Sound Signal Processing Based on Seq2Tree Network

Weicheng Ma[1], *Kai Cao*[2], *Zhaoheng Ni*[3], *Xiuyan Ni*[4], *Sang Chin*[5]

[1]Computer Science Department, Boston University, USA
[2]Computer Science Department, New York University, USA
[3]Graduate Center, City University of New York, USA
[4]Graduate Center, City University of New York, USA
[5]Computer Science Department, Boston University, USA

wcma@csa1.bu.edu, kcao@cs.nyu.edu, zni@gradcenter.cuny.edu, nixiuyan0823@gmail.com,
spchin@cs.bu.edu

Abstract

Most state-of-the-art solutions to sound signal processing tasks such as the speech and noise separation task and the music style classification task are based on Recurrent Neural Network (RNN) architecture or Hidden Markov Model (HMM). Both RNN and HMM assume that the input is chain-structured so that each element in the chain is equally dependent on all its previous units. However in real-life scenes the units alone do not carry much meaning. Only when several units group to be segments will they be semantically informative. This characteristic of sound signals clearly prefers emphasizing dependencies among units in the same segment, which leads to a natural selection of tree-structured models instead of chain-structured ones. In this paper we introduce Seq2Tree network and two models based on Seq2Tree architecture solving 1) speech and noise separation task and 2) music style classification task, respectively. Experiments show that our Seq2Tree-based models outperform the state-of-the-art systems in both tasks, which agrees with our hypothesis that sound signals have potential tree-structured dependencies among their sound elements. Also the experiment results prove the advancement of the Seq2Tree network architecture in sound signal processing tasks.

Index Terms: speech and noise separation, music signal processing, deep learning, seq2tree network

1. Introduction

Traditional models on sound signal processing tasks rely heavily on the global chain-structured dependency among the units in the signal. Such models include models based on Recurrent Neural Network (RNN) or Hidden Markov Model (HMM). It is true that the temporally successive units are related to each other, but the assumption of the chain-structured dependency requires the units to be semantically meaningful. In most sound signal processing tasks this is not fulfilled. For example phonemes in a piece of speech cannot express any meaning without being grouped to be words. This property of units in sound signal processing tasks violates the prerequisite of using chain models so undermines their performances on these tasks.

We call the tasks in which units have to be combined with their neighbors to carry semantic meanings segment processing tasks. The two tasks we choose to tackle in this paper are both segment processing tasks. In the speech and noise separation task clearly the signal has to be modeled by a sequence of words in speech, intertwined with noise fractions from various sources. In the music classification task the characteristics of music styles are uncovered by the patterns of combinatorial use of chords so it is important to abstract segments of related chords into one object in the signal.

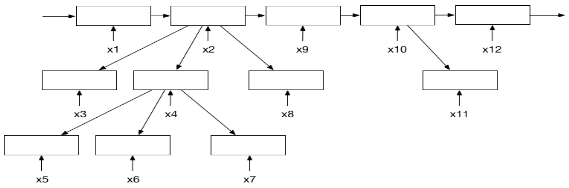

Figure 1: *A three-layered tree structure generated from a Seq2Tree model. Hidden states of lower-level nodes are inherited from parent nodes. The root node of every subtree summarizes the output from its children nodes.*

In segment processing tasks since local relatedness is emphasized, the signals naturally prefer a tree structure where only locally related nodes appear in the same subtree. The start node of a segment appears as the root node of a subtree formed by the following nodes in the segment. Nodes on the same level obeys the chain dependency rule. An example of this tree structure is shown in Figure 1. Existing networks like Tree-LSTM[1] and Multilayer Seq2Seq [2] either require known dependency trees or can only build trees with fixed height at every time step, so neither of them are appropriate in solving segment processing problems. Thus we introduce a neural network structure which builds up such a tree, as is shown in Figure 1, from sequential input. We call this neural network structure Seq2Tree network. The Seq2Tree network should be the standard solution to segment processing tasks since the tree it builds correctly models the signals in segment processing tasks.

Seq2Tree network is able to build up the temporally expanding tree by passing the input data and the previous states to a direction selection gate before any other operation. The direction selection gate chooses the direct parent of the current state, thus decides the level on which the current state should be put. Seq2Tree network also has an update phase at the end of processing each state. These two features enable Seq2Tree network to deal with segment processing tasks better than any existing RNN network structure. We prove the correctness and efficiency of Seq2Tree network over Long Short Term Memory (LSTM), a most commonly used RNN variant, as well as the state-of-the-art models in both tasks using our experimental results generated from the same training/test separations of the corpora with the same set of network parameters. The data for the speech and noise separation task comes from the CHiME challenge [3]. The music data for our evaluation is sampled by ourselves from the Million Song Dataset [4].

Current state-of-the-art system for the speech and noise separation task is based on Bidirectional LSTM network.[5] Bidirectional LSTM makes the same assumption over the input data as the ordinary LSTM network. The difference is that Bidirectional LSTM incorporates future hidden states to the prediction at a previous step.[6] In the music style classification task most of the evaluations have been done over datasets collected by the author, so there exists a lot of randomness. From our selected papers we found that the state-of-the-art results in age- and region-based music classification tasks are both achieved by classifiers based on HMM[7], while the state-of-the-art system in the composer-based music style classification task uses deep feed forward network[8]. In the feed forward neural network model all weights on each layer are shared, which means that the model treats the note at each time step in the same way. This characteristic limits the feed forward neural network's ability to discover segment-level features of the input data. The RNN and HMM based systems also face this problem.

Experiments show that Seq2Tree network generates comparable or better results as LSTM network, while it also beats the state-of-the-art systems on both tasks. This proves that Seq2Tree network is more powerful in modeling noisy speech signals and music signals and could produce the state-of-the-art results in tasks related to these two fields.

2. Seq2Tree Network

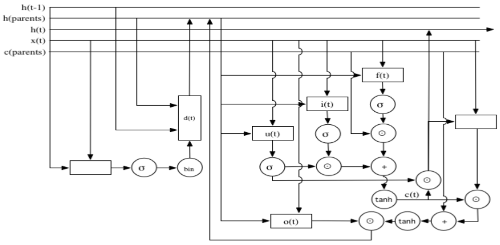

Figure 2: *Seq2Tree network architecture. Left part of the figure is the parent selection gate. $h_{parents}$ and $c_{parents}$ are inherited from the former state. The right half includes the hidden state calculation and parent hidden state updating mechanism.*

Original RNN's are based on the assumption that the input data is chain-structured. However this assumption does not agree with most real-life scenes. For example, in music signal streams, the chords make more sense when they are grouped to be bars or even longer segments. This weakens the performances of RNN's on sound signal processing tasks. In fact according to our experimental results, the model based on Long Short Term Memory(LSTM)[9], an RNN architecture, performs similarly to HMM models.

Thus we come up with a neural network which is better at modeling real-life sound signals by discovering the tree-structured dependency paths among input units. We call this neural network architecture Seq2Tree. In the tree Seq2Tree constructs, children nodes in a subtree inherit the state from their parent node, and the local nodes in the same level of a subtree passes hidden states. This structure efficiently emphasizes the connections among local nodes under the same parent node, which agrees with the features of real-life music signals.

Seq2Tree builds up the tree structure instead of a chain-structured output by adding one branching phase before getting the previous state. The branching operation gives each node more freedom in choosing the parent node to follow. A branching gate d is used to control the parent-selection operation. Transition functions of Seq2Tree network is as follows:

$$\begin{aligned}
d_{kt} &= \theta(\sigma(W^{(d)}x_t + U^{(d)}h_k + b^{(d)})),\\
h_{parent} &= \prod_{d_{kt}=0} d_{kt}h_k + d_{(k-1)t}h_{k-1},\\
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{parent} + b^{(i)}),\\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{parent} + b^{(f)}),\\
o_t &= \sigma(W^{(o)}x_t + U^{(o)}h_{parent} + b^{(o)}),\\
u_t &= tanh(W^{(u)}x_t + U^{(u)}h_{parent} + b^{(u)}),\\
c_t &= i_t \odot u_t + f_t \odot c_{t-1},\\
h_t &= u_t \odot tanh(c_t),\\
\Delta f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_t + b^{(f)}),\\
\Delta c_t &= \Delta f_t \odot c_t,\\
c_{lt} &= c_{lt} + \Delta c_t - \sum_{i=0}^{l-1} u_i \Delta c_t,\\
h_{lt} &= o_{lt} \odot tanh(c_{lt}).
\end{aligned}$$

where θ stands for a binary thresholding function, kt indexes all the ancestor nodes including the previous state and lt is in the range of the current node's ancestors. h_{parent} is the selected parent node from the set $h_{parents} \cup h_{t-1}$. i, f, o, u, c are the LSTM gates and the memory cell, respectively, and the W, U, b matrices are the weights. The new node keeps climbing up its ancestor path till the direction gate d_{kt} becomes non-negative. Node p_{k-1} is then selected to be the parent node of the current node when the algorithm stops, or when it reaches node on the highest level of the tree.

3. Tasks

3.1. Speech and Noise Separation

As is defined in the Second CHiME Challenge[3, 10], the goal of speech and noise separation task is to predict a time-frequency mask, when given a piece of noisy speech audio, that minimizes the energy of noise when applied to the original speech signal.

The current state-of-the-art system for this task uses Bidirectional LSTM network structure[5]. Compared to the original LSTM network, Bidirectional LSTM incorporates future information to the prediction at each time step. This helps the Bidirectional LSTM model bound the range of noise signals.

However different types of noises follow different sets of patterns. Bidirectional LSTM is not able to detect the noise type when multiple noise signals overlap with each other, which harms their prediction performance in more complex situations.

Based on the structural features of the speech signals, we found that Seq2Tree network is more suitable for the speech and noise separation task. The benefit of applying Seq2Tree architecture is that in the generated tree structure, overlapped noise signals can be separated into nested layers, which helps preserve the pattern of noise from a single source.

3.2. Music Style Classification

Compared to speech recognition tasks, input signals in music-related tasks are often more regularized and easier to predict.

To emphasize this feature and to be consistent with our baseline systems, we treat the input music signals as sequences of notes. The notes are represented using their pitch class.

Previous works in this task mainly focus on extracting features from fractions of the input sequence. This leads to a trend of combining HMM with a classification algorithm such as Naive Bayes Classifier or AdaBoost, as is used by the current state-of-the-art system. Based on the chain-structured input assumption we also implemented a LSTM classifier as our baseline neural network model. Experiments show that the LSTM classifier performs comparably well with the state-of-the-art AdaBoost classification system.

Again we noticed that the melody features in a music piece are decided by not single notes, but segments of them. This motivated us to apply Seq2Tree network on music style classification task, benefitting from the ability of Seq2Tree architecture to group notes in adjacent time steps.

4. Models

4.1. Speech and Noise Separation

Our solution to the speech and noise separation task is to build up the tree structure modeling a real-life speech scene first, then make a prediction based on the hidden state at each time step by training a softmax regressor:

$$mask_t = softmax(U^{(R)}h_t + b^{(C)})$$

where $U^{(R)}$ is the regression matrix which is trained on the CHiME data[10, 11].

The input sound signal is preprocessed into one feature vector containing the energy in all the frequency bins at each time step through Fourier Transform. As is suggested by Weninger et al.[12], we apply two-stage training with the following loss functions:

$$J_1(t) = -\frac{1}{C}\sum_{i=1}^{c}(mask_{ti} - label_{ti})^2$$

$$J_2(t) = -\frac{1}{C}\sum_{i=1}^{c}(\|x_t\| \cdot (mask_{ti} - label_{ti}))^2$$

where c denotes the number of frequency bins, $mask_{ti}$ is the predicted vector mask for bin i at time t, and $label_{ti}$ represents the gold-standard mask for bin i at time step t.

4.2. Music Style Classification

We apply a trained softmax classifier on top of the root node of the built tree structure to solve the music style classification task. The root node is an additional highest level node added to the tree generated by Seq2Tree network. Its hidden state is updated by every root state of a non-empty subtree. The final output is a probability distribution over all possible classes:

$$p(y|h_{root}) = softmax(U^{(C)}h_{root} + b^{(C)}),$$
$$\hat{y} = argmax_y\, p(y|h_{root})$$

where $U^{(c)}$ is the classification matrix, $b^{(c)}$ represents the bias, h_{root} is the hidden state at the root node, and $\hat{y}$ is the predicted label of the input.

The loss function we use in this model is the cross entropy loss of the predicted label $\hat{y}$:

$$J(\theta) = -\frac{1}{C}\sum_{i=1}^{C} p(\hat{y}) \cdot log\, p_\theta(\hat{y}|h_{root})$$

where C is the number of possible classes, $p(\hat{y})$ is the probability that the input music actually belongs to the class $\hat{y}$, and $p_\theta(\hat{y}|h_{root})$ represents the predicted probability that the music falls in the class $\hat{y}$ with the parameter set θ.

5. Experiments and Discussions

5.1. General Settings

In both models we set the dimension of hidden states to be 1024. For the speech and noise separation experiment we set the input shape to be 50 time steps with 513 frequency bins at each time step. In the music style classification task we slice the input note sequence into batches of size 300, since the input sequences are of variable length.

5.2. Speech and Noise Separation Experiment

5.2.1. Experiment Setup

In this task, the input is audio waveform, and our goal is to predict a mask which could minimize the energy of noise when the mask is applied to the input audio. The audio files are preprocessed using Short Time Fourier Transform (STFT), then the energy values are filtered into 513 frequency bins. The audio data is from the CHiMe dataset[10, 11], and we use a 80%/20% split for training and test sets.

For evaluation we implemented the prediction model described in Section 4.1 with Seq2Tree network. In the network, a prediction is made at each time step, but postorder in the tree generated by the model. The reason is that the parent state is ready to output only when the segment is entirely processed.

The overall results for both our model and the baseline system are evaluated in terms of Mean Squared Error (MSE), while we also compare their performances on single audio files using Overall Perceptual Score (OPS) in this task[13, 14]. The score is calculated by comparing the energy distribution at each time interval to that of the gold standard noise-free audio files.

5.2.2. Results and Analysis

The MSE for both our Seq2Tree-based model and the baseline model are listed in Table 1. The results are recorded after 10-fold cross validation over the sampled CHiME data.

Model	**MSE**
BLSTM	0.0445331
Seq2Tree	0.0204997

Table 1: *Speech and noise separation evaluation results. Average performance of each model after ten-fold validation.*

Clearly, in terms of average performance our model outperforms our baseline Bidirectional LSTM model. This agrees with our hypothesis that the speech and noise separation task is a segment processing task instead of sequence processing task, and that in this task, local dependencies inside each segment are more important than global dependency paths.

Though our model using Seq2Tree architecture performs better than the Bidirectional LSTM model in general, our model suffers from low performance in the worst case. The best and worst results of both systems can be found in Table 2. This should have been caused by incorrect branching operations when building the tree. Further tuning for the threshold of the branching gate is needed.

Model	OPS(dB)
BLSTM (Worst Case)	25.01
BLSTM (Best Case)	40.96
Seq2Tree (Worst Case)	26.17
Seq2Tree (Best Case)	62.09

Table 2: *Speech and noise separation evaluation results. Extreme case performances of each model.*

5.3. Music Style Classification Experiment

5.3.1. Experiment Setup

For the music style classification task we designed three independent experiments in order to compare the performance of our model with the baseline systems. The three experiments are composer-based music style classification, age-based music style classification, and country-based music style classification, respectively. Each dataset for the three tasks contain 800 music pieces. Data for every experiment is subject to a 80%/20% training/test separation.

We use the same model for the three experiments. The model is implemented according to the specifications in Section 4.2. The input is a sequence of notes represented by their pitch class, and the prediction is made at the root node of the tree built by our Seq2Tree-based model, after all updates are done. The results are recorded in terms of classification precision.

5.3.2. Results and Analysis

The country-based music style classification results are reflected in Table 3. The overall precision is calculated over all six classes. We also give the average precision of our model on binary classification tasks over every of the six countries, so as to make our results comparable to those of our baseline systems[15, 8, 7, 16]. Regarding the classification results, clearly our model performs better than the baseline systems, with an improvement of nearly 30 in terms of binary classification precision. Moreover, to prove the advancement of our model over neural network approaches, we evaluated the performance of one LSTM-based model on the same task and corpus. The results show that in both multi-class and binary classification tasks our model outperforms the LSTM model.

Model	Precision
HMM (Binary)	77%
HMM (3-class)	63%
LSTM (Binary)	83.33%
Seq2Tree (Binary)	98.26%
LSTM	76.67%
Seq2Tree	94.74%

Table 3: *Country-based music style classification results. The binary classification result is the average precision for every class.*

The encouraging results produced by our Seq2Tree model which beat all the chain-structured models with the same hyper parameters support our hypothesis that music signals are tree-structured. Also they help prove the ability of the Seq2Tree network to preserve tree-structured dependencies.

For the composer-based music style classification problem we compare the performance of our model with the state-of-the-art results by the system of Giuseppe Buzzanca [8] using deep feed forward network. They did only binary classification experiments so we added two experiments using our LSTM model on both the ten-class classification task and the binary classification task, as our baseline. Meanwhile, we are the first to publish evaluation results in an age-based music style classification task so we compare our results only with the LSTM system. Results for the two experiments are shown in Table 4 and Table 5, respectively. In both tasks our Seq2Tree-based model outperforms the state-of-the-art results and our baseline LSTM system, which is very consistent with the experimental results in the country-based music style classification task.

Model	Precision
Feed Forward Network (Binary)	97.09%
LSTM (Binary)	90.08%
Seq2Tree (Binary)	99.46%
LSTM	81.27%
Seq2Tree	96.76%

Table 4: *Composer-based music style classification results.*

Model	Precision
LSTM (Binary)	80.13%
Seq2Tree (Binary)	98.75%
LSTM	74.33%
Seq2Tree	96.86%

Table 5: *Age-based music style classification results.*

6. Conclusion

In this paper we introduce Seq2Tree network to sound signal processing tasks. Seq2Tree network is able to learn tree-styled dependency structure from sequential input without the help of syntactic rules. Compared to other tree-construction neural networks such as multilayer Seq2Seq network, Seq2Tree architecture allows the tree it builds to be arbitrarily deep at each time step, which agrees with real-life sound signals more. Thus we think that Seq2Tree network is a better choice in sound signal processing tasks. To show this we designed two experiments and compared the performances of a Seq2Tree model with the state-of-the-art system and an LSTM baseline system in each task. The tasks are 1) a speech and noise separation task and 2) a music style classification task. Experimental results show that in both tasks the Seq2Tree-based models beat the reported state-of-the-art results and the performance of the LSTM baseline system. This clearly supports our claim that Seq2Tree network is more proper for sound signal processing tasks than the popular sequential systems. Besides sound signal processing tasks, we believe that Seq2Tree network should work well on any segment processing task. Further experiments are needed to prove this.

7. References

[1] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[2] O. Vinyals, Ł. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a foreign language," in *Advances in Neural Information Processing Systems*, 2015, pp. 2773–2781.

[3] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 162–167.

[4] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 708–712.

[6] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[7] W. Chai and B. Vercoe, "Folk music classification using hidden markov models," in *Proceedings of International Conference on Artificial Intelligence*, vol. 6, no. 6.4. sn, 2001.

[8] G. Buzzanca, "A supervised learning approach to musical style recognition," in *Music and Artificial Intelligence. Additional Proceedings of the Second International Conference, ICMAI*, vol. 2002, 2002, p. 167.

[9] W. Zaremba and I. Sutskever, "Learning to execute," *arXiv preprint arXiv:1410.4615*, 2014.

[10] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'chime' speech separation and recognition challenge: Datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 126–130.

[11] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The pascal chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013.

[12] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*. IEEE, 2014, pp. 577–581.

[13] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[14] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 430–437.

[15] G. Velarde, T. Weyde, C. C. Chacón, D. Meredith, and M. Grachten, "Composer recognition based on 2d-filtered piano-rolls," *brain*, vol. 4, no. 10, p. 6, 2016.

[16] J. Vlegels and J. Lievens, "Music classification, genres, and taste patterns: A ground-up network analysis on the clustering of artist preferences," *Poetics*, 2016.