INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

On the Use of Pitch Features for Disordered Speech Recognition

Shansong Liu[1*], *Shoukang Hu*[1*], *Xunying Liu*[1], *Helen Meng*[1]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
{ssliu,skhu,xyliu,hmmeng}@se.cuhk.edu.hk

Abstract

Pitch features have long been known to be useful for recognition of normal speech. However, for disordered speech, the significant degradation of voice quality renders the prosodic features, such as pitch, not always useful, particularly when the underlying conditions, for example, damages to the cerebellum, introduce a large effect on prosody control. Hence, both acoustic and prosodic information can be distorted. To the best of our knowledge, there has been very limited research on using pitch features for disordered speech recognition. In this paper, a comparative study of multiple approaches designed to incorporate pitch features is conducted to improve the performance of two disordered speech recognition tasks: English UASpeech, and Cantonese CUDYS. A novel gated neural network (GNN) based approach is used to improve acoustic and pitch feature integration over a conventional concatenation between the two. Bayesian estimation of GNNs is also investigated to further improve their robustness.

Index Terms: pitch, disordered speech, speech recognition

1. Introduction

Pitch, as a perceptual measurement of fundamental frequency (F0) of speech signals [1], is a powerful prosodic cue for auditory perception. Pitch features have long known to be useful for recognition of normal speech, especially for tonal languages, such as Mandarin [2, 3, 4], Cantonese [5, 6], Vietnamese [7, 8] and Thai [9, 10], since pitch can serve as an informative source to distinguish different tones in tonal languages [11]. In non-tonal languages, for instance, English [12, 13, 14] and Japanese [15, 16], it is also feasible to treat pitch as an auxiliary information by concatenating with acoustic features to improve speech recognition performance.

However, for disordered speech, the deterioration on voice quality renders the prosodic features, such as pitch, not always useful. Medical conditions, for example, damages to cerebellum in the Ataxic dysathria [17], can cause uncoordinated muscle movements of articulation organs to affected patients, introducing a large effect on prosody control [18]. Another example is the Spastic dysarthria, which is caused by damages to the pyramidal tract [17]. In this case, a low pitch with pitch breaks often occurs. Therefore, an appropriate integration of acoustic and pitch features under such harsh acoustic conditions is a pressing need for disordered speech recognition.

Speech recognition for disordered speech is a challenging task in general [19]. Acoustic features play a major role in early and recent studies of disordered speech recognition [20, 21, 22, 23]. To the best of our knowledge, the vast majority of speech recognition systems using pitch features are conducted on normal speech, very limited research incorporating pitch features has been found for disordered speech. In this paper, we investigate multiple approaches to explore a robust integration of pitch features to improve the performance of two disordered speech recognition tasks: English UASpeech [24] and Cantonese CUDYS [25]. A novel gated neural network (GNN) [26] is used to improve the integration of acoustic and pitch features over the conventional feature concatenation between the two. A more advanced form, Bayesian gated neural network (BGNN) using Bayesian estimation of GNNs [27], is also investigated to further improve their robustness.

The main contributions of this paper are summarized below. First, as far as we know, this is the first work that incorporates pitch features for disordered speech recognition. Second, this paper presents the first attempt to leverage GNN and BGNN approaches for prosodic feature selection to improve the performance of disordered speech recognition and speech recognition systems in general. Experiments conducted on the two corpus, UASpeech and CUDYS, suggest that a selection mechanism is needed for a robust integration of acoustic and pitch features for disordered speech recognition tasks.

The rest of the paper is structured as follows. Section 2 and section 3 describe the pitch extraction algorithm and the models investigated in this paper. The details of experimental setup, results and analysis are described in section 4. The last section concludes the paper.

2. Pitch Extraction

There has been a long history developing the pitch extraction algorithm in previous research [28, 29, 30, 31, 32]. In this paper, we apply the Kaldi pitch tracker [32] in the Kaldi speech recognition toolkit [33] to obtain pitch and Probability of Voicing (POV) features for the experiments conducted on the aformentioned two disordered speech recognition tasks. The Kaldi pitch tracker is a highly modified version of getf0 (RAPT) algorithm introduced in [29]. Unlike the original getf0 algorithm where a hard decision has to be made to determine whether a given frame is voiced or unvoiced, the Kaldi pitch tracker assigns a pitch to all the frames. The algorithm also produces a quantity that can be used as a probability of voicing measure, which is based on the normalized autocorrelation measure. The detailed computational steps of the pitch features can be found in [32].

Framel level Normalized Cross Correlation Function (NCCF) values are further processed to obtain POV features. The resulting pitch features are normalized by the short-time mean subtraction approach [34] with POV weighting. Then the delta-log-pitch features computed directly from the unnormalized log pitch is added to the POV and normalized pitch features to form a 3-dimension pitch feature vector, which was used throughout the experiments of this paper.

3. Incorporating Pitch Feature

In our neural network acoustic models, we concatenate 40-dimension filter banks (FBKs) with 3-dimension pitch feature vectors, then further concatenate them with their first order dif-

[*]Equal contribution.

http://dx.doi.org/10.21437/Interspeech.2019-2609

ferential parameters to obtain 86-dimension feature vectors as the network input, shown in Fig. 1 and Fig. 2. Given an input vector $\mathbf{z}_t^{(l-1)}$ from $(l-1)$-th layer at t-th frame, a standard DNN system computes the output $h_i^{(l)}(\mathbf{z}_t^{(l-1)})$ of the i-th node in the l-th layer using Eqn. (1).

$$h_i^{(l)}(\mathbf{z}_t^{(l-1)}) = \phi\left(\boldsymbol{\theta}_i^{(l)} \bullet \mathbf{z}_t^{(l-1)}\right) \tag{1}$$

where $\mathbf{z}_t^{(l-1)} = \left[h_1^{(l-1)}(\mathbf{z}_t^{(l-2)}), \cdots, h_d^{(l-1)}(\mathbf{z}_t^{(l-2)}), 1\right]$ is the input vector fed into the l-th hidden layer, $\boldsymbol{\theta}_i^{(l)} = \left[w_{i,1}^{(l)}, \cdots, w_{i,d}^{(l)}, b^{(l)}\right]$ denotes the node's weight vector, $\phi(\cdot)$ is the activation function, and $\bullet$ denotes the dot product.

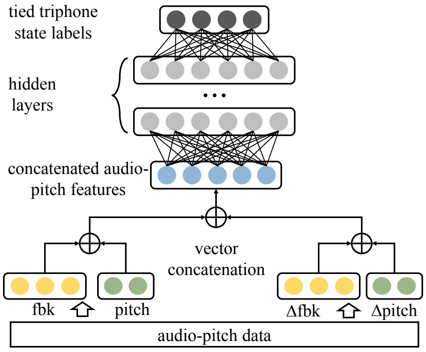

Figure 1: *The framework of the standard DNN based architecture. Acoustic and pitch features are concatenated with their first order differential parameters at the input layer before they are fed into the subsequent hidden layers.*

Acoustic and pitch features are concatenated at the input layer, where $\mathbf{z}_t^{(0)} = [\mathbf{x}_t^a \oplus \mathbf{x}_t^p, 1]$. $\mathbf{x}_t^a$ and $\mathbf{x}_t^p$ are the acoustic and pitch feature vectors of t-th frame, respectively. $\oplus$ denotes the vector concatenation operation. The output layer targets are tied triphone state labels.

In this paper, we focus on the incorporation of distorted pitch features to improve speech recognition robustness, hence the gated control is only applied to the pitch features. The gated input layer outputs $\mathbf{z}_t^{(0)}$ are computed as:

$$\begin{aligned} \mathbf{z}_t^p &= [\mathbf{x}_t^p, 1] \\ h_i^{(0)}(\mathbf{z}_t^p) &= \phi\left(\boldsymbol{\theta}_i^{(0)} \bullet \mathbf{z}_t^p\right) \\ \mathbf{z}_t^{(0),p} &= \mathbf{x}_t^p \otimes h^{(0)}(\mathbf{z}_t^p) \\ \mathbf{z}_t^{(0)} &= [\mathbf{x}_t^a \oplus \mathbf{z}_t^{(0),p}, 1] \end{aligned} \tag{2}$$

where the gating layer is denoted as the 0-th hidden layer. $\otimes$ and $\oplus$ denote the element-wise multiplication and vector concatenation, respectively. The activation function $\phi(\cdot)$ is a sigmoid function, whose outputs vary between 0 and 1. The pitch and Δpitch shown in Fig. 2 share the same gating parameters.

The investigated Bayesian gated neural network (BGNN) proposed in [27], which was also submitted to interspeech 2019, is described in Fig. 2. A more detailed model description can be found in this submitted paper [27]. Compared to standard DNN system, a gating layer is placed at the input layer to dynamically weight the contributions from pitch features. A posterior distribution over the gating parameters is also applied to model the uncertainty given limited and variable disordered speech data.

The general form of the hidden output with Bayesian learning is as follows:

$$h_i^{(l)}(\mathbf{z}_t^{(l-1)}) = \int \phi\left(\boldsymbol{\theta}_i^{(l)} \bullet \mathbf{z}_t^{(l-1)}\right) p(\boldsymbol{\theta}_i^{(l)}) d\boldsymbol{\theta}_i^{(l)} \tag{3}$$

where $p(\boldsymbol{\theta}_i^{(l)}) = p(\boldsymbol{\theta}_i^{(l)} \mid \{\mathbf{x}_t, \widehat{\mathbf{y}}_t\})$ denotes the node dependent activation parameter posterior distribution to be learned from training data $\{\mathbf{x}_t, \widehat{\mathbf{y}}_t\}$ ($\mathbf{x}_t, \widehat{\mathbf{y}}_t$ are the input data and its corresponding triphone state label at t-th frame). In our scenario, we only perform Bayesian learning on the gating parameters, hence the Eqn. (3) can be rewritten as the specialized form Eqn. (4).

$$h_i^{(0)}(\mathbf{z}_t^p) = \int \phi\left(\boldsymbol{\theta}_i^{(0)} \bullet \mathbf{z}_t^p\right) p(\boldsymbol{\theta}_i^{(0)}) d\boldsymbol{\theta}_i^{(0)} \tag{4}$$

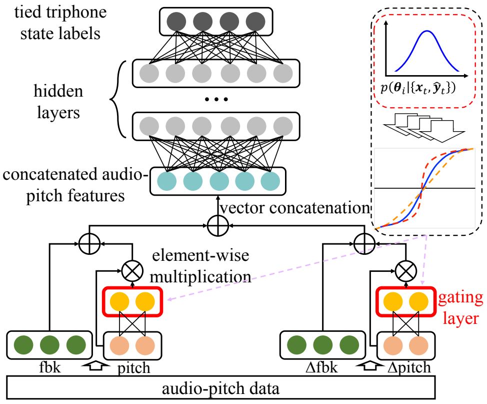
tied triphone state labels
hidden layers
concatenated audio-pitch features
$p(\boldsymbol{\theta}_i|\{x_t, \hat{y}_t\})$
vector concatenation
element-wise multiplication
gating layer
fbk
pitch
Δfbk
Δpitch
audio-pitch data

Figure 2: *A Bayesian gated DNN based system architecture. In contrast to the conventional gated DNN, a posterior distribution (top right corner) over the gating parameters is used to model the uncertainty given limited and variable pitch data.*

To estimate the hyper-parameters of $p(\boldsymbol{\theta}_i^{(0)})$, two addtional steps need to be included in the standard back-propagation algorithm. One is to calculate the variational lower bound approximate over the integration of the model parameters, the other is to perform a sampling step on the first term of the lower bound to obtain the gradient statistics required for updating the hyper-parameters. These changes allow all layers including the gating layer of the network to be updated using back-propagation. The variational inference approach [35] is used to approximate the integration in Eqn. (4). For notation simplicity, we consider the parameters $\boldsymbol{\theta} = \boldsymbol{\theta}_i^{(0)}$ as the gating parameters at the i-th gating layer node. By applying Jensen's inequality, we calculate the evidence lower bound of the cross-entropy criterion, or equivalently the log-likelihood (see Eqn. (5)) of tied HMM state sequence Y given input acoustic feature vector sequence X, with pitch features optionally appended.

$$\begin{aligned} &\log P(\mathbf{Y} \mid \mathbf{X}) = \log \int P(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X}) P_r(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &\geq \underbrace{\int q(\boldsymbol{\theta}) \log P(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{X}) d\boldsymbol{\theta}}_{\mathcal{L}_1} - \underbrace{KL(q(\boldsymbol{\theta}) \| P_r(\boldsymbol{\theta}))}_{\mathcal{L}_2} = \mathcal{L} \end{aligned} \tag{5}$$

where T is the total number of frames in the training data. $P_r(\boldsymbol{\theta})$ denotes gating parameters prior distribution, $q(\boldsymbol{\theta})$ is the variational approximation of gating parameters posterior distribution $p(\boldsymbol{\theta})$. We assume that the variational distribution q and the prior distribution P_r are both Gaussian distributions, following [36], i.e. $P_r(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r^2)$, $q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. $KL(q \| P_r)$ is the Kullback-Leibler (KL) divergence between q and P_r.

The equations for updating the hyper-paramters of the posterior distribution are as in [37]:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu_j} &= \frac{1}{N}\sum_{t,k=1}^{T,N}\frac{\partial \log P(\mathbf{y}\,|\,\mathbf{x},\lambda,\boldsymbol{\epsilon}_k)}{\partial \mu_j} - \frac{T_b}{T}\frac{(\mu_j - \mu_{r,j})}{\sigma_j^2}\\ \frac{\partial \mathcal{L}}{\partial \sigma_j} &= \frac{1}{N}\sum_{t,k=1}^{T,N}\frac{\partial \log P(\mathbf{y}\,|\,\mathbf{x},\lambda,\boldsymbol{\epsilon}_k)}{\partial \sigma_j} - \frac{T_b}{T}\left(\frac{\sigma_j^2 - \sigma_{r,j}^2}{\sigma_j \sigma_{r,j}^2}\right)\end{aligned} \qquad (6)$$

where T_b is the number of frames in a minibatch [38]. Backpropagation method can be applied to the calculation of the two gradient terms $\frac{\partial \log P(\mathbf{y}|\mathbf{x},\lambda,\boldsymbol{\epsilon}_k)}{\partial \mu_j}$ and $\frac{\partial \log P(\mathbf{y}|\mathbf{x},\lambda,\boldsymbol{\epsilon}_k)}{\partial \sigma_j}$ for updating hyper-parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}$. During model evaluation, the mean of the gating parameter $\boldsymbol{\mu}$ of the posterior distribution $p(\boldsymbol{\theta})$ is used as the drawn sample to compute the gating layer outputs.

Table 1: *Performance of baseline DNN systems without (w/o) pitch features v.s. other NN systems with (w/) pitch features on the UASpeech control (healthy) and dysarthric speakers with audio data available. The abbreviations CP and SP in the Dysarthria column represent Cerebral Palsy and Spastic. Mixed means two symptoms were both detected for the speaker.*

ID	WER%				Dysarth
	w/o Pitch	w/ Pitch			
	DNN	DNN	GNN	BGNN	
Control	7.3	7.3	7.0	**6.7**	–
M10	11.2	10.6	9.7	**9.4**	Mixed
M12	65.8	64.3	**63.3**	64.1	Mixed
F04	29.2	**27.0**	28.7	28.2	CP
M11	32.4	**29.2**	29.5	29.7	CP
F02	49.6	46.1	45.8	**43.9**	SP
F03	55.4	51.4	51.9	**50.9**	SP
F05	9.7	7.6	8.1	**7.2**	SP
M01	84.7	74.8	**70.6**	72.1	SP
M04	87.1	**85.8**	87.2	87.4	SP
M05	26.6	21.1	**20.9**	22.4	SP
M06	33.1	32.9	35.8	**32.7**	SP
M07	24.1	23.9	23.0	**22.9**	SP
M08	11.6	10.6	10.4	**10.1**	SP
M09	11.9	10.0	10.0	**10.0**	SP
M14	20.3	16.8	17.0	**16.5**	SP
M16	29.5	27.7	26.1	**25.9**	SP
Avg	33.8	31.4	31.4	**31.0**	–

4. Experiments

4.1. UASpeech Corpus

4.1.1. Experimental Setup

The UASpeech is an isolated word recognition task including 16 dysarthric speakers. All speakers were required to repeat 455 distinct words, comprising of 155 common words and 300 uncommon 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 feed forward DNN based acoustic model without using pitch features is determined as the baseline system, since more advanced forms of neural networks based acoustic models such as time delayed neural networks [39] and long short-term memory recurrent neural networks [40] did not produce lower WER over feed forward DNN on the UASpeech recognition task [23]. All the investigated neural network models on the UASpeech corpus were built in Pytorch [41].

In the experiments, a 9-frame context window was used in both standard feed forward DNN systems and pitch incorporated neural network (NN) systems. Acoustic features are 80-dimension filter banks (FBKs)+Δ features. Pitch features consist of POV, normalized pitch and delta-log-pitch and their first derivatives. The fusion of acoustic and pitch features are shown in section 3. Target tied triphone state labels were produced by speaker dependent GMM-HMM models. Speaker dependent neural network acoustic models have 5 hidden layers with 500 neurons in each hidden layer. This is applicable for all the experimented NN models. For the gated NN and Bayesian gated NN systems incorporating pitch features, an additional hidden layer (or gating layer) was added to perform dynamic selection on pitch features. Layer-wise pretraining was applied to the training process, following a fine-tune for the whole network using SGD optimization with a NEWBOB learning rate scheduler. All the models were trained from scratch. The frame level output probability tables were fed to HDecode [42] to obtain recognition outputs for performance evaluation.

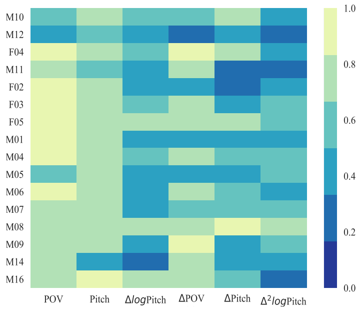

Figure 3: *The pitch feature selection pattern using GNN model on the 16 UASpeech dysarthric speakers*

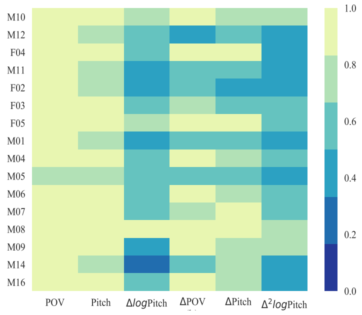

Figure 4: *The pitch feature selection pattern using BGNN model on the 16 UASpeech dysarthric speakers*

4.1.2. Results and Analysis

The performance of baseline DNN systems without pitch features and other NN systems with pitch features is shown in Table 1. We first briefly study the systems' performance on control (healthy) speakers. We can see that Bayesian gated NN works the best. In general, for dysarthric speakers as well, the BGNN systems with pitch features outperform baseline DNN systems with and without pitch features and GNN systems with pitch features on average WER. The BGNN systems provide a 2.8%[1] absolute (8.3% relative) WER reduction over baseline DNN systems without pitch features. Compared to DNN and

[1] The average WER of the DNN ASR systems [23] built by HTK on the 16 speakers is 33.9%. The authors of [23] provide the speaker level WERs. Our baseline DNN ASR systems produced a competitive result.

GNN systems with pitch features, the BGNN systems produce a 0.4% absolute (1.3% relative) WER reduction. Considering the individual speakers' performance, although inconsistent results are observed (see speakers F04, M11 and M04), the selection mechanism on the pitch features still work for most of the speakers, especially for speakers F02 and M01.

To further investigate the selection mechanism of the gating layer on pitch features for both GNN and BGNN systems, we draw heatmaps (see Fig. 3 and Fig. 4) using the activation values extracted from the gating layer's sigmoid function outputs. A polysyllabic word is randomly selected from B2 accross all 16 dysarthric speakers. The frames of the word of each speaker are fed to its corresponding trained GNN and BGNN models to obtain $m * n$ activation value matrices, where m is the frame number of the word and n is the dimension of pitch features. By column averaging, we finally get the average activation value vectors of the selected word for each speaker. From Fig. 3 and Fig. 4, we observe that for BGNN models, the first two dimensions of pitch features are more heavily used than the GNN models. In this case study, using pitch features, the recognition performance of BGNN models is better than DNN and GNN models according to the word recognition result. Therefore it indicates that for this selected word, the BGNN model allows more 0-th order pitch information to flow into the network to improve integration between acoustic and pitch features, and ultimately recognition performance.

Table 2: *Performance of baseline DNN systems without (w/o) pitch features v.s. other NN systems with (w/) pitch features on the control and CUDYS dysarthric speakers with audio data available. The abbreviations SCA and CP in the Dysarthria column represent Spinocerebellar Ataxia and Cerebral palsy.*

ID	WER%			Dysarthria
	w/o Pitch	w/ Pitch		
	DNN	DNN	GNN	
Control	17.8	16.8	**16.5**	–
S0006	3.4	2.2	**2.2**	SCA
S0013	5.3	**4.6**	7.5	SCA
S0015	**83.8**	85.5	84.2	CP
S0019	0.5	**0.5**	1.1	SCA
S0027	2.2	1.6	**1.1**	SCA
S0030	32.7	23.8	**22.9**	SCA
S0031	76.2	74.8	**68.9**	CP
S0034	**63.3**	70.0	66.5	CP
Avg	29.7	29.0	**28.2**	–

4.2. CUDYS Corpus

4.2.1. Experimental Setup

CUDYS [25] corpus is a Cantonese Dysarthric speech corpus collected by the Chinese University of Hong Kong. This corpus contains three tasks, which are words, sentences and paragraphs. In this paper, we only consider the word-level task. In order to improve the recognition performance on disordered speech, an extra large Cantonese normal speech corpus (SpeechOcean, 205.8 hours from 500 healthy subjects) was incorporated and mixed with CUDYS (3.3 hours) to train our ASR system. The training set contains 23 speakers from CUDYS and 441 speakers from SpeechOcean while the test set contains 10 speakers from CUDYS and 29 speakers from SpeechOcean.

Maximum Likelihood Linear Transform (MLLT) estimation [43] was used to train the GMM-HMM system on top of Heteroscedastic Linear Discriminant Analysis (HLDA) [44] transformed Perceptual Linear Prediction (PLP) coefficients. The input 39-dimension PLP features include differential parameters up to the second order. Approximatedly 7000 triphone states were tied according to the phonetic decision tree on the speaker independent GMM-HMM models.

The neural network acoustic model training on the mixed data set is the same as described in Section 4.1.1, except that the speaker independent neural network acoustic models was used consisting of 6 hidden layers with 2000 neurons in each hidden layer. The input acoustic features are 80-dimension filter banks (FBKs)+Δ features. For the GNN and BGNN acoustic models with pitch features, an additional hidden layer (or gating layer) was added to perform dynamic selection on the pitch features.

Table 3: *Performance of baseline DNN systems without (w/o) pitch features v.s. other NN systems with (w/) pitch features on the 10 CUDYS dysarthric speakers with audio data available. The abbreviations SCA and CP in the Dysarthria column represent Spinocerebellar Ataxia and Cerebral palsy.*

Dysarthria	WER%		
	w/o Pitch	w/ Pitch	
	DNN	DNN	GNN
SCA	9.8	**7.1**	7.5
CP	75.2	77.2	**73.9**

4.2.2. Results and Analysis

In this subsection, we compare the performance of baseline DNN systems without pitch features and other NN systems with pitch features on the CUDYS word-level task[2]. The results of control (SpeechOcean) and disordered speakers are shown in Table 2 and Table 3. Experiments in Table 2 show that the GNN systems outperform the DNN systems with (without) pitch by 0.8% (1.5%) absolute WER reduction on average for disordered speakers. For recognition performance of speakers with Cerebral palsy in Table 3, we see a performance degradation in DNN systems with pitch features since the pitch information is inaccurate. Then by dynamically selecting pitch features for CP in GNN system, 1.3% absolute WER reduction was obtained over the DNN systems without pitch features. However, we did not see the same conclusion from the speakers who have SCA.

5. Conclusion

To the best of our knowledge, this paper presents the first work using pitch features for disordered speech recognition. In addition, this is the first attempt to use GNN and BGNN models on exploring robust integration of pitch features to improve recognition performance of disordered speech. Experiments conducted on two disordered recognition tasks, UASpeech and CUDYS, suggest that a selection mechanism on pitch features is required for disordered speech recognition tasks.

6. Acknowledgement

This research is supported by Hong Kong Research Grants Council General Research Fund No. 14200218 and Shun Hing Institute of Advanced Engineering Project No. MMT-p1-19.

[2] In CUDYS setup, the experiments of BGNN system were not conducted, which need further investigation.

7. References

[1] H. Duifhuis, L. F. Willems, and R. Sluyter, "Measurement of pitch in speech: An implementation of goldsteins theory of pitch perception," *J ACOUST SOC AM*, vol. 71, no. 6, pp. 1568–1580, 1982.

[2] C. J. Chen, R. A. Gopinath, M. D. Monkowski, and et al., "New methods in continuous mandarin speech recognition," in *EUROSPEECH*, 1997.

[3] H.-m. Wang, T.-H. Ho, R.-C. Yang, and et al., "Complete recognition of continuous mandarin speech for chinese language with very large vocabulary using limited training data," *IEEE T SPEECH AUDI P*, vol. 5, no. 2, pp. 195–200, 1997.

[4] E. Chang, J. Zhou, S. Di, and et al., "Large vocabulary mandarin speech recognition with different approaches in modeling tones," in *ICSLT*, 2000.

[5] A. Y. P. Ng, L.-W. Chan, and P. Ching, "Automatic recognition of continuous cantonese speech with very large vocabulary," in *EUROSPEECH*, 1997.

[6] T. Lee, W. Lau, Y. W. Wong, and et al., "Using tone information in cantonese continuous speech recognition," *TALIP*, vol. 1, no. 1, pp. 83–102, 2002.

[7] N. T. Vu and T. Schultz, "Vietnamese large vocabulary continuous speech recognition," in *ASRU*, 2009, pp. 333–338.

[8] H. Q. Nguyen, T. D. Le, and et al., "Automatic speech recognition for vietnamese using htk system," in *RIVF*, 2010, pp. 1–4.

[9] S. Kasuriya, S. Kanokphara, N. Thatphilhakkul, and et al., "Context-independent acoustic models for thai speech recognition," in *ISCIT*, vol. 2, 2004, pp. 991–994.

[10] S. Suebvisai, P. Charoenpornsawat, A. Black, and et al., "Thai automatic speech recognition," in *ICASSP*, vol. 1, 2005, pp. I–857.

[11] Y. W. Wong and E. Chang, "The effect of pitch and lexical tone on different mandarin speech recognition tasks," in *EUROSPEECH*, 2001.

[12] T. A. Stephenson, J. Escofet, M. Magimai-Doss, and et al., "Dynamic bayesian network based speech recognition with pitch and energy as auxiliary variables," in *NNSP*, 2002, pp. 637–646.

[13] M. Magimai-Doss, T. A. Stephenson, and H. Bourlard, "Using pitch frequency information in speech recognition," in *EUROSPEECH*, 2003.

[14] M. Płonkowski and P. Urbanovich, "The use of pitch in large-vocabulary continuous speech recognition system," 2016.

[15] A. Cutler and T. Otake, "Pitch accent in spoken-word recognition in japanese," *J ACOUST SOC AM*, vol. 105, no. 3, pp. 1877–1888, 1999.

[16] I. Masuda-Katsuse, "Contribution of pitch-accent information to japanese spoken-word recognition," *Acoustical science and technology*, vol. 27, no. 2, pp. 97–103, 2006.

[17] P. McCaffrey, "The neuroscience on the web series: Cmsd 642 neuropathologies of swallowing and speech," 2013.

[18] K. G. Nicholson, S. Baum, L. L. Cuddy, and et al., "A case of impaired auditory and visual speech prosody perception after right hemisphere damage," *Neurocase*, vol. 8, no. 4, pp. 314–322, 2002.

[19] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *ASSIST TECHNOL*, vol. 22, no. 2, pp. 99–112, 2010.

[20] P. Green, J. Carmichael, A. Hatzis, and et al., "Automatic speech recognition with sparse training data for dysarthric speakers," in *EUROSPEECH*, 2003.

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

[22] 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, pp. 3642–3645.

[23] J. Yu, X. Xie, S. Liu, and et al., "Development of the cuhk dysarthric speech recognition system for the uaspeech corpus," *INTERSPEECH*, pp. 2938–2942, 2018.

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

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

[26] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE T AUDIO SPEECH*, vol. 26, no. 7, pp. 1286–1298, 2018.

[27] 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*.

[28] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J ACOUST SOC AM*, vol. 111, no. 4, pp. 1917–1930, 2002.

[29] D. Talkin, "A robust algorithm for pitch tracking (rapt). speech coding and synthesis, ed. wb kleijn and kk paliwal," 1995.

[30] B. S. Lee, "Noise robust pitch tracking by subband autocorrelation classification," Ph.D. dissertation, Columbia University, 2012.

[31] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *ICASSP*, vol. 1. IEEE, 2002, pp. I–361.

[32] 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.

[33] D. Povey, A. Ghoshal, G. Boulianne, and et al., "The kaldi speech recognition toolkit," Tech. Rep., 2011.

[34] X. Lei, *Modeling lexical tones for Mandarin large vocabulary continuous speech recognition*, 2006, vol. 67, no. 11.

[35] A. Graves, "Practical variational inference for neural networks," in *ADV NEUR IN*, 2011, pp. 2348–2356.

[36] D. Barber and C. M. Bishop, "Ensemble learning in bayesian neural networks," *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, vol. 168, pp. 215–238, 1998.

[37] S. Hu, M. W. Lam, X. Xie, S. Liu, J. Yu, X. Wu, X. Liu, and H. Meng, "Bayesian and gaussian process neural networks for large vocabulary continuous speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6555–6559.

[38] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *NIPS*, 2015, pp. 2575–2583.

[39] A. Waibel, T. Hanazawa, G. Hinton, and et al., "Phoneme recognition using time-delay neural networks," in *Readings in speech recognition*, 1990, pp. 393–404.

[40] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013, pp. 6645–6649.

[41] K. Nikhil, *Introduction to PyTorch*, Berkeley, CA, 2017, pp. 195–208.

[42] S. Young, G. Evermann, M. Gales, and et al., "The htk book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[43] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[44] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition," *Speech communication*, vol. 26, no. 4, pp. 283–297, 1998.