CONVOLUTIONAL NEURAL NETWORK FOR ROBUST PITCH DETERMINATION

Hong Su, Hui Zhang, Xueliang Zhang, Guanglai Gao

Department of Computer Science, Inner Mongolia University, Hohhot, China, 010021

suhong90_imu@qq.com, alzhu.san@163.com, cszxl@imu.edu.cn, csggl@imu.edu.cn

ABSTRACT

Pitch is an important characteristic of speech and is useful for many applications. However, pitch determination in noisy conditions is difficult. In this paper, we propose a supervised learning algorithm to estimate pitch using a convolutional neural network (CNN). Specifically, we use a CNN for pitch candidate selection, and dynamic programming for pitch tracking. Our experimental results show that the proposed method can obtain accurate pitch estimation and they show good generalization ability to new speakers and noisy conditions. We credit the success to the use of CNN, which is suitable for modeling the shift-invariant spectral feature for pitch detection.

Index Terms— Pitch determination, convolutional neural network, dynamic programming

1. INTRODUCTION

Pitch, or fundamental frequency (F0), is an important characteristic of speech. It is useful for many applications, such as speech separation, speech or speaker recognition [1, 2]. Many algorithms are designed to determine pitch in noise-free environments, however, there is still challenge in the present of strong noise [3]. The most prominent difficulty is the corruption of the speech harmonic structure, since most of the existing algorithms rely on a clear harmonic structure [4].

In general, the pitch determination task can be divided into two steps: pitch candidate selection and pitch tracking. Firstly, possible pitches of each frame are selected as candidates. These candidates are selected independently without consideration of other frames. Then a continuous pitch contour is generated by tracking the selected pitch candidates with the temporal continuity constraint. Dynamic programming [5] or hidden Markov models (HMMs) [6] are often adopted for pitch tracking. For pitch candidate selection, signal processing methods, statistical models [7, 8], and the summary autocorrelation function (ACF) [9] are popular. These methods are mostly based on empirical parameters which are not guaranteed to be optimum, or a priori assumption on the noise which limits the application. Inspired by the success of deep learning [10, 11], some researchers select pitch candidates with deep models. Han and Wang investigate the use of a deep neural network (DNN) and recurrent neural network (RNN) for pitch candidate selection [12]. In this study we propose using the convolutional neural network (CNN). To our best knowledge, this is the first study using CNN for robust pitch determination.

We employ the CNN because of its shift-invariant property, which means a pattern can be recognized regardless of its position in the input. This shift-invariant property could be useful in pitch determination. Figure 1 clarifies the idea. There are many parallel lines in the spectrogram indicating harmonics. We can see that the local patterns of harmonic structure are similar along time and frequency axis. Therefore, CNN can model the shift-invariance of local patterns seen in a spectrogram.

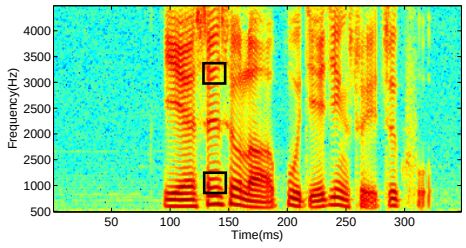

Fig. 1. Harmonic structure in spectrogram. The patterns in small windows are shift-invariant (see the ones in the two black boxes).

In this study, we utilize CNN for pitch detection. The experimental results show that the proposed method can obtain accurate pitch estimation and good generalization ability to new speakers and noisy conditions.

This paper is organized as follows. We list the related works in the next section. Section 3 gives the details of the proposed method. The experimental results are presented in section 4. We conclude the paper in section 5.

2. RELATED WORKS

Numerous robust pitch detection algorithms have been developed. These studies analyze the harmonic structure in the frequency domain, in the time domain or in the time-frequency domain.

The studies in the frequency domain extract the pitch candidates from the spectrogram of the speech by assuming that each peak in the spectrogram is the potential pitch harmonic [13, 14]. Chu and Alan [5] propose a probabilistic framework to model the effect of noise on voiced speech spectra. The PEFAC [15] algorithm combines nonlinear amplitude compression to attenuate narrow-band noise components, with a comb-filter applied in the log-frequency power spectral domain, whose impulse response is chosen to attenuate smoothly varying noise components.

Some other methods consider the periodicity of the speech in the time domain. YIN [16] uses the autocorrelation-based squared difference function and the cumulative mean normalized difference function calculated over voiced speech, with little post-processing to acquire pitch candidates. RAPT [17] and YAPPT [18] generate pitch candidates by extracting local maxima of the normalized cross-correlation function which is calculated over voiced speech.

A variety of temporal approaches extract pitch using the periodicity of individual frequency subbands in the time-frequency domain. In [8], Wu et al. model pitch period statistics in less corrupted channels and then use a HMM for extracting continuous pitch contours. Jin and Wang [6] use cross-correlation to select reliable channels and derive pitch scores from a constituted summary correlogram. Lee and Ellis [19] utilize Wu et al.'s algorithm to extract the ACF features and train a neural network on the principal components of the ACF features for pitch detection. Huang and Lee [7] compute a temporally accumulated peak spectrum to estimate pitch.

3. SYSTEM DESCRIPTION

Similar to other studies, we divide the pitch determination task into pitch candidate selection and pitch tracking. We use CNN to select the pitch candidates, as described in the following subsection. Then we use dynamic programming for pitch tracking, as described in section 3.3.

3.1. Pitch Candidate Selection

Pitch candidate selection chooses the pitch values with high probability. We model this probability distribution with a CNN under a set of observed features. The harmonic structure of spectrum is badly corrupted by the noise. Therefore the feature used in PEFAC [15] is adopted, which shows robustness to noise. We rearrange the original PEFAC features from a logarithmic scale to a linear scale, since they are shift-invariant in linear scale. By this, the harmonic structure is represented by these parallel lines (Fig. 1), and the distance between two adjacent parallel lines indicates the pitch. With a linear scale, this distance is a constant, so that these features in linear scale are shift-invariant. Furthermore, the very exact location of the harmonics is not relevant in our study, we just need to ascertain the pitch bins of the speech.

We set the target pitch frequency from 80 to 415 Hz, a typical range that covers both male and female speech in daily conversations. To simplify the modeling task, we quantize the plausible pitch frequency into "pitch states" by using 24 bins per octave in a logarithmic scale using [12].

$$s = \left\lceil \log_2 \left(\frac{p}{60} \right) \cdot 24 \right\rceil \quad (1)$$

where p is the plausible pitch frequency, and s is the corresponding state. We also incorporate a non-pitched state corresponding to an unvoiced or speech-free frame. Therefore, we have 59 pitch states: 1 state for the non-pitched frame and the other 58 states for the pitched frame.

The output of the CNN is the probability on pitch states, where each pitch state corresponds to a range of pitch values. We convert this probability on pitch states into the probability distribution on real pitch values by adopting a Gaussian mixture model (GMM) framework. Probability density function $p(z)$ for a GMM can be written as:

$$p(z) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(z; \mu_k, \sigma_k^2), \text{where} \sum_{k=1}^{K} \alpha_k = 1, \alpha_k \geq 0 \quad (2)$$

where α_k are the coefficients, $\mathcal{N}(z; \mu_k, \sigma_k^2)$ is a Gaussian distribution, μ_k and σ_k^2 denote the mean and variance of it. K is the number of components.

To select the pitch candidates, we first model each pitch state with a Gaussian distribution whose mean, μ_k, is the center frequency of this state, and the standard deviation, σ_k, is half of its bandwidth. Then we select the top K pitch states using CNN outputs. The corresponding (normalized) CNN outputs are set to the GMM coefficients. We set $K = 3$ according to the development set. Finally, the probability of real pitch values, $p(z)$, is calculated with equation (2). This probability will be utilized for pitch tracking by dynamic programming, which is described in section 3.3. In next subsection, we will describe the CNN used in this study.

3.2. CNN for Pitch State Estimation

In a standard CNN, a convolutional layer is followed by a pooling layer. These layers are stacked up one by one into a deep architecture. And the outputs of the last pooling layer are feed into a fully-connected multi-layer perception (MLP) for classification. The CNN used in this study is illustrated in Fig. 2.

In this study, the speech sampling rate is 8000Hz and window size is 320 of each frame. Since neighboring frames contains useful information for pitch tracking, we also

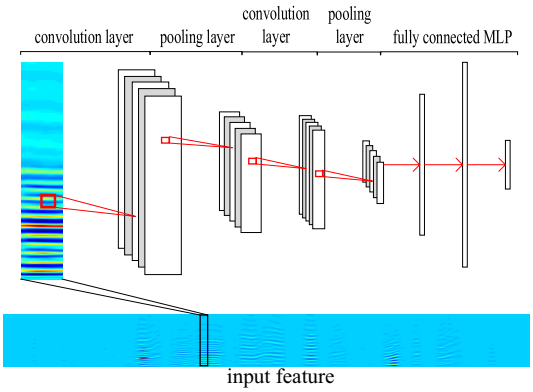

Fig. 2. Structure of the proposed CNN.

incorporate the frontal 7 frames and posteriori 8 frames into the feature vector. Therefor, the feature size of the input is 160×16. The CNN has 2 convolutional layers and 2 pooling layers. The kernel size of the first convolutional layer is 5×5, which is suitable for capturing a discriminative and shift-invariant feature from the inputs. The first convolutional layer contains 10 kernels corresponding to 10 feature maps, and the second convolutional layer contains 20 kernels, whose size is 5×5. After the second convolutional operation, 200 feature maps are generated by 20 kernels fully connected with the 10 feature maps. The pooling layers in the study are mean-pooling, whose size is 2×2. At last, outputs from the last pooling layer are flattened into a vector and feed into a MLP. The MLP contains a sigmoid hidden layer with 500 nodes, and the output function of last layer is softmax. The whole CNN is trained with RMSprop [20] against the cross-entropy loss function. The architecture of CNN is selected by a development set.

3.3. Pitch Tracking

Pitch tracking generates a continuous pitch contour by maximizing the pitch probability under the temporal continuity constraint of speech. The calculation of the pitch probability on each frame was described in section 3.1. The other thing is modeling the temporal continuity constraint, which does not allow the pitch to change by a large amount. As suggested in [8], it can be modeled by a Laplacian distribution:

$$p_t(\Delta) = \frac{1}{2\sigma} exp\left(-\frac{|\Delta - \mu|}{\sigma}\right) \tag{3}$$

where Δ represents the change of pitch period from one frame to the next. We limit $|\Delta| \leq 20$ to further reduce search space. μ is a location parameter and σ>0 is a scale parameter. Here, we set $\mu = 0.4$ and $\sigma = 2.4$ by data analysis.

We generate the final continuous pitch contour by maximizing both the pitch probability and the transfer probability. This process is implemented by a dynamic programming algorithm.

4. EVALUATION

4.1. Dataset

We use the Chinese National Hi-Tech Project 863 corpus for our evaluation. The noises are: n1-machine operation, n2-cocktail party noise, n3-factory noise, n4-siren, n5-speech shaped noise, n6-white noise, n7-bird chirp, n8-cock crow, n9-crowd cheer, n10-babble noise, n11-sound of engine start, n12-alarm, n13-sound in playground, n14-traffic noise, n15-sound of the flowing water and n16-sound of wind, which are selected from [21]. These noises cover a variety of daily noises. To take a further evaluation on the generalization ability, another noise set from the IEEE AASP audio classification challenge [22] is included. This noise set is widely used and includes 10 types of noises, which are denoted as n17-n26.

For our experiments, we setup the training set by randomly selecting a female and a male speaker from corpus and 50 utterances from each. These 100 utterances are mixed with 6 types of noises (n1-n6) at 0 dB. Three test sets are setup: speaker-dependent, speaker-independent and an audio classification challenge (ACC) set. For the speaker-dependent test set, another 40 utterances are selected from the same two speakers (20 new utterance for each) as in the training set. For the speaker-independent and the ACC test sets, another 40 speakers are used and 1 utterance is selected from each speaker. All utterances are mixed with noises at -10, -5, 0 and 5 dB to generate the test set. The speaker-dependent and speaker-independent test sets use the first 16 types of noises (n1-n16). And the ACC test set uses the last 10 types of noises (n17-n26). The noises n7-n26 are not included in the training set. These noises form the unseen noisy conditions.

The ground truth pitch is extracted from the clean utterance using Praat [23].

4.2. Evaluation Metrics

We evaluate the pitch tracking results in terms of two measurements: accuracy rate(AR) on the voiced frames, i.e. a pitch estimate is selected if the deviation of the estimated F0 is within $\pm 5\%$ of the ground truth F0. Another measurement is the voicing decision error (VDE) [19] indicating the percentage of frames are misclassified in terms of pitched and non-pitched as defined:

$$AR = \frac{N_{0.05}}{N_p}, VDE = \frac{N_{p\to n} + N_{n\to p}}{N} \tag{4}$$

where, $N_{0.05}$ denotes the number of frames with the pitch frequency deviation smaller than 5% of the ground truth frequency. $N_{p\to n}$ and $N_{n\to p}$ denote the number of frames misclassified as non-pitched and pitched, respectively. N_p

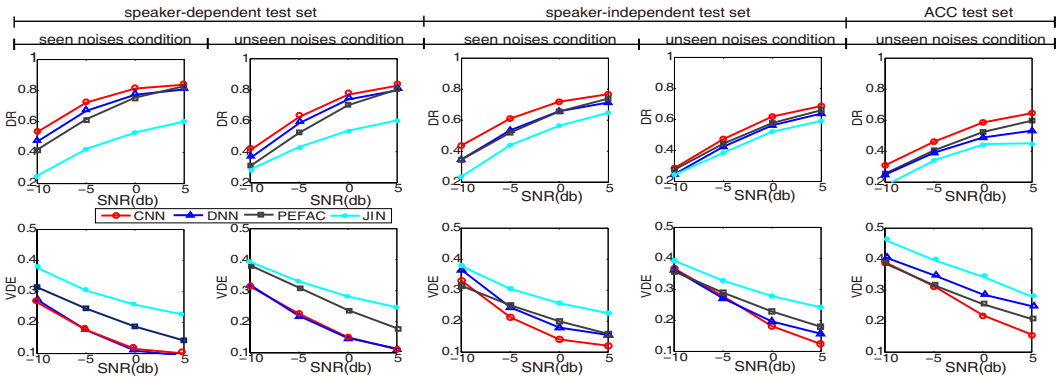

Fig. 3. Performance comparisons. First row: accuracy rate. Second row: voicing decision error. 1st and 2nd column: speaker-dependent test set. 1st column: seen noises condition. 2nd column: unseen noises condition. 3rd and 4th column: speaker-independent test set. 3rd column: seen noises condition. 4th column: unseen noises condition. 5th column: audio classification challenge test set.

and N are the number of pitched frames and total frames in a sentence. High AR and low VDE indicate better pitch estimation.

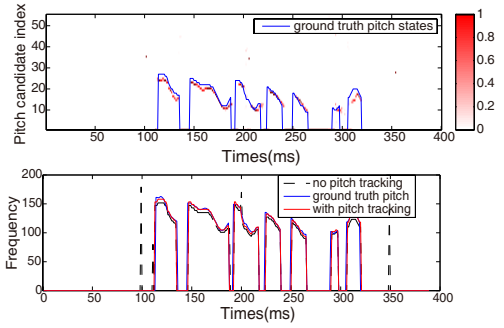

Fig. 4. Example output of the proposed pitch detection method. (a) CNN output. (b) Pitch tracking output. The example mixture is a male utterance which is mixed with machine noise at 0 dB.

4.3. Evaluations

We compare our approaches with three recently proposed pitch determination algorithms: Jin and Wang [6] (denoted as 'Jin'), PEFAC [15] (denoted as 'PEFAC') and a DNN method [12] (denoted as 'DNN'). The code of the first two methods is provided by their authors and we implement the DNN method based on [12].

We first give an example output of our pitch detection method in Fig. 4. Figure 4(a) shows the CNN output, which is the estimated pitch states. We can see that the highest probability of CNN outputs are almost always on the blue line which is the ground truth pitch state. It indicates that the CNN can generate high accuracy pitch state estimates. In Fig. 4(b), we simply select the pitch state with the highest probability, and output its center frequency as the final output. This result shows in black dotted line. We can see some outliers which are caused by errors in pitch state estimations. These outliers break the continuity of the final output. With the pitch tracking, the output gets more continuous, which is shown in the red line. It indicates that the pitch tracking can correct some errors from the pitch state estimation by the CNN.

Then the systematic evaluation results are listed in Fig. 3. It can be clearly seen that the proposed method (the red line in Fig. 3) almost always obtains the highest accuracy rate and lowest voicing decision error. From the left to right, the test condition is less similar to the training condition, where more and more unmatched factors are added. We see that the advantage of the proposed method becomes more obvious. It indicates that the proposed method has a good generalization ability.

5. CONCLUSION

In this study, we employ CNN for robust pitch determination. With shift-invariant characteristics, the CNN models the harmonic structure well. Experimental results show that the proposed method produces promising results and generalizes well to new speakers and noisy conditions.

6. ACKNOWLEDGMENT

This research was supported in part by the China National Nature Science Foundation (No.61365006 and No.61263037), and the Postgraduate Scientific Research Innovation Foundation of Inner Mongolia (No. 1402020201).

7. REFERENCES

[1] Kun Han and Deliang Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[2] Xiaojia Zhao, Yang Shao, and Deliang Wang, "CASA-based robust speaker identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1608–1616, 2012.

[3] Dushyant Sharma and Patrick A Naylor, "Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment," in *Proc European Signal Processing Conf*. Citeseer, 2009, pp. 2514–2518.

[4] Zhengwei Huang, "Multi-pitch estimation," *In Proceedings of the ACM International Conference on Multimedia*, vol. 1, pp. 801–804, 2014.

[5] Wei Chu and Abeer Alwan, "SAFE: a statistical approach to F0 estimation under clean and noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 933–944, 2012.

[6] Zhaozhang Jin and Deliang Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1091–1102, 2011.

[7] Feng Huang and Tan Lee, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 99–109, 2013.

[8] Mingyang Wu, Deliang Wang, and Guy J Brown, "A multipitch tracking algorithm for noisy speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 229–241, 2003.

[9] Lawrence Rabiner, "On the use of autocorrelation analysis for pitch detection," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 25, no. 1, pp. 24–33, 1977.

[10] Dan Ciresan, Ueli Meier, and Jürgen Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

[11] Zhengwei Huang, "Speech emotion recognition using cnn," *In Proceedings of the ACM International Conference on Multimedia*, vol. 1, pp. 801–804, 2014.

[12] Kun Han and Deliang Wang, "Neural networks for supervised pitch tracking in noise," in *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1502–1506.

[13] Dik J. Hermes, "Measurement of pitch by subharmonic summation," *The journal of the acoustical society of America*, vol. 83, no. 1, pp. 257–264, 1988.

[14] Manfred R. Schroeder, "Period histogram and product spectrum: New methods for fundamental-frequency measurement," *The Journal of the Acoustical Society of America*, vol. 43, no. 4, pp. 829–834, 1968.

[15] Sira Gonzalez and Mike Brookes, "PEFAC - A pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[16] Alain De Cheveigné and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[17] David Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[18] Kavita Kasi and Stephen A. Zahorian, "Yet another algorithm for pitch tracking," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–361.

[19] Byung Suk Lee and Daniel P.W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[20] T. Tieleman and G. Hinton, "Lecture 6.5 - rmsprop," COURSERA: Neural Networks for Machine Learning, 2012.

[21] Guoning Hu, "100 nonspeech sounds," http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html, 2006.

[22] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[23] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.