2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 978-1-7281-6994-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCSLP49672.2021.9362063

Multi-Scale Model for Mandarin Tone Recognition

Linkai Peng, Wang Dai, Dengfeng Ke, Jinsong Zhang

School of Information Science, Beijing Language and Culture University

penglinkai96@gmail.com, daiwang_ai@163.com, 8363331@qq.com, jinsong.zhang@blcu.edu.cn

Abstract

Tone plays an important role in tonal languages such as Mandarin and tone classification is an essential component of speech evaluation of Mandarin Chinese. Previous methods for tone classification rarely take into account that different tones possess different scales along both time and frequency axis. Meanwhile, tone contours are subject to many sorts of variation and therefore information from multiple scales can help models to determine the unclear boundary of tones in continuous speech. In this work, we propose a Multi-Scale model which can gather information at multiple resolutions to better capture the characteristics of tone variations effected by complex phonetic and linguistic rules. The experimental results showed that our method achieves competitive results on the Chinese National Hi-Tech Project 863 corpus with TER of 10.5%.

Index Terms: Mandarin tone recognition, spectrogram, sequence processing, deep learning, CTC

1. Introduction

Tone is a pitch pattern being reflected in F0 contours and used for distinguishing ambiguous words and syllables. In Mandarin Chinese, there are four basic tones and a "neutral" one, i.e., Tone 0 (neutral), Tone 1 (flat and high), Tone 2 (rising and middle), Tone 3 (low and dipping) and Tone 4 (falling). The words for "west" (xi1), "learn" (xi2), "lave" (xi3), "thin" (xi4) are shared the same syllable and only can be distinguished by their tones in spoken Mandarin. A good pronunciation of tone can accurately express the meaning of the speaker and facilitate the understanding of the listener. Besides, a well-designed tone recognition model can provide a solid foundation for intonation information processing, prosodic labeling system and computer-aided language learning system. High performance is easy to achieve in the tone recognition of isolated syllables or short words because the speaker produces them more carefully. In [1], Gao *et al.* train a convolution neural network (CNN)[2] to take Mel-spectrogram as the input feature for a signal tone syllable and achieve 99.16% of accuracy. In contrast, continuous speech present difficulties that result in a much lower performance [3]. The F0 contour pattern in continuous speech is often affected by complex phonetic and linguistic rules. First, tone sandhi is a phonological phenomenon that makes changes to tones in certain situations [4]. Second, the tone of one syllable is smoothed to some extent, because human's articulatory organs cannot achieve transient move, which is known as tone co-articulation [5]. In addition, other factors, such as variables emphasis, topic-shift effects and so on, can bring uncertainty to the tone shape [6, 7, 8].

There is a considerable amount of previous work on Mandarin tone recognition. One straightforward approach is splitting the input into segments [9, 10] or frames [11, 12], and then making predictions. [12] take a window of features as input and use CNN and Bidirectional Long Short Term Memory (Bi-LSTM) to extract features on the frequency and temporal dimension, respectively. [10] split the input and predict tone for each segment. These methods require forced alignment to get the frame/segment-level label and another separate model is trained to achieve it. This problem has been eased by connectionist temporal classification (CTC) [13]. [14] use a conceptually simple CNN-CTC model to extract features from cepstrogram and predict tone sequences directly.

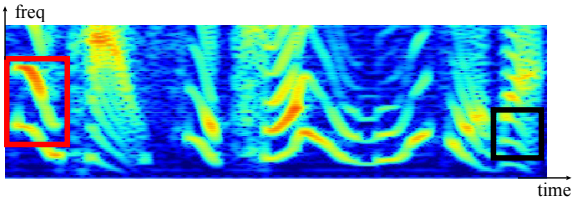

Figure 1: *Mel-spectrogram of F03A161 in dataset 863. Transcript of this part is "... Anger can be understood (... fen4 kai3 shi4 ke3 yi3 li3 jie3 de0)".*

However, most of the previous methods have been applied to *single-scale* features (using single fixed windows width or receptive field), without considering multi-resolution processing. As shown in Figure 1, the tone closed by the red box covers a wider range than the one closed by the black box in the time-frequency domain. Therefore models performed on a single resolution cannot utilize enough information to capture the discriminative properties of tones. Meanwhile, it is difficult for models to determine the boundary between tones without explicit alignment information. In order to help models decide how to assign tone entities, we can provide model temporal information at multiple resolutions to capture temporal relations between neighboring tones.

In this paper, we presented a multi-scale method for recognizing tones. Multi-scale feature representations have proven successful for many vision and speech recognition tasks compared to single-scale methods [15, 16]. Inspired by Inception [17], FCN [18] and U-Net [19], we use CNN to generate multi-scale feature representations and fuse them for later recognition. Specifically, we use a bottom-up structure with a small kernel of convolution to extract multi-scale features and then concatenate the low-scale feature representation and the high-scale feature representation to the standard one. The combined features can capture the properties of tones with different temporal and frequential scales and guide the model to recognize the tone entity and the locations of tones.

2. Proposed Method

In this section, we explain what kind of input we used and why. Then, we describe the model structure (Figure 2) used in this paper.

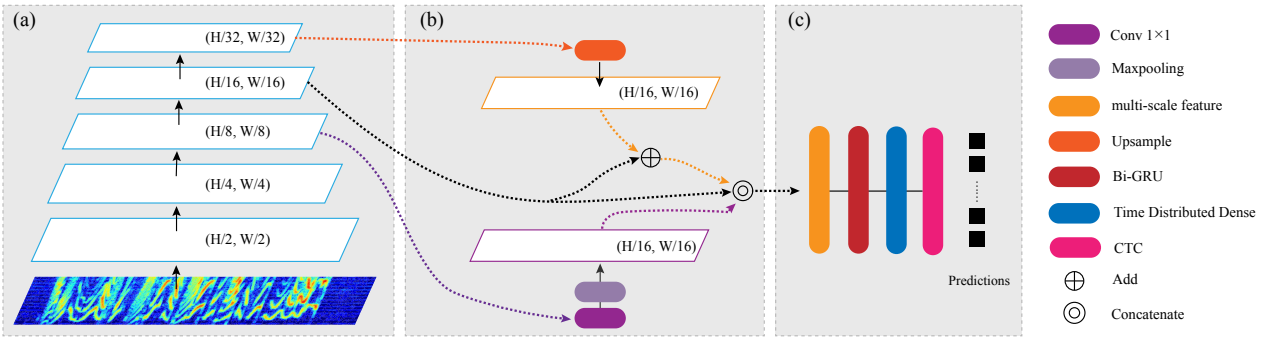

Figure 2: *Model architecture of our method. (a) Bottom-up CNN structure. (b) Merge the deeper branch (high-scale feature representation) and the shallower one (low-scale feature representation) to the standard one. (c) A bi-directional GRU is applied to capture temporal information and a fully-connected layer will predict tone sequences.*

2.1. Input Features

Both Mel-spectrogram and cepstrogram can provide enough information for tone recognition and we can observe tone contour in them. In cepstrogram, the pitch appears as a single peak at the same time, which means that the whole 2D time-frequency representation should be input to the model if we use it as the input. The harmonic structure in Mel-spectrogram makes it possible that we can only select the low-frequency area as input [1], which can gain a remarkable computational saving and improve stability during training.

2.2. Multi-Scale Convolutional Network

The identity of a tone is relatively stable if its overall pitch contour is translated in time or frequency and CNN has been proven to work well on detecting translation-invariant patterns. Thus, a natural way would be using CNN to capture tone patterns from input [1, 12]. Before talking about the multi-scale method, we consider a more common method (single-scale) firstly. In [14], Lugosch *et al.* use CNN to extract single-scale features and Recurrent Neural Network (RNN) to deal with the effect from neighboring entities. The architecture is also popular in handwriting recognition task [20] and license plate detection task [21].

Many methods have proposed multi-scale CNN-based architectures for object detection and recognition. For specific architecture, Inception [17] and Deeplab [22] utilize convolutional layers of varying kernel sizes in parallel to form features at different scales. FCN [18], Yolo [23] and U-net [19] take the outputs from different convolution blocks and fuse them together to extract multi-scale features. For the merging method, Inception [17], U-net [19] and DenseNet [24] combine features by element-wise addition while FPN [25] concatenates multiple features along the channal dimension. These excellent works inspire our multi-scale model for Mandarin tone recognition.

The intention of using multi-scale structures are two folds: *(1)* different tones and their variations possess different scales along the time-frequency domain. In Mandarin Chinese, for example, tone 4 lasts shorter than tone 1 and covers narrower range in the frequency domain generally. *(2)* without force alignment, features at fixed temporal scale are unfriendly to those tones which last longer and shorter than the fixed scale and it is correlative with the insertion and deletion error made by the model. Note that these two ideas are not independant.

As shown in Figure 2, the network of the proposed method consists of three parts: *(a)* we utilize a bottom-up structure to extract spatial features at different scales from the input. It involves five CNN modules which are followed by a max-pooling operation with pooling size and stride of 2×2. While downsampling features map to lower and lower resolutions, each CNN module corresponds to a single unique feature resolution (or scale). As a result, five different scales are gained: $(H/2 \times W/2)$, $(H/4 \times W/4)$, $(H/8 \times W/8)$, $(H/16 \times W/16)$, $(H/32 \times W/32)$, where H and W is the size of input feature. We refer to the feature at scale $(H/16 \times W/16)$ as *standard-resolution* branch (the black dotted line), the deeper one as *low-resolution* (the orange dotted line) and the shallower one as *high-resolution* (the purple dotted line) . *(b)* to merge features from these three different scales, we should resample them back to the *stanard-resolution*. Specifically, we use a 1×1 convolution operation followed by max-pooling operation (the same configuration as previous) to downsample the *high-resolution* features and Nearest-Neighbor interpolation to upsample *low-resolution* features with skip connection. There are two options for merging these features: element-wise addition or concatenation. In this work we chose concatenation as our merging approach and then present the combined features to the recurrent network. To sum up, our multi-scale method is based on the CNN architecture.

2.3. Recurrent Neural Network

RNN-based architectures have strong capability of modeling time sequence which have proven powerful when being combined with the CNN architecture. Part *(c)* of our proposed method uses a bi-directional Gated Recurrent Unit (GRU) layer to deal with the bi-directional effect from neighboring tones in tone recognition. Then a fully-connected layer with 6 outputs (5 Mandarin tones and a CTC 'blank' label) and sigmoid function are applied to get the prediction at each time step. CTC is an approach for sequence labeling that uses Markov assumptions to efficiently solve sequential problems by dynamic programming without additional arduous step of aligning. To get the final tone sequences, we use CTC to encode time step sequences.

3. Experiment

3.1. Dataset

We conduct experiments on Chinese National Hi-Tech Project 863 corpus [26]. The dataset consists of 48373 utterances, contributed by 166 speakers, including 83 females and 83 males, whose total duration is 107 hours. Table 1 lists the distribution of the tones in the corpus. The dataset was divided into training set and test set at the ratio of 9:1. The training set and testing set have not any overlap at the speaker-level and utterance-level.

Table 1: *Numbers of samples of the speech corpus(K)*

Data	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
Total	64.5	235.0	250.9	183.9	384.3

3.2. Experiment Setup

Our spectrogram features are extracted by tool librosa [27], a professional voice processing library. Mel-spectrogram is performed in the whole audio from 20 to 8000Hz, by using a frame length of 2048 sampling points (120ms) and a frameshift of 100 sample points (6ms). The number of samples used each time to compute Fourier Transform (n-fft) is set to 2048 and the number of Mel bins used is set to 512 to achieve a clearer Mel-spectrogram. According to [1], we save the spectrogram as a picture and crop the low frequency part of the picture. In this work, we only reserve the low 128 pixels in the spectrogram with frequency range of [50, 350] Hz and then resize the remain part's height to 64 pixels (at the same time, the width reduce by half). In this way, the training of models can be computationally efficient and stable. For the CNN module, we choose residual block [28] (we refer to it as ResNet for brevity) consisting of two convolution layers with 3×3 kernel size which is more sensitive to fine-grained information. The number of channels are set to [128,256,256,512,512]. [1] use a convolution layer followed by a batch normalization [29] layer and a max-pooling layer as basic feature extract architecture which is taken as the baseline in this paper. Besides, three more CNN modules are explored to compare their performance: i.e. Inception [17] (note that we use it to replace the part-a and the part-b of proposed method because it itself is multi-scaled), Inception-v4-A [30], ResNet pre-act [31], and ensure their parameters are similar. We also apply batch normalization after the convolution layer. The Bi-GRU layer had 512 units in each direction. In the model training stage, models are trained for 50 epochs using Adam optimizer [32] and the learning rate was fixed 0.0001. About the decoding strategy, we use simple greedy decoding to compare with other systems' performance. We use tone error rate (TER) metric for tone recognition evaluation, which is defined as the average Levenshtein distance between predictions and labels.

4. Results and Discussion

4.1. Results

Effect of CNN module. As can be seen in Table 2, multi-scale based models yield better performance over single-scale based models(standard) using the same CNN module, clearly demonstrating the advantage of combining low- and high-resolution features. In addition, the multi-scale model using ResNet as CNN module gains 2.02% improvement in TER but a worse performance than the baseline is observed on Inception-v4-A. It maybe needs some careful adaptions to apply Inception-v4-A. We further evaluated a different approach to model multi-scale by using Inception and it produces better results against the baseline model as expected, confirming that multi-scale learning can improve the performance.

Table 2: *Performance comparison of using various CNN Modules.*

CNN module	Low	Standard	High	TER
Baseline		√		12.53%
ResNet		√		11.55%
ResNet-preact		√		11.38%
Inception-v4-A	√	√	√	16.87%
ResNet	√	√	√	**10.51%**
ResNet-preact	√	√	√	10.79%
Inception	-	-	-	11.35%

Ablation studies. Furthermore, there are more results of ablation studies shown in Table 3. Models combining low- or high-resolution features separately produce better performance than the model using only standard features. The single-scale model combining low-resolution outperforms the one combining high-resolution, which suggests that the high-level network provides more useful information. Model using three scales achieves a much more improvement in TER of 1.04%.

Table 3: *Ablation study of our method.*

CNN module	Low	Standard	High	TER
ResNet		√		11.55%
		√	√	10.95%
	√	√		10.74%
	√	√	√	**10.51%**

Comparison with previous methods. The results are shown in Table 4. Our proposed method significantly outperforms all related works and improves the TER by $\sim$7%. Note that Huang *et al.* [10] use forced alignment to segment the input and predict tone for each segment. However, our method could perform better even without explicit alignment information. This is mainly because the tone shape of a syllable is affected by the F0 contours of its neighboring syllables which requires models to pay attention to a larger scope. This result is consistent with the result shown in ablation studies: the high-level network can provide more useful information.

Table 4: *Comparison performance among other types of networks.*

Method	Model	TER
Chao *et al.* [33]	DBN-SVM	16.97%
Huang *et al.* [10]	Bi-RNN	17.1%
Proposed	MS CNN+BiRNN+CTC	**10.51%**

Effect of feature fusion method. In Table 5, we empirically show that the concatenation approach performs better than element-wise addition, which leaves concatenation as a better choice for our proposed method in tone recognition tasks.

Table 5: *Performance comparison between feature fusion methods.*

Operation	TER
element-wise addition	11.13%
concatenation	10.51%

Error analysis. We use the backtrace of the edit distance between the predictions and the labels to give a more detailed analysis of the types of errors made by the recognizers. From Table 6, it can be seen that our models make fewer deletion and substitution errors than the baseline while the multi-scale model provides an additional gain. The overall number of errors gains substantial reduction while the performance degradation on insertion error could be compromised. Table 7 shows the recognition accuracy for each tone. We can see the proposed method can improve the detection accuracy for all tones by 9%, 0.7%, 1.5% and 0.8%, respectively. Especially for the Tone 0, our proposed method improves the accuracy of it by 9%. This is mainly because that tone 0 in Mandarin is neutral one which does not have a specific pitch pattern and our multi-scale method can capture its variation. In addition, we believe that the multi-scale method can be applied to other features. [14] train a single scale model to extract features from cepstrogram and evaluate their method on AISHELL-1 dataset [34]. Table 8 shows [14] combined with our multi-scale method also can improve the performance.

Table 6: *Breakdowns of errors.*

Method	Insertions	Deletions	Substitutions
Baseline	**79**	1207	7479
Standard	187	752	7211
Multi-Scale	169	**530**	**6696**

Table 7: *Pre-tone accuracy.*

Method	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4
Baseline	68.8%	93.6%	90.1%	81.5%	94.4%
Proposed	**77.2%**	**94.3%**	**91.6%**	**82.0%**	**95.2%**

4.2. Visualization

We use Grad Cam++ [35] to visualize the concerns of features at multiple resolutions. As shown by the Mel-spectrogram (a), an utterance from testing set is taken as an example. From (b) to (d), we present the outputs of our proposed model from three scales, where (b) is the high-resolution representation and (d) is the low-resolution representation. Interestingly, it is obvious that there is a complementary relationship between (c) and (d), as evidenced by the outline closed by black lines, which suggests that the performance improvements are mainly due to the extra information that the standard resolution cannot provide. In contrast, if analysis is performed using only standard resolution, the model might miss some useful information to capture the distinct features of each tone, just like the standard model in Table 2 and Table 3.

Table 8: *Performance on aishell using multi-scale module.*

Method	TER
[14]	11.7%
[14]+Multi-scale	11.4%

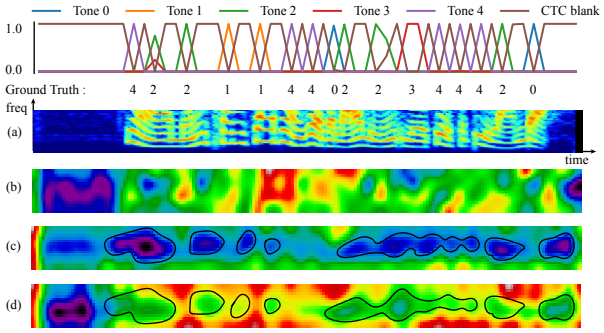

Figure 3: *Visualization of multi-scale features map using Grad-cam++. (a) shows an input sample, i.e. low frequency area of Mel-spectrogram extracted from an utterance (id M72D575). From (b) to (d) are the outputs of high-resolution, standard-resolution and low-resolution. The top one is the output of sigmoid which is consistent with the ground truth. We outline some parts of (c) and (d) with the black line for explicitly indicating the relation between features at different scales in our model.*

4.3. Honorable mentions

We notice that there are some works for tone recognition that worth mentioning. In [36], Lei *et al.* directly extract the tone sequences from the transcript yielded by an ASR system, receiving a TER of 9.3%. Lin *et al.* [37] utilize the toneless syllable sequences to construct an extended recognition network for further tone recognition resulting in a TER of 7.17%. These works rely on additional information such as word label and language model which can correct some tone errors made by the acoustic model. In this work, we pay more attention to the model based on acoustic information only, and what our model can see are the inputs and the tone labels. In addition, Yang *et al.* in [12] take a window of MFCC plus POV as input and train Bi-LSTM network with attention mechanism, achieving a TER of 9.30%. It is not entirely fair to compare our results with theirs because we notice that their scores are calculated on the frame level.

5. Conclusions

This paper has proposed a multi-scale model based on integrating multiple feature representations at different scales for Mandarin tone recognition. Both the low-resolution branch and the high-resolution one can extract more meaningful features and enrich the standard one. Furthermore, when using the proposed method, we achieve state-of-the-art results on the 863 corpus. We hope this work can provide insight for researchers on adopting multi-scale method to tone-related tasks. This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Fundamental Research Funds for the Central Universities (19YBT12, 20YJ040002) and the Research Funds of Beijing Language and Culture University(20YCX136), Jinsong Zhang is the corresponding author.

6. References

[1] Q. Gao, S. tao Sun, and Y. Yang, "Tonenet: A cnn model of tone classification of mandarin chinese," in *INTERSPEECH*, 2019.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[3] J. Zhang, S. Nakamura, and K. Hirose, "Tone nucleus-based multi-level robust acoustic tonal modeling of sentential f0 variations for chinese continuous speech tone recognition," *Speech Communication*, vol. 46, pp. 440–454, 07 2005.

[4] L. Lee, C. Tseng, and C. Hsieh, "Improved tone concatenation rules in a formant-based chinese text-to-speech system," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 3, pp. 287–294, 1993.

[5] Y. Xu, "Contextual tonal variation in mandarin chinese," *ETD Collection for University of Connecticut*, 01 1993.

[6] C. Shih, "The phonetics of the chinese tonal system," *AT&T Bell Labs technical memo*, 1987.

[7] N. Umeda, "F0 declination is situation dependent," 1980.

[8] Y. Xu, "Effects of tone and focus on the formation and alignment of f0contours," *Journal of phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[9] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly accurate mandarin tone classification in the absence of pitch information," 2014.

[10] H. Huang, Y. Hu, and H. Xu, "Mandarin tone modeling using recurrent neural networks," *arXiv preprint arXiv:1711.01946*, 2017.

[11] N. Ryant, J. Yuan, and M. Liberman, "Mandarin tone classification without pitch tracking," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4868–4872.

[12] L. Yang, Y. Xie, and J. Zhang, "Improving mandarin tone recognition using convolutional bidirectional long short-term memory with attention." in *Interspeech*, 2018, pp. 352–356.

[13] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[14] L. Lugosch and V. S. Tomar, "Tone recognition using lifters and ctc," *arXiv preprint arXiv:1807.02465*, 2018.

[15] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] L. Toth, "Multi-resolution spectral input for convolutional neural network-based speech recognition," in *International Conference on Speech Technology & Human-computer Dialogue*, 2017.

[17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[20] K. Dutta, P. Krishnan, M. Mathew, and C. Jawahar, "Improving cnn-rnn hybrid networks for handwriting recognition," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 80–85.

[21] H. Li, P. Wang, and C. Shen, "Toward end-to-end car license plate detection and recognition with deep neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1126–1136, 2018.

[22] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[23] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[26] S. Gao, B. Xu, H. Zhang, B. Zhao, C. Li, and T. Huang, "Update progress of sinohear: advanced mandarin lvcsr system at nlpr," in *Sixth International Conference on Spoken Language Processing*, 2000.

[27] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[30] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," *CoRR*, vol. abs/1603.05027, 2016. [Online]. Available: http://arxiv.org/abs/1603.05027

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[33] H. Chao, C. Song, B.-Y. Lu, and Y.-L. Liu, "Feature extraction based on dbn-svm for tone recognition." *JIPS*, vol. 15, no. 1, pp. 91–99, 2019.

[34] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: an open-source mandarin speech corpus and A speech recognition baseline," *CoRR*, vol. abs/1709.05522, 2017. [Online]. Available: http://arxiv.org/abs/1709.05522

[35] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

[36] X. Lei, M. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for mandarin broadcast news speech recognition," in *Ninth International Conference on Spoken Language Processing*, 2006.

[37] J. Lin, W. Li, Y. Gao, Y. Xie, N. F. Chen, S. M. Siniscalchi, J. Zhang, and C.-H. Lee, "Improving mandarin tone recognition based on dnn by combining acoustic and articulatory features using extended recognition networks," *Journal of Signal Processing Systems*, vol. 90, no. 7, pp. 1077–1087, 2018.