EXPERIMENTS ON RECOGNITION OF LAVALIER MICROPHONE SPEECH AND WHISPERED SPEECH IN REAL WORLD ENVIRONMENTS

Kiyoshi Tatara, Taisuke Ito, Parham Zolfaghari, Kazuya Takeda & Fumitada Itakura

Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603 Japan
tatara@itakura.nuee.nagoya-u.ac.jp
<http://www.itakura.nuee.nagoya-u.ac.jp/people/tatara>.

ABSTRACT

In this paper, we present corpora and recognition experiments of the speech recorded in everyday life for the real world speech recognition. A speech corpus of 8,600 sentences from 53 speakers recorded through lavalier microphones in four different environments is built. The data was collected in an office space, a sound-proof room, in cars of different sizes, and on the street. Another corpus consisting of whispered and normal speech of more than 6,000 sentences from 100 speakers recorded through a close-talking microphone is devised. Continuous speech recognition experiments using acoustic models trained by the speech corpus in each environment, attain a recognition accuracy of above 80%. For the whispered speech corpus, the recognition accuracy obtained was 74% using the whispered speech model.

1. INTRODUCTION

In everyday life, we use speech for communication and other purposes in various environments. In recent times, speech recognition systems have improved considerably and speech data acquisition is an important factor in this improvement [1]. At some point, we expect to utilize such speech recognition capabilities as we speak in everyday life in real world conditions.

One of the aims in this research is to construct a corpus of speech recorded through lavalier microphones. The reason for considering the use of lavalier microphones is due to the uncomfortable nature of the close-talking microphones. Users have to use a headset for such microphones while lavalier microphones can be attached to any part of the body. They are also lightweight and the SNR difference between them and close-talking microphones is not significant if they are placed near the mouth area.

Recording in various environments plays an important part in building speech recognizers applicable in such environments. Here, we have used lavalier microphones to record phonetically balanced Japanese sentences in a sound-proof room, in an office space, on the street and inside cars of varying sizes. We will present recognition results using HMM based recognizers for these real world environments.

During communication, we sometimes whisper when speaking aloud is a nuisance in public, or when we do not want the contents of the conversation to be heard by others. In such a quiet or private communication, the use of whispered speech is considered effective. Therefore, we also construct a database of whispered speech in order to make recognition systems for such private communication scenarios.

In this paper, we first describe the details of the lavalier microphone speech corpus followed by the whispered speech corpus. Recognition experiments using these corpora are then given in sections 4 and 5.

2. LAVALIER MICROPHONE SPEECH CORPORA

Speech was recorded by using Sony ECM77B lavalier microphones. This is an ultra mini omnidirectional electret condenser lavalier suitable for many different applications, ranging from recording of news and interviews to recording in theaters and for instrument pick-up. Its frequency response is 40-20 kHz with upper range lift for extra presence. Directivity is optimized to ensure uniform output, regardless of direction of the sound source. The metal mesh windshield effectively eliminates both outdoors wind noise and “popping” in close miced situations. It is 5.6mm in diameter and 12.55mm in length.

Each subject was equipped with two lavalier microphones. One was attached to the frame of the provided spectacles and the other was attached to the subject’s shirt around the chest area. The recording scene is as shown in Figure 1. Input speech was quantized to 16 bits, and sampled at 48 kHz.
2.1. Recording Environments

This database was constructed to carry out recognition experiments for real world applications. Speech was recorded in four different environments which included recording in a sound-proof room, in an office space, on a street, and inside a car. The noise level of the sound-proof room was about 22 dB(A), and the reverberation time was approximately 150 ms. Three types of cars were used: a Sedan, a station wagon, and a one-box type car. A driver was instructed to drive these cars in the Nagoya city suburbs. The subjects sat in the passenger seat and were instructed to read a list of phonetically balanced sentences. The traffic on the street was relatively heavy.

2.2. Sentence Composition

For constructing this speech corpus, we used 10 sets of ATR phonetically balanced Japanese sentences [2], consisting of a total of 503 sentences. The Acoustical Society of Japan (ASJ) continuous speech corpus (Japanese Newspaper Article Sentences: JNAS) [3], consisting of 100 sentences in total was also used.

Each subject read one set from the ATR phonetically balanced sentences and 5 sentences from 100 JNAS sentences. Phonetically balanced sentences were used for to build the training data, and the JNAS sentences were used for the test data. Based on this recording method, each speaker read 60 sentences in each environment. In total, 53 speakers (26 males and 27 females) participated in building this database.

3. WHISPERED SPEECH CORPORAS

Whispered speech is produced by speaking without vibration of the vocal cords. Since exhalation is the source of sound in whispered speech, its acoustic characteristics differ from those of normal speech. In particular, the magnitude (power) in the low-frequency region of whispered speech is weaker than that in normal speech. Therefore, the signal-to-noise ratio (SNR) of whispered speech in a real environment where the background noise is present is low. Accordingly, whispered speech recognition is considered to be more difficult.

We have built a speech corpus consisting of whispered speech and normal speech of more than 6,000 sentences from 123 speakers.

3.1. Recording Method

Whispered and normal speech utterances were recorded in the same sound-proof room as the lavalier microphone corpus, using a DV camera and a close-talking microphone (Sennheiser HMD410). The sampling rate used was 48 kHz with 16 bit quantization. Whispered speech was also recorded from a cellular-phone with and without covering the mouth (Figure 2).

3.2. Sentence Composition

One hundred and twenty three speakers (68 males and 55 females) participated in speech recording. They produced both normal speech and whispered speech. Each speaker read one set (50 sentences) from sets A to I in ATR phonetically balanced Japanese sentences. For the test data, 50 sentences from newspaper articles were used.

4. RECOGNITION OF LAVALIER MICROPHONE SPEECH

In this section, we conduct recognition experiments using the lavalier microphone speech corpus. The recognition model is the hidden Markov Model (HMM).
4.1. Recognition Model

Four HMMs were trained for each environment. For training each model 2150 sentences were used from 43 speakers (21 males and 22 females). The other remaining conditions are summarized in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Recognition Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling rate</td>
</tr>
<tr>
<td>Window</td>
</tr>
<tr>
<td>Frame length</td>
</tr>
<tr>
<td>Frame shift</td>
</tr>
<tr>
<td>Feature Vector</td>
</tr>
<tr>
<td>HMM</td>
</tr>
<tr>
<td>Number of states</td>
</tr>
<tr>
<td>Training data</td>
</tr>
</tbody>
</table>

4.2. Recognition Experiments

HMM based continuous speech recognition experiments were performed using the lavalier microphone speech corpus with a vocabulary size of 20,000. The test data were 50 JNAS sentences spoken by 10 speakers (5 males and females each) different from those who recorded the training data. The decoder used was julius3.1 [4]. Recognition performance is given by the word correct rate (% correct):

\[
\%\text{Corr} = \frac{N - S - D}{N} \times 100[\%]
\]

where \(N\) is the number of words, \(S\) is the number of substitutions, and \(D\) is the number of deletions. For comparison, we considered the IPA\(^1\) standard Japanese HMM. This model has 2,000 states, with 16 mixtures, and is of gender independent triphone type.

4.3. Recognition Results

The results obtained from the HMM continuous speech recognition tests are shown in Figures 3 and 4. In these figures, ‘IPA’ refers to the IPA standard Japanese HMM. ‘Train’ refers to the HMM trained using 2,150 sentences in the same environment as sentences in the test set. ‘Train ALL’ refers to the HMM trained by 8,600 sentences in all environment conditions. Compared with the results of IPA standard HMM, it is evident that the recognition rates using environment-dependent HMMs are better than those using IPA HMMs in all environments except for the sound-proof room. Especially, at such noisy environments as the street and inside the car, the recognition rates are over 10% better than those using IPA and only 2,150 sentences were used. The recognition rate of speech utterances recorded in the sound-proof room is about 6% higher if the IPA standard HMM is used rather than the HMM trained by lavalier microphone speech utterances. This is because

1. In the sound-proof room, there is little difference between lavalier microphone speech and close-talking microphone speech.
2. IPA standard HMM is trained with about 20 times more training data than the HMMs trained here.

Fig. 3. Recognition results for speech in different environment recorded by a lavalier microphone attached to the spectacles.

5. WHISPERED SPEECH RECOGNITION

In this section, we conduct recognition experiments using the close-talking microphone whispered speech corpus described above[5]. The recognition model is again the hidden Markov Model (HMM).

5.1. Recognition Model

HMMs trained by normal speech and HMM trained by whispered speech were built. For training the whispered speech model, 4,000 sentences were used from 80 speakers (40 males and 40 females). Other remaining conditions are summarized in Table 2.

For comparison with the whispered HMM, a normal speech HMM was also trained using 14,000 phonetically balanced Japanese sentences of 276 speakers (138 males and 138 females) from JNAS speech database [3]. This HMM has 129 states, with 16 mixtures, and is of gender independent monophone type.
Fig. 4. Recognition results for speech in different environments recorded by a lavalier microphone attached to the shirt around the chest area.

<table>
<thead>
<tr>
<th>Table 2. Recognition Model Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Sampling rate</td>
</tr>
<tr>
<td>Window</td>
</tr>
<tr>
<td>Frame length</td>
</tr>
<tr>
<td>Frame shift</td>
</tr>
<tr>
<td>Feature Vector</td>
</tr>
<tr>
<td>HMM</td>
</tr>
<tr>
<td>Number of states</td>
</tr>
</tbody>
</table>

5.2. Recognition Experiments

Using the whispered and normal speech HMMs, continuous speech recognition experiments were performed with a vocabulary size of 20,000. The test data comprised of 200 whispered and normal speech JNAS sentences spoken by 4 speakers (2 males and females each). The decoder used was julius-3.1.

5.3. Recognition Results

The recognition rates of normal speech and whispered speech using the normal speech model and whispered speech model is shown in Table 3. The benchmark result for this experiment was the 87% recognition rate obtained for normal speech using the normal speech model. However, a 74% recognition rate was obtained for whispered speech using the whispered speech model.

We also used the normal speech model for recognizing the whispered speech. This gave a 27% recognition rate which was significantly lower than those in the cases where the same speech style was used in the acoustic model and the evaluation sentences. Also this reduction in recognition rate was found to be larger than the case where the whispered speech model was used for recognizing normal speech.

<table>
<thead>
<tr>
<th>Table 3. Recognition rates of normal speech and whispered speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Speech Models</td>
</tr>
<tr>
<td>Normal Speech Model</td>
</tr>
<tr>
<td>Whisper Speech Model</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

In this paper, we have presented recognition results for real world environments and whispered speech. We built normal speech corpora of 8,600 sentences recorded through lavalier microphones in four different environments. The whispered speech corpora consisted of both whispered speech and normal speech of more than 6,000 sentences recorded through a close-talking microphone in a sound-proof room.

From the experimental results, the effectiveness of using lavalier microphone for obtaining the speech in everyday life was demonstrated. We also showed that even whispered speech can be modeled by HMM when a large training corpus is available.

7. REFERENCES