On-line Chord Recognition Using FifthNet with Synchrosqueezing Transform

Rikuto Ito[†] Natsuki Akaishi[†] Kohei Yatabe[‡] Yasuhiro Oikawa[†]

[†]Waseda University, Tokyo, Japan [‡]Tokyo University of Agriculture and Technology, Tokyo, Japan

***Abstract*—Automatic chord recognition is a fundamental and important task in music information processing. FifthNet is a recently proposed chord recognition framework based on a deep neural network (DNN). Its aim is to reduce the computational and memory loads by taking advantage of knowledge on music signals. Since FifthNet achieved the state-of-the-art performance and requires less computational resource compared to the other DNN-based methods, it seems suitable for real-time applications. However, the original FifthNet cannot be directly used in real-time because it applies the spectral reassignment method in its feature extractor. The reassignment method requires some look-ahead frames that results in unavoidable latency. In this paper, we propose to replace the reassignment method of FifthNet with the synchrosqueezing transform to reduce the amount of required look-ahead frames and the computational requirement. Our modification makes FifthNet on-line and removes the obstacle towards real-time execution. In addition, the experimental result shows that the accuracy of chord recognition can be improved by the proposed modification.**

***Index Terms*—Automatic chord recognition, deep neural network (DNN), sparse time-frequency representation, spectral reassignment, synchrosqueezing transform.**

I. Introduction

Chord is one of the most fundamental components of music, and thus automatic recognition of chord is an important task in music information processing. Automatic chord recognition has been realized by applying machine-learning-based classifiers on some acoustic features extracted from music signals [1]–[19]. Classical acoustic features were derived from the short-time Fourier transform (STFT) due to its popularity in acoustic signal processing. However, STFT treats the frequency axis linearly, which does not match with the structure of music signals. Therefore, the constant-Q transform (CQT) is usually applied to handle the frequency axis logarithmically [20] because musical notes have the base-2 logarithmic relationship. Typically, spectral features obtained by CQT are further processed to construct chroma features which compactly represent the energy components related to chord [6]–[14]. Calculated spectral or chroma features are inputted to a classifier to estimate the underlying chord, where recently proposed methods often use a deep neural network (DNN) for the classifier [1]–[14].

FifthNet is a recently proposed DNN-based chord recognition framework [14]. It consists of an input feature extractor and a structured DNN, where both of them are specifically designed for automatic chord recognition. By considering the general structure of music signals, FifthNet achieved the state-of-the-art performance while reducing the number of parameters [14]. Such a small and lightweight DNN can broaden the application of DNN-based chord recognition thanks to the lower computational and memory loads. For example, a sufficiently small DNN can be applied to recognize chord in real-time using a mobile device. Therefore, reducing the computational and memory loads of DNN-based chord recognition systems, while maintaining the recognition accuracy, is one important direction of research.

However, even though the computational and memory loads of FifthNet are small compared to those of the other DNN-based methods, it cannot be directly used in real-time applications. This is because the feature extractor of FifthNet uses the spectral reassignment method [25] that modifies spectral components not only in the frequency direction but also in the time direction. The reassignment method provides sparse time-frequency representation that is known to be effective for automatic chord recognition [12]–[15]. Since it reassigns the positions of time-frequency bins in both time and frequency directions, the feature extractor of FifthNet requires information from both past and future when recognizing chord at the current frame. This property naturally requires some look-ahead frames that would result in undesirable latency on a real-time chord recognition system based on FifthNet. It is desirable for a real-time system to minimize its latency, and therefore reconsidering the feature extractor of FifthNet is necessary for some of its potential applications.

In this paper, we propose to replace the spectral reassignment used in FifthNet with the synchrosqueezing transform to make FifthNet on-line (i.e., allowing step-by-step sequential execution). The synchrosqueezing transform is a variant of the reassignment method [21], where the positions of time-frequency bins are preserved in the time direction. That is, our modification removes the requirement on the look-ahead frames of FifthNet, while maintaining the positive effect of the reassignment method, and reduces the computational and memory loads. In addition, it will be shown in our experiment that the proposed modification of FifthNet can perform similar or better than the original FifthNet. Therefore, our contributions on FifthNet in this paper are three-folds: (1) removing the requirement on the look-ahead frames, (2) reducing the computational and memory loads, and (3) improving the chord recognition accuracy. Although we have not implemented a real-time application, our proposal provides one important and unavoidable step toward real-time execution of FifthNet.

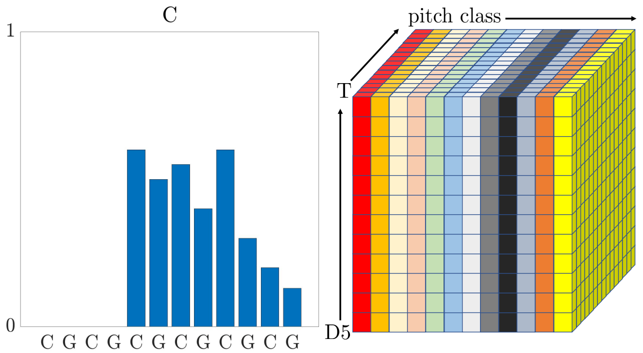

Fig. 1. Example of D5 unit and input feature of FifthNet. An example of the D5 unit for note C at single time frame is shown on the left, and an illustration of the input feature tensor obtained by concatenating D5 units for all time frames and pitch classes is shown on the right.

II. FifthNet and Spectral Reassignment

FifthNet is a chord recognition framework consisted of small DNNs and the specially designed input feature extractor. By incorporating knowledge on the structure of music signals, FifthNet can reduce the computational and memory loads while achieving the state-of-the-art accuracy. Since this paper focuses on the input feature, details related to DNNs are left in the reference. Please see the original paper [14] for the whole details of FifthNet. Also, see Section IV for the network architecture adopted in this paper.

A. Input Feature of FifthNet

The DNNs introduced in the original paper of FifthNet [14] are not special but just a simple combinations of ordinary layers (e.g., convolutional layers, average pooling layers, batch normalization layers, dropout, etc.). The most important aspect that makes FifthNet distinct from the other models is the specially designed input feature obtained through the subfeatures called D5 (fifth data) units.

The D5 unit for each note (and for each time frame) represents the CQT spectral coefficients[1] corresponding to that note, its octaves, and its fifth notes, where the minimum and maximum of the considered notes are A0 and G♯7, respectively. An example for note C is shown on the left side of Fig. 1. Such D5 units for all time frames around the target frame and for all 12 pitch classes are concatenated to construct an input feature tensor for the target frame, as shown on the right side of Fig. 1. This feature is inputted to a DNN that processes and summarizes all the frames, followed by another DNN that provide the recognition result. Once D5 units are calculated from an audio signal, construction of the input feature is straightforward as explained in [14]. In this paper, we focus on the process before that: computation of the pitch feature using CQT with the spectral reassignment.

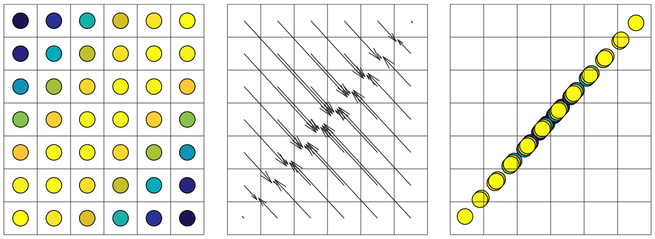

Fig. 2. Illustration of reassignment applied to linear chirp signal. Each bin of the usual spectrogram (left) is moved according to reassignment vectors (center) and placed at the center of gravity of the signal component (right).

B. Spectral Reassignment and Feature Cleaning

FifthNet uses the reassigned CQT [13] for calculating the pitch feature. It is a combination of the spectral reassignment and ordinary CQT, and therefore the reassignment method is explained in this subsection.

The reassignment method [24], [25] is a non-linear post-processing applied to spectrogram[2] for obtaining sparse time-frequency representation. Fig. 2 gives an illustration of the reassignment method. The usual spectrogram in the left figure is reassigned to obtain sparse time-frequency representation in the right figure using the reassignment vectors in the middle figure. The reassignment vectors are calculated using partial derivatives of phase of the spectrogram as follows.

Let STFT of a signal x with a window w be defined as

$$X_w(\omega,t) = \int_{\mathbb{R}} x(\tau)\, w(\tau - t)\, \mathrm{e}^{-\mathrm{i}\omega(\tau-t)}\, \mathrm{d}\tau, \tag{1}$$

where i is the imaginary unit. The polar coordinate representation of a complex number $X_w = A\,\mathrm{e}^{\mathrm{i}\phi}$ provides the amplitude $A(\omega,t) \geq 0$ and phase $\phi(\omega,t) \in [0, 2\pi)$ of an STFT coefficient $X_w(\omega,t) \in \mathbb{C}$. Then, the reassignment method moves $X_w(\omega,t)$ to the new position $(\omega_{\mathrm{new}}, t_{\mathrm{new}})$ given by

$$\omega_{\mathrm{new}}(\omega,t) = \frac{\partial\phi}{\partial t}(\omega,t), \qquad t_{\mathrm{new}}(\omega,t) = t - \frac{\partial\phi}{\partial\omega}(\omega,t), \tag{2}$$

which can perfectly localize sinusoid, Dirac delta, and linear chirp signals. Since the partial derivatives of phase can be computed using two specific windows $w_{\mathrm{D}}(t) = (\mathrm{d}w/\mathrm{d}t)(t)$ and $w_{\mathrm{T}}(t) = t\,w(t)$, Eq. (2) can be rewritten as[3]

$$\omega_{\mathrm{new}}(\omega,t) = \omega - \mathrm{Im}\left[\frac{X_{w_{\mathrm{D}}}(\omega,t)}{X_w(\omega,t)}\right], \tag{5}$$

$$t_{\mathrm{new}}(\omega,t) = \ t + \mathrm{Re}\left[\frac{X_{w_{\mathrm{T}}}(\omega,t)}{X_w(\omega,t)}\right], \tag{6}$$

[1] Note that this is not the total amount of the corresponding component because non-sinusoidal components are discarded based on Eq. (7) before computing the pitch feature and the corresponding D5 unit.

[2] Note that the reassignment method can be applied to some general time-frequency representation other than spectrogram [25], [26].

[3] This representation may seem different from the equations in other papers [13], but such difference is due to the definition of STFT. For example,

$$\begin{aligned}\frac{\partial X_w}{\partial t}(\omega,t) &= -\int_{\mathbb{R}} x(\tau)\,\frac{\mathrm{d}w}{\mathrm{d}t}(\tau-t)\,\mathrm{e}^{-\mathrm{i}\omega(\tau-t)}\,\mathrm{d}\tau \\ &\quad + \mathrm{i}\,\omega\int_{\mathbb{R}} x(\tau)\,w(\tau-t)\,\mathrm{e}^{-\mathrm{i}\omega(\tau-t)}\,\mathrm{d}\tau \qquad (3)\\ &= -X_{w_{\mathrm{D}}}(\omega,t) + \mathrm{i}\,\omega X_w(\omega,t) \qquad (4)\end{aligned}$$

and $(1/X_w)(\partial X/\partial t) = (1/A)(\partial A/\partial t) + \mathrm{i}(\partial\phi/\partial t) = -(X_{w_{\mathrm{D}}}/X_w) + \mathrm{i}\,\omega$ gives Eq. (5) for the STFT defined as in Eq. (1).

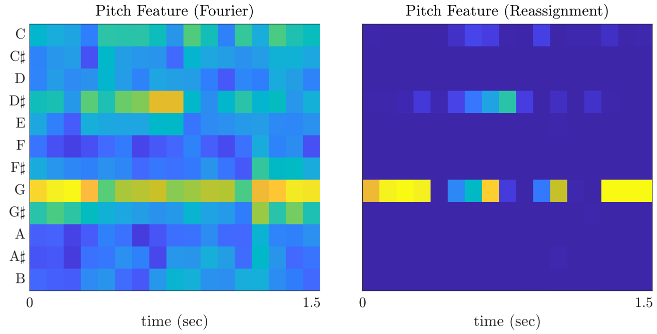

Fig. 3. Example of pitch features with (right) and without (left) reassignment. Note that these are excerpts, and only an octave is shown for visbility.

where $\mathrm{Re}[\cdot]$ and $\mathrm{Im}[\cdot]$ represent the real and imaginary parts, respectively. By computing STFTs using three windows w, w_{D} and w_{T}, the spectral component smeared by the uncertainty principle can be placed back to the original time-frequency position before the smearing.

In addition to the reassignment, FifthNet applies cleaning of the feature by removing non-sinusoidal components. This feature cleaning is based on the following relationship:

$$\frac{\partial^2\phi}{\partial\omega\partial t}(\omega,t) = \begin{cases} 0 & \text{(if signal is sinusoid)} \\ 1 & \text{(if signal is Dirac delta)} \end{cases}, \tag{7}$$

where the mixed partial derivative can be calculated as[4]

$$\frac{\partial^2\phi}{\partial\omega\partial t}(\omega,t) = \mathrm{Re}\left[\frac{X_{w_{\mathrm{TD}}}(\omega,t)}{X_w(\omega,t)} - \frac{X_{w_{\mathrm{T}}}(\omega,t)}{X_w(\omega,t)}\frac{X_{w_{\mathrm{D}}}(\omega,t)}{X_w(\omega,t)}\right] + 1 \tag{8}$$

using the window $w_{\mathrm{TD}}(t) = t\,(\mathrm{d}w/\mathrm{d}t)(t)$. Based on Eq. (7), FifthNet discards non-sinusoidal components determined by the condition $|(\partial^2\phi/\partial\omega\partial t)(\omega,t)| \geq \lambda$, where λ is a threshold and was set to $\lambda = 0.4$ [13].

After applying the reassignment and cleaning, CQT is computed by dividing the frequency axis with equal bins per octave and aggregating the reassigned spectrogram according to the time-frequency positions given by Eqs. (5) and (6) [13]. The use of the reassigned CQT is able to improve the chord recognition performance as shown in [13]. This can be expected by comparing pitch features with and without the reassignment method as shown in Fig. 3.

However, the reassignment method moves time-frequency components not only in the frequency direction but also in the time direction, which naturally requires look-ahead frames. For example, in the case of Fig. 2, the bottom right bin is reassigned to the position that is three frames before, which requires at least three look-ahead frames. The requirement of look-ahead frames directly increases the processing latency which is undesirable for real-time application. Therefore, the amount of the look-ahead frames should be reduced for broaden the application range of FifthNet.

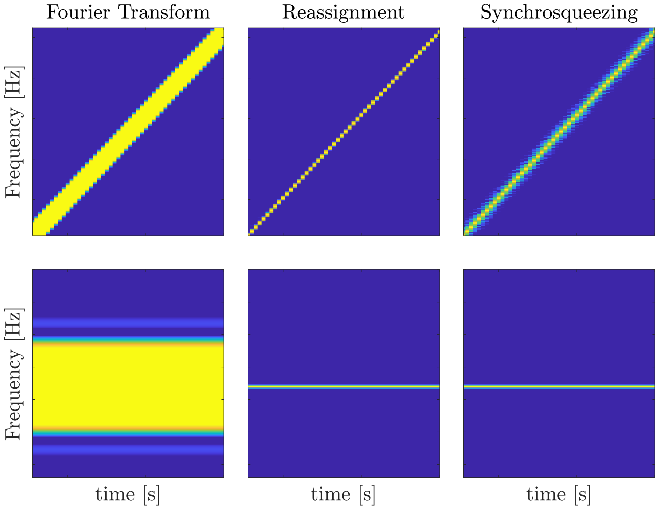

Fig. 4. Example of reassignment of chirp (top) and sinusoid (bottom). The middle column performed both time- and frequency-directional reassignment, while the right column performed only frequency-directional reassignment.

III. Proposed Method

The inconvenience of the reassignment method used in FifthNet comes from the fact that the reassigned temporal position t_{new} moves back and forth relative to the original position t. In this paper, we propose to omit the time-directional reassignment and show that only the frequency-directional reassignment is sufficient for automatic chord recognition.

A. Feature Extraction by Synchrosqueezing Transform

The synchrosqueezing transform [21] is another sparse time-frequency analysis method that is proposed independently from the reassignment method. Although the context and motivation of the proposals were different, the synchrosqueezing transform can be viewed as a variation of the reassignment method where only the frequency position is reassigned[5].

Our proposal is simple: omitting the time-directional reassignment in Eq. (6) from FifthNet. In other words, we perform only the frequency-directional reassignment using Eq. (5), which results in the synchrosqueezing transform. This modification allows us to remove the requirement of look-ahead frames because synchrosqueezing can be calculated using only a current time frame. Our proposal might sound ad hoc, but we have two logical reasons for doing so.

The first reason is that the target of FifthNet is chord. As illustrated in Fig. 2, the time-directional reassignment is necessary for localizing a chirp signal (or any other rapidly varying component). However, it is not necessary for sinusoidal components. As shown in Fig. 4, a sinusoidal signal can be localized by only the frequency-directional reassignment. Since the components related to chord are mainly sinusoidal (i.e., frequency modulation of musical instruments for chord is not so large), omitting the time-directional reassignment should not degrade the chord recognition performance.

[4] These equations may also be different from those in other papers due to the definition of STFT in Eq. (1) (see also Footnote 3).

[5] Even though we do not consider signal reconstruction in this paper, we use the term "synchrosqueezing" merely for convenience.

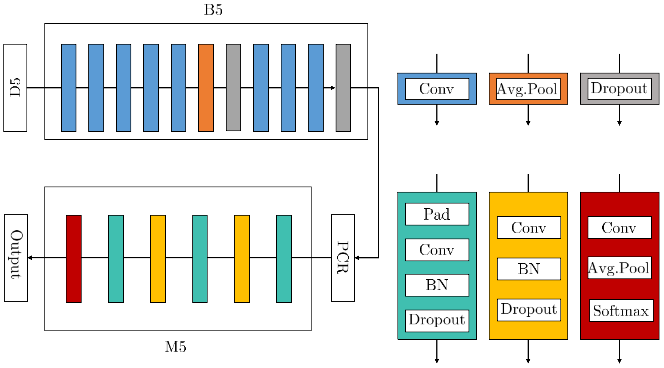

Fig. 5. Network structure of FifthNet. "Conv", "Avg.Pool", "Pad", and "BN" stand for convolution, average pooling, padding, and batch normalization, respectively. The ReLU activation is applied after each convolutional layer. The color of each block on the left side corresponds to those of the details shown on the right side. The number of parameters (e.g., kernel sizes of the convolutional layers) was set to the same as that in [14].

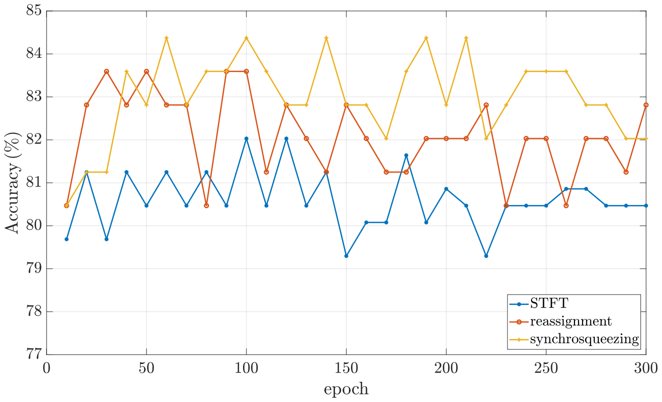

Fig. 6. Median of chord classification accuracy during training. The accuracy was calculated at every 10 epoch for every batch, and median was taken. The original FifthNet and the proposed FifthNet are labeled as "Reassignment" (red) and "Synchrosqueezing" (yellow), respectively. As a baseline, FifthNet without reassignment was also tested and labeled as "STFT" (blue).

The second reason is that FifthNet applies the feature cleaning based on Eq. (7). Using the mixed partial derivative of phase, non-sinusoidal components are removed from pitch features. Therefore, after the feature cleaning, most of the remaining components are sinusoidal. In such case, the time-directional reassignment can be omitted because rapidly varying components are removed by the feature cleaning. The experiment in the next section will show that the proposed omission of the time-directional reassignment does not have negative effect on chord recognition performance.

IV. Experiment

An experiment was performed to see how the proposed omission of the time-directional reassignment affects chord recognition performance.

The network architecture of FifthNet is shown in Fig. 5. The input feature is first processed by a network called B5, and then its output is processed by another network called M5. The intermediate feature is called pitch class representation (PCR). The final output is a 25-dimensional vector, where each dimension corresponds to major or minor chord with 12 root positions, and the last dimension is reserved for "no chord". The number of parameters (e.g., kernel sizes of the convolutional layers) was set to the same as that in [14][6].

The Schubert Winterreise dataset [29] was used for the training. It consists of 48 piano songs for about 4 hours duration and contains audio signals and annotated chord ground truths. We used 40 songs for training, and 8 songs for testing. To prevent overfitting, data augmentation was performed using a pitch shifter, where the pitch of the input signal was randomly shifted for -5 to $+6$ semitones.

The trained task was classification of major and minor triad chords. Since the dataset contains other chords like 7th chords, the annotated ground truths chords were modified as follows: all minor and diminished chords are mapped to the closest minor chords, while all the other chords were mapped to the closest major chords, which is a common way to perform major-minor classification tasks.

The networks were trained for 300 epochs using the Adam optimizer with batch size of 128 and learning rate of 0.001. The accuracy was calculated as follows.

$$\text{Acc}\,[\%] = \frac{\text{Total number of correctly estimated segments}}{\text{Total number of segments}}. \quad (9)$$

First, we tested the accuracy during the training. For every 10 epoch, the accuracy was calculated for each batch, and the median of calculated accuracy values was taken. The result is summarized in Fig. 6. The proposed FifthNet is labeled as "Synchrosqueezing" (yellow) and the conventional FifthNet is labeled as "Reassignment" (red). As a baseline, FifthNet without using the reassignment method was also tested and is labeled as "STFT" (blue). As in the figure, the accuracy of the baseline (blue) reached the ceiling around 20 epochs. Interestingly, the original FifthNet (red) and the proposed FifthNet (yellow) required 10 and 20 more epochs, respectively, to reach the ceiling. Although the accuracy changes for about ±1 or ±2 % as training proceed, the original FifthNet (red) and the proposed FifthNet (yellow) outperformed the baseline in most epochs. In addition, the proposed FifthNet (yellow) outperformed the original FifthNet (red) in many cases. This result indicates that the proposed omission of time-directional reassignment can improve the accuracy of FifthNet even though the computational effort is reduced.

For seeing the detail of the result, we divided the test data into 60 sub groups where each group contains 200 test data. The accuracy values for the 60 trials are summarized in Fig. 7, where the networks at the 100th, 200th, and 300th epoch were used for classification. The median values of the baseline model (blue) were 81.25 %, 80.75 %, and 81.25 % at 100th, 200th, and 300th epoch, respectively. Those of the original FifthNet (red) were 82.25 %, 82.0 %, and 81.0 %, and those of the proposed FifthNet (yellow) were 83.25 %, 82.75 %, and 81.75 %, respectively. In all epochs, the proposed FifthNet (yellow) achieved the highest accuracy in terms of median.

[6] See Table 2 of [14] for B5 and Table 3 of [14] for M5.

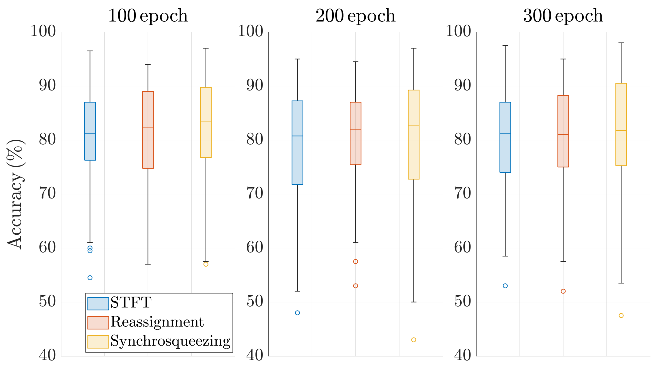

Fig. 7. Box plot of accuracy values calculated using 60 sub groups of test data. The labels and colors are the same as those in Fig. 6. The networks at 100th (left), 200th (center), and 300th (right) epoch were used for classification.

Roughly speaking, the original FifthNet (red) and the proposed FifthNet (yellow) obtained about 1 % and 2 % higher accuracy, respectively, compared to the baseline (blue).

One possible reason for this improvement is that the time-directional reassignment can mix up several components when each component is not isolated. Since music signal is naturally a mixture of sinusoidal components, components in the time-frequency bins are often mixed. Then, the reassignment vectors may point incorrect positions unlike those for a single-component signal. This can result in less accurate pitch features and hence less accurate classification. Our proposal can avoid such mixing of components during reassignment.

Note that our motivation was not to improve the accuracy but to remove the look-ahead frames of FifthNet for making it on-line. Fortunately, our proposal unexpectedly improved the accuracy. From our result, we recommend to omit the time-directional reassignment from FifthNet in all aspects.

V. Conclusion

In this study, we proposed to omit time-directional reassignment used in FifthNet. This simplification allows us to remove the look-ahead frames of FifthNet, which makes it more suitable for real-time execution. As an additional benefit, the proposed modification improved the accuracy of chord recognition, which was confirmed by the experiment. Although our proposal is simple, it provides one important and unavoidable improvement toward real-time execution of FifthNet. In the future works, we will consider real-time implementation of the proposed FifthNet for real-time applications.

References

[1] E. J. Humphrey and J. P. Bello, "Rethinking automatic chord recognition with convolutional neural networks," *Int. Conf. Mach. Learn. Appl.*, pp. 357–362, 2012.

[2] N. Boulanger-Lewandowski, N. Bengio, and P. Vincent. "Audio chord recognition with recurrent neural networks," *Int. Soc. Music Inf. Retr. (ISMIR)*, pp. 355–340, 2013.

[3] S. Sigtia, N. Boulanger-Lewandowski, and S. Dixon, "Audio chord recognition with a hybrid recurrent neural network," *Int. Soc. Music Inf. Retr. (ISMIR)*, pp. 127–133, 2015.

[4] F. Korzeniowski and G. Widmer, "A fully convolutional deep auditory model for musical chord recognition," *IEEE Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pp. 1–6, 2016.

[5] D. Jun-qi and K. Yu-Kwong, "Large vocabulary automatic chord estimation with an even chance training scheme," *Int. Soc. Music Inf. Retr. (ISMIR)*, pp. 531–536, 2017.

[6] B. McFee and J. P. Bello, "Structured training for large-vocabulary chord recognition," *Int. Soc. Music Inf. Retr. (ISMIR)*, pp. 188–194, 2017.

[7] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," *Int. Soc. Music Inf. Retr. (ISMIR)*, pp. 37–43, 2016.

[8] J. Park, K. Choi, S. Jeon, D. Kim, and J. Park, "A bi-directional transformer for musical chord recognition," *Proc. Int. Soc.Music Inf. Retr. (ISMIR)*, pp. 620–627, 2019.

[9] Y. Wu and W. Li, "Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF sequence decoding model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 355–366, Feb. 2019.

[10] L. Rowe and G. Tzanetakis, "Curriculum learning for imbalanced classification in large vocabulary automatic chord recognition," *Int. Soc. Music Inf. Ret. (ISMIR)*, pp. 586–593, 2021.

[11] J. Pauwels, K. O'Hanlon, E. Gómez, MB. Sandler, "20 years of automatic chord recognition from audio," *Int. Soc. Music Inf. Retr. (ISMIR)*, pp. 54–63, 2019.

[12] J. Miller, K. O'Hanlon, and M. B. Sandler, "Improving balance in automatic chord recognition with random forests," *Eur. Signal Process. Conf. (EUSIPCO)*, pp. 244–248, 2022.

[13] K. O'Hanlon and M. B. Sandler, "Comparing CQT and reassignment-based chroma features for template-based automatic chord recognition," *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 860–864, 2019.

[14] K. O'Hanlon and M. B. Sandler. "FifthNet: Structured compact neural networks for automatic chord recognition." *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2671–2682, 2021.

[15] M. Khadkevich and M. Omologo, "Reassigned spectrum-based feature extraction for GMM-based automatic chord recognition," *J. Audio, Speech, Music Process.*, vol. 2013, no. 1, pp. 1–15, 2013.

[16] R. Chen, W. Shen, A. Srinivasamurthy, and P. Chordia, "Chord recognition using duration-explicit hidden Markov models," *Int. Soc. Music Inf. Retr. (ISMIR)*, pp. 445–450, 2012.

[17] L. Kyogu, "Automatic chord recognition from audio using enhanced pitch class profile," *Int. Conf. Math. Comput.*, 2006.

[18] L. Oudre, Y. Grenier and C. Fevotte, "Chord recognition by fitting rescaled chroma vectors to chord templates," *IEEE Trans. Audio, Speech, Langu. Process.*, vol. 19, no. 7, pp. 2222–2233, Sep. 2011.

[19] E. J. Humphrey, T. Cho and J. P. Bello, "Learning a robust Tonnetz-space transform for automatic chord recognition," *IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 453–456, 2012.

[20] J. C. Brown, "Calculation of a constant-Q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 425–434, 1991.

[21] I. Daubechies and S. Maes, "A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models," *Wavelets Med. Biol.*, pp. 527–546, 1996.

[22] E. Gomez, "Tonal description of music audio signals," *Ph.D. dissertation, Univ. Pompeu Fabra*, 2006.

[23] T. Cho and J. P. Bello, "On the relative importance of individual components of chord recognition systems," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 22, pp. 477–492, 2014.

[24] K. Kodera, R. Gendrin, and C. Villedary, "Analysis of time-varying signals with small BT values," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, pp. 64–76, 1978.

[25] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Process.*, vol. 43, no. 5, pp. 1068–1089, May 1995.

[26] N. Holighaus, Z. Průša, P. L. Søndergaard, "Reassignment and synchrosqueezing for general time-frequency filter banks, subsampling and processing," *Signal Process.*, vol. 125, pp. 1–8, Aug. 2016.

[27] T. Oberlin, S. Meignen and V. Perrier, "The Fourier-based synchrosqueezing transform," *IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 315–319, 2014.

[28] M. Müller and S. Ewert, "Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features," *Int. Soc. Music Inf. Retr. (ISMIR)*, 2011.

[29] C. Weiß, F. Zalkow, V. Arifi-Müller, M. Müller, H. V. Koops, A. Volk and H. G. Grohganz, "Schubert Winterreise Dataset: A multimodal scenario for music analysis," *ACM J. Comput. Cult. Herit*, vol. 14, no. 2, pp.1–18, 2021.