resources.

Speech quality is affected by the transmission impairments found on telephone connections. Sometimes the intelligibility and naturalness of speech degrade to an intolerable extent. The significant impairments and their effect on speech signals are described in [1]. These impairments include loudness loss, circuit noise, sidetone loudness loss, room noise, attenuation distortion, talker echo, listener echo, quantizing distortion and phase jitter. In addition, such user interfaces terminating the transmission channel as mobile handsets and hands-free terminals are likely to pick up background noise. More details will be discussed in the next section.

Some work had been done to telephone speech simulation to obtain telephone speech data. Morone et al. obtained simulated databases by passing clean speech through a hardware telephone channel simulator allowing a rather wide range of telephone conditions [2]. Tarcisio et al. simulated the transmission channel by filtering with a measured impulse response and adding recorded background noise [3]. Guiliani et al. has proposed the similar technique by modeling hands-free terminals [4]. Deficiencies in such techniques arise from the fact that transmission impairments found in telephony tend to be various in nature. Möller et al. presented a model to investigate the impact of specified telephone channel characteristics on ASR the performance of a speech recognizer [5]. The transmission functions were implemented mainly by DSP boards and other hardware equipments. For example, they used three DSP boards to simulate asynchronous codec tandems with up to three coding–decoding processes. The method had the advantage of good controllability and being real-time but also a weakness for hardware.

This paper presents a software implementation for telephone simulation based on the International Telecommunication Union’s (ITU-T) planning configuration [6] and using the real-time system above for reference. The next section deals with the simulation approach in detail. Section 3 describes the training and test sets for experiments. Section 4 provides the comparing experimental results about testing the simulated data with real telephone speech, and some
analyses on the results have been made in this section.

2 Simulation Method

2.1 Method Description

Digital telephony in the Public Switched Telephone Network (PSTN) is now commonplace except for the basic telephone services provided via analog subscriber lines. Figure 1 shows the speech flow over the end-to-end PCM channel with the PSTN (and associated 64 kbits/s architecture). The method under discussion in this paper is implemented completely in software form. This simulation method tends to simulate as exactly the real transmission characteristics of telephone channel as possible to obtain speech similar to telephone quality. Most of transmission impairments are taken into consideration (e.g. loss, circuit noise, echo, delay and quantization distortion arising from the codec schemes, code and decode), besides which the impulse noise and crosstalk disturbance are also simulated in this system. Especially, analogue transmission facilities in telephone connection are modeled as various filters regarding their physics characteristics, which reflect the frequency changes to the speech signal flowing through them.

2.2 Signal Preprocessing

The Mandarin speech database DB863 has time sample of 16 kHz and quantization precision of up to 16 bits. The high-quality clean speech from the database is first downsampled to 8 kHz. Digital speech codec is sensitive to the level of the input speech, consequently affecting the intelligibility of the output speech. To reduce the effect of speech level merely due to codec, the next processing is to normalize the speech level. When assessing the performance of a codec it is usual to set the input levels to be relative to the overload point of the codec. The overall gain calibration factor $S_{\text{cal}}$ is defined by:

$$
S_{\text{cal}} = \frac{\sum_{m=0}^{M} \sum_{n=0}^{N} x_{mn}^2 [n]}{X_{ov}^2 \sum_{m=0}^{M} N_m}
$$

where $x_{mn}$ is the $n$th sample value of the $m$th voiced segment in a speech portion, $X_{ov}$ is the overload point sample, $M$ is the number of voice segments, and $N_m$ denotes the number of samples in the $m$th voiced segment. Each speech utterance gets the same power level by operating on speech data with the factor $S_{\text{cal}}$.

2.3 IRS Filtering

The speech signal is passed through the sending part including the handset microphone and analog subscriber loop. The ITU-T recommends the use of the modified Intermediate Reference System (IRS) sending characteristic that describes the property of handset devices [7]. The filtering action simulates the frequency response of a typical telephone handset microphone. The similar process is carried out for the receiving part to simulate the frequency response of the handset earpiece. In order to reduce computational cost, the characteristics of handset in this simulation system are simulated by means of the Least P-norm optimal IIR filter method when ambient room noise is on a very low level.

2.4 Line Filter

The line filter represents analog subscriber line’s electrical characteristics that introduce frequency changes
to the carried signal. Transmission line theory can be used to qualify signal distortions [8]. For a given load impedance \(Z_L\), the transfer function of the line is determined as follows:

\[
T(w) = \frac{Z_L}{Z_L \cosh(\gamma d) + Z_0 \sinh(\gamma d)}
\]  

(2)

where \(d\) is the length of the line, \(\gamma\) and \(Z_0\) are determined by the series inductance \(L\), shunt capacitance \(C\), shunt conductance \(G\) and series resistance \(R\) of line and they are all defined per unit length of cable. The signal attenuation becomes more significant at higher frequencies for longer cables. With the line length set at a new value, new attenuation and frequency changes to the speech flow are obtained for the simulation behaviors.

### 2.5 PCM Encoder and Decoder

In the 64 kbits/s PCM structure, a PCM Encoder implements anti-alias filtering, sampling the received signal, and convert the sample into A-law/u-law binary codewords, which are decoded into analog signal and passed through a low-pass filter in a PCM Decoder. Elliptic filters offer steeper rolloff characteristics than Butterworth and Chebyshev filters, but they are equiripple in both the passband and the stopband. Elliptic filters usually meet a given set of filter performance specifications with the lowest filter order. In our model, the anti-alias filter implemented in the PCM encoder consists of a 5-order elliptic low-pass filter cascaded with a 3-order elliptic hi-pass filter. The low-pass filter \(H_l(z)\) and hi-pass filter \(H_h(z)\) in the simulation system are given in the following expressions:

\[
H_l(z) = \frac{0.4931 - 1.4931z^{-1} + 2.3699z^{-2} + 4.6484z^{-3} + 4.6484z^{-4} + 2.3699z^{-5}}{1 - 3.4636z^{-1} + 5.0773z^{-2} + 3.8205z^{-3} + 1.4535z^{-4} + 0.2079z^{-5}}
\]  

(3)

\[
H_h(z) = \frac{1.4931 - 2.3699z^{-1} + 2.3699z^{-2} + 4.6484z^{-3} + 4.6484z^{-4} + 2.3699z^{-5}}{1 - 2.5480z^{-1} + 2.1875z^{-2} - 6.259z^{-3}}
\]  

(4)

The magnitude response and group delay time of an anti-alias filter composed of such two filters are shown in Figure 2, which meets the requirements of ITU-T Rec. G712 [9]. The same elliptic low-pass filter as \(H_l(z)\) is used in the PCM Decoder. A desired frequency response can be obtained to a certain degree by adjusting the coefficients of the two filters and the channel filter combined with IRS filters.

### 2.6 Codec Schemes

In telephone transmission simulation, speech codec methods should be taken into consideration as an indispensable part. Moreover, the methods besides A-law/u-law PCM (G.711) standardized by the ITU-T and are likely to be used in the digital network. Such codec schemes include ADPCM series like G721, G726 and G.729. The ETSI (European Telecommunication Standards Institute) GSM6.10 is also incorporated into the system in order to identify the performance of different codecs schemes. Various codecs and their possible tandems are implemented in the simulation system.

Compared with the distortion caused by codec, the effect of signal-correlated noise generated by a modulated noise reference unit (MNRU) is also investigated on the location of the codec in this work. The distortion introduced by signal-modulated noise produces a similar cue as in typical speech codecs.

The experiments below will show that the encoding schemes, especially at low bit-rates, introduce quantizing distortion, which is a particular form of nonlinear distortion.

### 2.7 Echo and Sidetone Computation

In addition to the forward transmission path, Figure 1 shows that the impairments of sidetone loss and echo losses on the flowing speech signal. Sidetone losses occur in the sidetone path from microphone to earphone in the same handset. The listener and talker echoes occur at the far and near end hybrid circuit of interface in center office due to impedance mismatch. Listener sidetone with attenuation \(Lst\) is represented as follows:

\[
LSTR = STMR + D
\]  

(5)

where \(STMR = SLRset + RLSet + Lst - I\) reflects the loudness loss between a subscriber’s mouth and his ear(earphone) via the electric sidetone path. \(SLRset\) and
RLRset denote the handset characteristics as described before. D is a default value.

Talker echo with one-way delay T and attenuation Le can be computed by:

\[ TELR = SLR + RLR + Le \]  \hspace{1cm} (6)

where SLR (send loudness rating) and RLR (receive loudness rating) in this model are expressed as channel filters and IRS filters.

2.8 Noise Simulation

Speech quality are impaired by different kinds of noise during transmission, mainly including circuit noise, noise floor, impulse noise ambient room noise and crosstalk noise. These noise sources in this study are described by psophometrically or A-weighted power levels regarding telephone transmission planning.

In this model, circuit noise is modeled as narrow-band white noise distributed in a telephone connection, while noise floor at the receive side described as wide-band. The effects of ambient room noise or Hoth noise in receiving environment is simulated by adding the noise to the listener’s side in frequency domain.

Due to the weak influence of far end crosstalk (FEXT), the system only considers the effect on signal arising from the near end crosstalk (NEXT). NEXT is considered in simulation to be increased with \( f^{1.5} \), where \( f \) denotes frequency (Hz).

The impulse noise is characterized as a random pulse waveform whose amplitude is much higher than the background noise. The frequency of impulses is set between 1 and 5 per minute in our simulation approach.

Figure 3 shows an example of changing a portion of female speech from the DB863 database into simulated telephone version given a parameter set during one simulation operation. The figure indicates the background noise added into the speech as well as the change in frequency bandwidth.

3 Speech Database

A training set for the telephone speech recognizer consisting of 43,096 utterances of 180,626 seconds from 80 speakers was derived by means of a dual channel telephone board from the male DB863 database (The speech passed through such a processing board is assumed to be closer to the realistic situation). Six test sets were used for the following experiments. A speech set containing 240 utterances of 1,032 seconds from four male speakers extracted from the remaining of the DB863 was used as the original set. The first test set (+) was obtained by passing the original test set through a local telephone channel. The latter five test sets (A, B, C, D and E) for the experiments are listed in Table 1. They were formed by means of the simulation method described in the previous section with different parameter value sets. The five sets were selected for their similar quality to the real telephone situations by 20 persons in an informal listening test prior to recognition experiments.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Circuit Noise: Ne (dBm0p)</th>
<th>Hoth Noise: Pr (dBA)</th>
<th>Codec Mode</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-55</td>
<td>45</td>
<td>G711</td>
<td>Low Noise</td>
</tr>
<tr>
<td>B</td>
<td>-45</td>
<td>40</td>
<td>G711</td>
<td>Noise</td>
</tr>
<tr>
<td>C</td>
<td>-55</td>
<td>40</td>
<td>G726 (32kbits/s)</td>
<td>Noisy DCME</td>
</tr>
<tr>
<td>D</td>
<td>-55</td>
<td>45</td>
<td>GSM610</td>
<td>Noisy Mobile</td>
</tr>
<tr>
<td>E</td>
<td>-70</td>
<td>35</td>
<td>MNRU, Q=20dB</td>
<td>Sig. Cor. N</td>
</tr>
</tbody>
</table>

Table 1: Test sets derived from the clean set with different parameters set in the simulation model. The other parameters have been set to default values

4 Experiments and Results

The experimental objective is to discuss the feasibility of converting simulated data from clean speech and to verify the effective of the software simulation method. Acoustic features consist of Mel cepstral coefficients of 12 dimensions, the energy feature of one dimension, the normalized pitch of one dimension, and their corresponding first and second order time derivatives, thus forming a feature vector of 42 dimensions.
Acoustical models used 10,801 decision tree based tonal class-triphone models and 2,591 output pdfs were defined by 16 Gaussian mixtures. Language model consisting of backoff Trigrams of about 80Mb (45Mb Trigram and 35Mb Bigram) was trained from the whole DB863 database.

The experiments verified the effectiveness of the simulation method in the view of recognition test. The corresponding performance of each situation had been evaluated and the recognition results are listed in Table 2.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>corr</th>
<th>sub</th>
<th>del</th>
<th>ins</th>
<th>err</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>74.3</td>
<td>22.8</td>
<td>2.9</td>
<td>0.0</td>
<td>25.7</td>
</tr>
<tr>
<td>A</td>
<td>74.1</td>
<td>22.5</td>
<td>3.5</td>
<td>0.3</td>
<td>26.3</td>
</tr>
<tr>
<td>B</td>
<td>68.0</td>
<td>29.7</td>
<td>4.3</td>
<td>1.9</td>
<td>35.9</td>
</tr>
<tr>
<td>C</td>
<td>76.8</td>
<td>19.8</td>
<td>3.3</td>
<td>0.1</td>
<td>23.2</td>
</tr>
<tr>
<td>D</td>
<td>73.7</td>
<td>23.0</td>
<td>3.3</td>
<td>0.3</td>
<td>26.5</td>
</tr>
<tr>
<td>E</td>
<td>74.3</td>
<td>22.6</td>
<td>3.1</td>
<td>0.1</td>
<td>25.8</td>
</tr>
</tbody>
</table>

Table 2: Recognition performance, obtained on the different test sets

As is seen in the table, the recognition correct rate (corr) of the B set is merely 68.0% while 74% or so is for the others with the same recognizer. When the recognition rate of the real set (+) is contrasted with those of test sets generated by the software simulation, the results indicate that the approximate recognition rates to real telephone speech test set (+) can be obtained by setting different parameters for the model. In other words, this simulation method can generate telephony noisy speech approximate to the real telephone speech by adjusting parameters.

By comparing B with A, C and D, the simulation results show that circuit noise in telephone transmission channel and ambient room noise determine the telephone speech quality in an important degree. Properly simulating different noises is a key problem in this simulation system. With A (G711) and D (GSM610) compared, the values show that code methods affect the recognition performance of one recognizer, but the difference is not prominent due to the effect of background noise. For the tests that have similar correction rates, the error rates substitution (sub), deletion (del), insertion (ins) and all errors (err) from the table are close in numerical value too.

5 Concluding Remarks

In this paper, we addressed our approach to generate simulated telephone-quality speech from clean corpus. Recognition results show the effectiveness and feasibility of the software simulation. The design emphasis focused on simulating the frequency characteristics of analogue facilities and different noise behaviors in telephone transmission channel. By controlling the parameter values, the approach can approximate clean speech to different actual telephone situations, and demonstrating the tendency towards variety in telephone channel characteristics in itself. It provides an operable and economical solution to the problem on insufficiency in real telephone data. To simulate telephone speech precisely, more investigations must be directed at actual transmission physical properties, noise effects and channel time-variant characteristics, and some methods used in voice conversion technology may be employed in telephone speech simulation.

6 References