The CUHK Dysarthric Speech Recognition Systems for English and Cantonese

Shoukang Hu1*, Shansong Liu1*, Heng Fai Chang2*, Mengzhe Geng1, Jiani Chen1, Lau Wing Chung1, To Ka Hei1, Jianwei Yu1, Ka Ho Wong1, Xinying Liu1, Helen Meng1

1The Chinese University of Hong Kong, Hong Kong SAR, China
2University of Cambridge, Cambridge, UK

{akhu,ssliu,xyliu,hfmc38}@se.cuhk.edu.hk, hfc38@cam.ac.uk

Abstract

Speech disorders affect many people around the world and introduce a negative impact on their quality of life. Dysarthria is a neural-motor speech disorder that obstructs the normal production of speech. Current automatic speech recognition (ASR) systems are developed for normal speech. They are not suitable for accurate recognition of disordered speech. To the best of our knowledge, the majority of disordered speech recognition systems developed to date are for English. In this paper, we present two disordered speech recognition systems for both English and Cantonese. Both systems demonstrate competitive performance when compared with the Google speech recognition API and human recognition results.

Index Terms: dysarthria, speech disorders, speech recognition

1. Introduction

Speech disorders affect many people around the world. The difficulty in communication introduces a negative impact on their quality of life. People with speech disorders often co-occur with disabilities and thus it is difficult for them to use computers and touch-screen devices. Despite the degradation of voice quality, speech-based assistive technologies provide a natural solution to the problem. However, due to the large mismatch between disordered and normal speech, current automatic speech recognition (ASR) systems primarily designed for healthy speakers are not suitable for disordered speech.

There has been increasing research interest for improving the performance of disordered speech recognition systems over the years. The underlying techniques have evolved from the early generation of hidden Markov model (HMM) based framework [1, 2] to the current deep neural network (DNN) based techniques [3, 4]. A range of systems have been developed and defined state-of-the-art disordered ASR performance. These include the University of Sheffield systems [2, 5] using speaker and cross-domain adaptation techniques, and the Chinese University of Hong Kong system [4] constructed using a combination of multiple neural network models. However, these systems were only developed for English. To the best of our knowledge, there has been very limited development of disordered ASR systems for Cantonese.

In this paper, we present two disordered speech recognition systems developed for the English UASpeech database [5] and Cantonese CUDYS corpus [6]. A common modelling technique used in both systems and based on gated neural network (GNN) [7] was used to allow robust integration of acoustic features with visual features as well as optionally the prosody features based on pitch. A schematic overview of these two systems is shown in figure 1. The demo systems compare the performance of the CUHK systems for both languages against the Google speech recognition API and human recognition results based on multiple listening subjects.

2. Task Description

The English UASpeech [5] is an isolated word recognition task including 16 dysarthric speakers, and 8 out of the 16 speakers were provided with both audios and videos. We used these 8 speakers’ audio-visual data to construct our recognition systems. All speakers were required to repeat 455 distinct words. These words were distributed into three blocks. Block 1 (B1) and block 3 (B3) were treated as the training set, leaving the block 2 (B2) as the test set.

The CUDYS [6] is a Cantonese dysarthric speech corpus collected by the Chinese University of Hong Kong, containing three tasks, which are words, sentences and paragraphs. Word-level and sentence-level tasks were used in this paper. To facilitate training due to limited data source, an external Cantonese normal speech corpus (SpeechOcean, 205.8 hours) was incorporated and mixed with CUDYS (3.3 hours). The training set contains 23 speakers from CUDYS and 441 speakers from SpeechOcean. The test set contains 10 speakers from CUDYS.

3. UASpeech System for English

In the UASpeech disordered speech recognition task, visual features were incorporated to provide complementary information to the speech recognizer. The key difficulty of using visual features is that the underlying medical conditions of the dysarthric speakers can cause quality degradation of the recorded videos due to situations such as uncontrollable head movements. To address this issue, a face alignment network [8] was used to track the lip motions followed by an affine transformation to make the detected lip regions horizontal. Autoencoder and linear discriminant analysis were then applied sequentially on the 128×128 pixels lip figures to generate visual feature vectors.

In order to obtain a robust integration of acoustic and visual features, a novel Bayesian GNN ASR architecture was employed to dynamically weight the contributions from visual features as well as model the uncertainty given limited and variable disordered speech data. Details of this model can be found in [9]. Speaker dependent systems were trained using concatenated 105-dimension audio-visual features for the 8 dysarthric speakers. The networks were set to 5 hidden layers with 500 neurons in each hidden layer. For recognition, a word grammar network was used to produce recognition results.

4. CUDYS System for Cantonese

In the CUDYS disordered speech recognition task, we used pitch features as an additional source to assist standard acoustic features. Medical conditions like damages to cerebellum can cause uncoordinated muscle movements of articulators, in-
introducing a large effect on prosody control. Hence, simple concatenation of the two can lead to recognition performance degradation. Therefore, a GNN based approach was explored to improve the integration of acoustic and pitch features and ultimately improve the recognition performance [10].

Speaker independent GNN acoustic model was trained for the mixed training set using concatenated 80-dimension log filter bank plus pitch features. The network consists of 6 hidden layers with 2000 neurons in each hidden layer. The lexicon used in this task contains 84K words. For continuous speech recognition, a four-gram interpolated language model was trained on a mixed transcription set containing SpeechOcean (2M), CUDYS (281 words) and web text sources (30M). For isolated word recognition, a word grammar network was used, in common with the UASpeech task.

5. Demo System and Results Analysis

In our demo, we present two disordered speech recognition systems for both English UASpeech and Cantonese CUDYS. To protect the speakers’ privacy, the faces were obscured. We compare the performance of our systems with the Google speech recognition API and human recognition results. The Google API receives audios and produces recognized outputs, while human recognition results are produced by multiple listening subjects. For each dysarthric speaker, we list the corresponding transcript and the symptoms name in our demo’s GUI interface. For each dysarthric speaker, we list the corresponding

Table 1: Performance of CUHK systems, Google speech recognition API and human recognition on 6 impaired speakers for UASpeech English and CUDYS Cantonese corpora. CP represents Cerebral palsy. SCA represents Spinocerebellar Ataxia.

<table>
<thead>
<tr>
<th>ID</th>
<th>CUHK</th>
<th>Google API</th>
<th>Human</th>
<th>Dysarthria</th>
</tr>
</thead>
<tbody>
<tr>
<td>F04</td>
<td>24.0</td>
<td>50.6</td>
<td>58.1</td>
<td>Athetoid CP</td>
</tr>
<tr>
<td>M12</td>
<td>54.0</td>
<td>100.0</td>
<td>94.0</td>
<td>Mixed CP</td>
</tr>
<tr>
<td>M14</td>
<td>15.6</td>
<td>35.8</td>
<td>16.1</td>
<td>Spastic CP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>CUHK</th>
<th>Google API</th>
<th>Human</th>
<th>Dysarthria</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0006</td>
<td>1.7</td>
<td>20.6</td>
<td>11.2</td>
<td>SCA</td>
</tr>
<tr>
<td>S0015</td>
<td>92.7</td>
<td>93.4</td>
<td>53.3</td>
<td>Unknown CP</td>
</tr>
<tr>
<td>S0030</td>
<td>23.6</td>
<td>63.1</td>
<td>5.6</td>
<td>SCA</td>
</tr>
</tbody>
</table>

Figure 1: A schematic overview of the two demo systems—English UASpeech and Cantonese CUDYS.

6. References


