INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

The CUHK Dysarthric Speech Recognition Systems for English and Cantonese

Shoukang Hu[1*], *Shansong Liu*[1*], *Heng Fai Chang*[2*], *Mengzhe Geng*[1], *Jiani Chen*[1], *Lau Wing Chung*[1], *To Ka Hei*[1], *Jianwei Yu*[1], *Ka Ho Wong*[1], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]University of Cambridge, Cambridge, UK
{skhu,ssliu,xyliu,hmmeng}@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

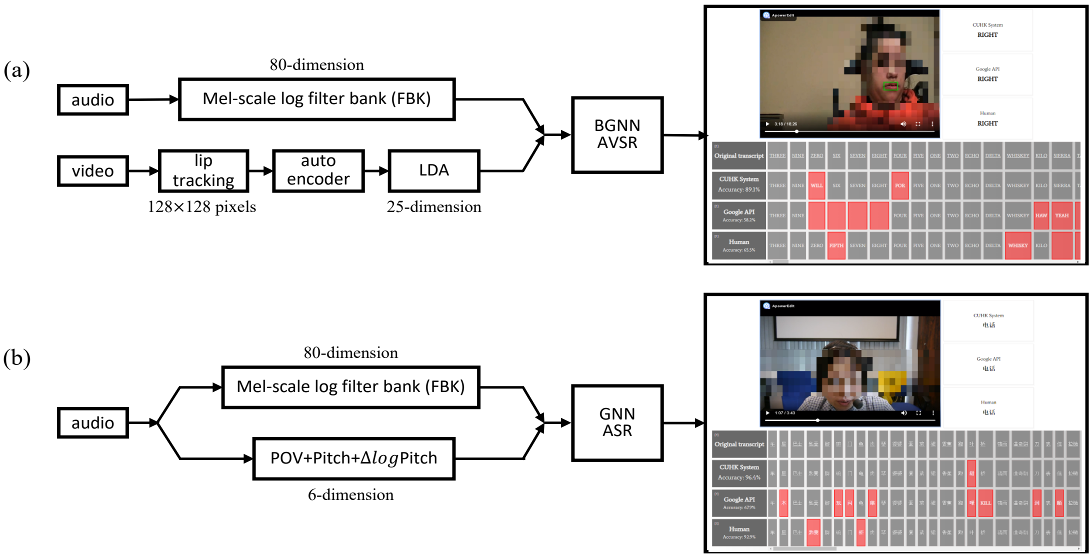
In the UASpeech disordered speech recognition task, visual features were incorporated to provide complementary information to the speech recognizer. The key difficulty of using visual features is that the underlying medical conditions of the dysarthric speakers can cause quality degradation of the recorded 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 AVSR 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-

[*]Equal contribution.



CUHK System RIGHT

Google API RIGHT

Human RIGHT

Original transcript	THREE	NINE	ZERO	SIX	SEVEN	EIGHT	FOUR	FIVE	ONE	TWO	ECHO	DELTA	WHISKEY	KILO	SIERRA
CUHK System Accuracy: 89.1%	THREE	NINE	WILL	SIX	SEVEN	EIGHT	FOR	FIVE	ONE	TWO	ECHO	DELTA	WHISKEY	KILO	SIERRA
Google API Accuracy: 58.2%	THREE	NINE					FOUR	FIVE	ONE	TWO	ECHO	DELTA	WHISKEY	HAW	YEAH
Human Accuracy: 65.5%	THREE	NINE	ZERO	FIFTH	SEVEN	EIGHT	FOUR	FIVE	ONE	TWO	ECHO	DELTA	WHISKY	KILO	

CUHK System 电话

Google API 电话

Human 电话

Original transcript	车	屋	巴士	匙羹	脚	碗	门	龟	肉	琴	婆婆	蛋	菜	裙	香蕉	鸡	叶	桥	落雨	曲奇饼	刀	表	信	拉链
CUHK System Accuracy: 96.4%	车	屋	巴士	匙羹	脚	碗	门	龟	肉	琴	婆婆	蛋	菜	裙	香蕉	鸡	甜	桥	落雨	曲奇饼	刀	表	信	拉链
Google API Accuracy: 67.9%	车	木	巴士	匙羹	脚	玩	问	龟	葉	琴	婆婆	蛋	菜	裙	香蕉	鸡	咩	KILL	落雨	曲奇饼	到	表	顺	拉链
Human Accuracy: 92.9%	车	屋	巴士	[illegible]	脚	碗	门	柜	肉	琴	婆婆	蛋	菜	裙	香蕉	鸡	叶	桥	落雨	曲奇饼	刀	表	信	拉链

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

troducing 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 86-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 [1] 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. All speakers in our two disordered speech recognition tasks are adults. No children were involved. Table 1 shows the word or character error rates (WERs or CERs) produced by our CUHK systems, Google API and human beings for different dysarthric speakers respectively. For both languages, the CUHK systems consistently outperform the Google speech recognition API. On the UASpeech English data, our system also outperforms human recognition across all speakers.

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.*

UASpeech(Word Error Rate%)				
ID	CUHK	Google API[1]	Human	Dysarthria
F04	**24.0**	50.6	58.1	Athetoid CP
M12	**54.0**	100.0	94.0	Mixed
M14	**15.6**	35.8	16.1	Spastic CP
CUDYS(Character Error Rate%)				
ID	CUHK	Google API[1]	Human	Dysarthria
S0006	**1.7**	20.6	11.2	SCA
S0015	92.7	93.4	53.3	Unknown CP [6]
S0030	23.6	63.1	5.6	SCA

6. References

[1] J. Deller Jr, D. Hsu, and L. J. Ferrier, "On the use of hidden markov modelling for recognition of dysarthric speech," *COMPUT METH PROG BIO*, vol. 35, no. 2, pp. 125–139, 1991.

[2] H. Christensen, S. Cunningham, C. Fox, and et al., "A comparative study of adaptive, automatic recognition of disordered speech," in *INTERSPEECH*, 2012.

[3] H. Christensen, M. Aniol, P. Bell, and et al., "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *INTERSPEECH*, 2013.

[4] J. Yu, X. Xie, S. Liu, and et al., "Development of the cuhk dysarthric speech recognition system for the uaspeech corpus," *INTERSPEECH*, 2018.

[5] H. Kim, M. Hasegawa, A. Perlman, and et al., "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008.

[6] K. H. Wong, Y. T. Yeung, E. H. Chan, and et al., "Development of a cantonese dysarthric speech corpus," in *INTERSPEECH*, 2015.

[7] P. Ghahremani, B. BabaAli, D. Povey, and et al., "A pitch extraction algorithm tuned for automatic speech recognition," in *ICASSP*. IEEE, 2014, pp. 2494–2498.

[8] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017.

[9] L. Shansong, H. Shoukang, W. Yi, and et al., "Exploiting visual features using bayesian gated neural networks for disordered speech recognition," in *submission to interspeech 2019*.

[10] L. Shansong, H. Shoukang, X. Liu, and et al., "On the use of pitch features for disordered speech recognition," in *submission to interspeech 2019*.

[1] The online Google speech recognition API can be found at https://cloud.google.com/speech-to-text