Interspeech 2018
2-6 September 2018, Hyderabad

Transfer Learning based Progressive Neural Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis

Ruibo Fu [1, 2], Jianhua Tao [1,2,3] , Yibin Zheng [1, 2] ,Zhengqi Wen [1]

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{ruibo.fu,jhtao,yibin.zheng,zqwen}@nlpr.ia.ac.cn

Abstract

The fundamental frequency and the spectrum parameters of the speech are correlated thus one of their learned mapping from the linguistic features can be leveraged to help determine the other. The conventional methods treated all the acoustic features as one stream for acoustic modeling. And the multi-task learning methods were applied to acoustic modeling with several targets in a global cost function. To improve the accuracy of the acoustic model, the progressive deep neural networks (PDNN) is applied for acoustic modeling in statistical parametric speech synthesis (SPSS) in our method. Each type of the acoustic features is modeled in different sub-networks with its own cost function and the knowledge transfers through lateral connections. Each sub-network in the PDNN can be trained step by step to reach its own optimum. Experiments are conducted to compare the proposed PDNN-based SPSS system with the standard DNN methods. The multi-task learning (MTL) method is also applied to the structure of PDNN and DNN as the contrast experiment of the transfer learning. The computational complexity, prediction sequences and quantity of hierarchies of the PDNN are investigated. Both objective and subjective experimental results demonstrate the effectiveness of the proposed technique.

Index Terms: speech synthesis, progressive neural networks, acoustic modeling, transfer learning

1. Introduction

Artificial speech synthesis, which is known as text-to-speech (TTS), has two domains. One domain is unit-selection and concatenative speech synthesis, which need large-scale speech corpora to achieve good quality [1-6]. Another domain is the statistical parametric speech synthesis (SPSS). Learning the mapping from the abstract linguistic features to acoustic parameters is one of its central tasks [7].

Recently, end-to-end SPSS methods have been proposed and achieved quite good results [8-11]. However, the end-to-end methods still need more delicate works in computational efficiency and robustness. On the contrary，the conventional SPSS methods, which have four pipelines including text analysis, prosody prediction, acoustic model and vocoder, are more robust and have small footprints. The application area of the SPSS is wide now.

The accuracy of the acoustic model affects the quality of synthetic speech. Early SPSS methods normally combined all the acoustic features (line spectral pair (LSP), fundamental frequency (F0), voiced/unvoiced (U/V), etc) to one features vector. For instance, SPSS based on the hidden Markov models (HMMs) used the decision tree to cluster the phone state [12]. Then, the deep belief networks (DBN) [13, 14] and deep neural networks (DNN) [15] were applied for the acoustic model. To enlarge the receptive field of speech, recurrent neural networks (RNN) [16] and its variant bidirectional long short term memory (BLSTM) [17] were applied to build the sequence mapping between linguistic features and acoustic features. However, the models of the above methods tended to learn the high-dimension spectral features and to ignore the low-dimension features like F0. The correlation between spectral features and F0 features was ignored in the training process.

To handle these problems, multi-task learning (MTL) methods [18, 19] were adopted. A combined weighted cost function was defined to balance the errors from the generation of F0 and spectrum in the training process. A structured output layer (SOL) [20] was applied to generate the F0 and spectral parameters separately. The above MTL methods separated the prediction of the F0 and the spectrum by algorithm and network structure. However, we still need to adjust the weights of each sub-target cost manually. And we have no controls on the parameters distribution ratio of the network for each sub-task. The networks could not fully concentrate on the sub-task, because the train criterion is to minimize the combined global weighted cost function.

In this paper, we investigate into modeling the correlation between the F0 and the spectrum by the transfer learning methods. One hypothesis that we make is that one of learned mappings from the linguistic features to the acoustic features can be transferred to another mapping learning process, which could improve the accuracy of the predicted acoustic features. However, the conventional pre-training and fine-tuning (PT/FT) method had very limited contribution because the sample size of the F0 and spectrum were equal and the model trained previously would be forgotten by epochs of training. On the contrary, we apply the transfer learning based progressive deep neural network (PDNN) with a sharing weights strategy to model the acoustic features (U/V, LSP, F0) step by step, which is a more thorough way to separate the predictions of F0 and spectrum. We investigate transfer learning between three types of acoustic features: U/V, F0, and LSP. In all cases, we investigate five methods: (1) DNN (2) DNN with MTL method (3) PDNN (4) PDNN with MTL method (5) PDNN with PT/FT. Furthermore, we also investigate prediction sequences and quantity of hierarchies of the PDNN

10.21437/Interspeech.2018-1265

on performance of the SPSS system. The computational complexity of the PDNN method is also discussed.

The rest of the paper is organized as follows: Section 2 proposes the PDNN framework for SPSS. The MTL methods are also included. Section 3 presents the experiments. The conclusions and future work are discussed in Section 4.

2. Methods

The production of speech is the cooperation of vocal folds and articulators [21]. In SPSS, the F0 parameters represent the state of vocal folds while the spectral parameters relate to articulators [22, 23]. The PDNN framework for SPSS is designed to model the correlation of F0 and spectrum by transfer learning method. Another MTL method is also introduced in this section.

2.1. Progressive neural networks

Progressive neural networks (ProgNets) [24] were first proposed by Google for the reinforcement learning tasks. ProgNets trained a new task by freezing the previous trained tasks. Compared with the conventional transfer learning methods that use the learned parameters as initial parameters, the ProgNets use the following strategies:

- Firstly, all the parameters of the old model are frozen when the new task begins.
- Secondly, the new model is initialized randomly.
- Thirdly, lateral connections are built between the new model and the frozen old model.
- Fourthly, the parameters of the new model is learned through backpropagation.

2.2. PDNN Framework for SPSS

The PDNN framework with 3 columns (3 colors) for SPSS is shown in Figure 1. The first task starts with a single column (green in the Figure 1): A deep neural network having 5 layers with hidden activations $h_i^{(1)} \in R^{n_i}$, with n_i the number of units at layer $i \leq 5$,and parameters $\Theta^{(1)}$ trained to convergence.

When switching to a second task, the parameters $\Theta^{(1)}$ are "frozen" and a new column (yellow in the Figure 1) with parameters $\Theta^{(2)}$ is instantiated with random initialization, where layer $h_i^{(2)}$ receives input from both $h_{i-1}^{(2)}$ and $h_{i-1}^{(1)}$ via lateral connections. This generalizes to K tasks as follows:

$$h_i^{(k)} = f\left(W_i^{(k)} h_{i-1}^{(k)} + \sum_{j<k} U_i^{(k:j)} h_{i-1}^{(j)} + b_i^{(k)}\right) \quad (1)$$

where $W_i^{(k)} \in R^{n_i \times n_{i-1}}$ is the weight matrix of layer i of column k, $U_i^{(k:j)} \in R^{n_i * n_j}$ are the lateral connections from layer $i-1$ of column j, to layer i of column k, $b_i^{(k)}$ are the biases and h_0 is the network input. f is the activation function.

In the construction of PDNN, it is important to carefully select a method for combining representations across network and to identify where these representations will be combined. Adaptation layers (as in the Figure 1) can be included to transform from one task to another. However, due to the limit of the computational complexity in the runtime SPSS system, a sharing weights strategy of the lateral connections $U_i^{(k:j)}$ is adopted in our framework. All the rows of the matrix $U_i^{(k:j)}$ are sharing the same row vector. Except the first layer, each node of in the same layer of the column k would receive the same bias. The real number of parameters in the matrix $U_i^{(k:j)}$is n_{i-1} instead of $n_i \times n_{i-1}$.

2.3. PDNN structure with Multi-task learning methods

MTL is also a way to train the PDNN model for the three different but related tasks. The above proposed PDNN structure can be applied by multi-task learning method if we change the training procedure and cost function. Squared loss is used as the cost function for each sub-task. All the three models are trained together with the global combined cost function:

$$F_g = \alpha F_s + (1-\alpha) F_p \quad (2)$$

where F_s and F_p are the error costs generated by the main task (spectrum) and the auxiliary task (F0) computed as mean squared errors (MSE). The coefficient α is a parameter that need manually adjusted. While in the training of PDNN, F_p and F_s are cost functions each task.

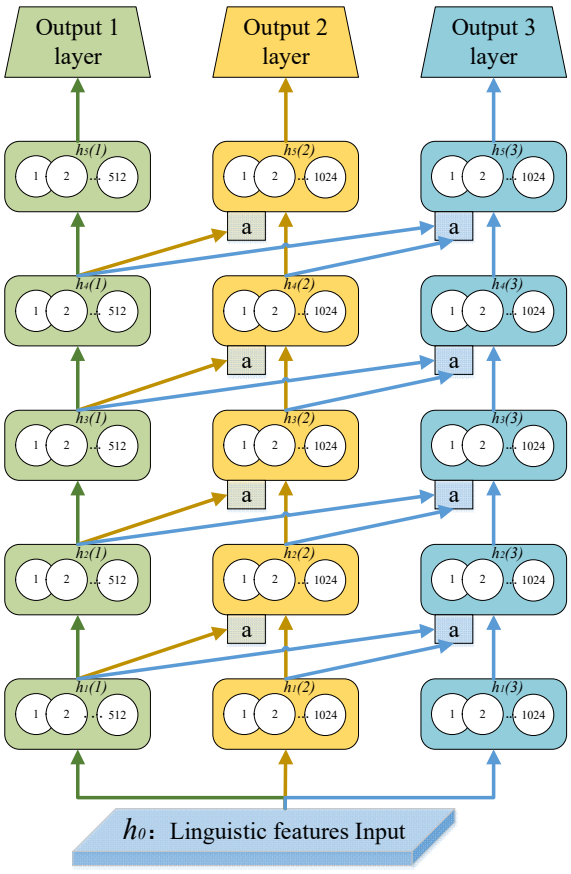

Figure 1 *PDNN framework for SPSS. The arrows represent dense connections among each layer. Blocks a represent the adapation layers as lateral connections. Each color of columns are 'frozen' after training. Output 1 to 3 are the acoustic features output U/V, F0, LSP accordingly. Input* h_0 *is the linguistic features, which is same for all the three columns.*

3. Experiments and Results

3.1. Database and features

A Mandarin database, which contains 10,000 phonetically rich sentences from a professional female broadcaster, is adopted in this paper: 9000 sentences as training set, 500 sentences as validation set, and the rest 500 sentences are reserved as test set. Each sentence has around 13 words.

- **Acoustic features:** All speech recordings are sampled at 16 kHz, windowed by a 25 ms window, and shifted every 5 ms. 40th-order line spectral pair (LSP) coefficients, the fundamental frequency (F0) in log scale and voiced/unvoiced (V/UV) flag are extracted with STRAIGHT [25].
- **Linguistic features:** The phonetic and prosodic contexts of Mandarin are included: The phone identity, the position of a phone, syllable and word in phrase and sentence, POS of word, prosodic phrase, intonational phrase and sentence, the length of prosodic word, prosodic phrase, intonational phrase and sentence, etc.

The input numerical features are normalized to the range of $(0, 1]$ and the frame level forced alignment upon the training data is processed with a HMM system implemented by HTS toolkit [12]. The target acoustic features are normalized to zero mean and unit variance before training. The dimension of the input linguistic features is 211. And the output contains acoustic features of each target, with 1 dimension of U/V, 9 dimensions of logF0 with their and 123 dimensions of LSP.

3.2. Experimental step

Six types of systems are implemented for comparison:

- **DNN-C:** Standard DNN-based approach. All the acoustic features are concatenated together and treated as one stream.
- **DNN-I:** Each type of the acoustic features is trained independently with separate DNN models.
- **MTL-DNN:** DNN approach with MTL method. One task is the prediction of U/V and F0. Another task is the prediction of spectral features (LSP). Different coefficients α are tested.
- **MTL-PDNN:** Using the similar PDNN structure to train all the targets synchronously by MTL method. We separate the F0 parameters into two parts: U/V and F0. Thus the combination considering the quantities of tasks and different prediction sequences can be concluded as following: (2: U/V F0, LSP), (2: LSP, U/V F0), (3: U/V, F0, LSP) and (3: LSP, U/V, F0). Different coefficients α are tested.
- **PDNN:** The proposed PDNN method. Similar with MTL-PDNN, different quantities of tasks and different prediction sequences are tried.
- **PDNN-FT:** After the training PDNN-3 (PDNN with 3 tasks), a fine-tuning procedure is done by the MTL method.

For testing, the outputs of all the systems are fed into a parameter generation module to generate smooth feature parameters with the dynamic constraints. Then formant sharping based on LSP frequencies is used to reduce the over-smoothing problem in modeling. The speech waveforms are synthesized by LPC synthesizer with generated speech parameters finally.

Our implementation is in TensorFlow [26] and we use the RMSProp optimizer with the global initial learning rate 0.0005 and its Tensorflow defaults parameters. We choose ReLU [27] as the activation function.

3.3. Objective evaluation

In objective evaluation, the generated features are assessed by comparing the distortions between the features extracted from natural speech in the test set and the generated ones predicted from different systems. Specifically, the duration extracted from natural speech is used directly in prediction. Table 1 shows the Objective measures of different models for speech synthesis. The architecture describes the number of nodes that DNN use in each type of systems. The number of the trained parameters is estimated. The coefficient α set in the table is the configuration that achieves the best performance in each module by MTL method.

As illustrated in Table 1, the LSD of the proposed PDNN method has reduced by thirteen percent compared to the standard DNN methods. The MTL-PDNN method performs better than the standard MTL-DNN method in the all objective measures. The MTL-PDNN separates the prediction of the U/V, F0 and LSP in different sub-networks. The knowledge transfer flow is defined by the lateral connections. Compared with the MTL-PDNN, the proposed PDNN have also achieved improvements in the objective measures. It illustrates that the goal of the MTL method is to minimize the combined global loss function, which is relevant to the objective measures of spectrum and F0. Through epochs of training, it would reach the optimum. But it won't be easy for each sub-target to reach its own optimum because other targets would also have an effect on the parameters of the entire networks. For MTL, it is hard to distinguish which parameters to learn the specific. On the contrary, it is easy to distinguish for PDNN because each task is trained by each sub-network. The transfer of memory depends on the lateral connections between these sub-networks. Compared with PDNN, the PDNN-FT does not achieve better performance. Some objective measures remain unchanged or turn worse. The fine-tuning procedure breaks the optimum that the PDNN has achieved already. It indicates that the PDNN step by step training mode, which is training separately by memorizing the knowledge from trained tasks, is better than MTL method with global optimal training.

One thing we need to consider is the sequence of targets. So we did sets of the experiments on the sequence of predictions. According to the results, we can draw the conclusion that which target is predicted later, the better performance we can get. Predicting the F0 parameters first has a better overall performance than predicting the spectral parameters first. We infer that it is more helpful for F0 trajectory to reconstruct the spectrum envelope.

Compared with PDNN-2, the PDNN-3 split the F0 parameters prediction into two steps, which is to predict LF0 based on the prediction results of U/V. Experimental results show that the performance of the U/V error and LF0 RMSE further improve after adding a task. It illustrates that the information process from text to human speech is a very complex mapping. With more intermediate variables and layers, the more information of speech can be provided for reconstruction.

3.4. Subjective evaluation

Figure 2 shows Mean opinion score (MOS) results for the naturalness of synthetic speech. 30 utterances from the test set

are randomly selected as the testing material. Each utterance is generated from the DNN-C, DNN-I, MTL-DNN, MTL-PDNN, PDNN and PDNN-FT systems, which is set to the configuration reaching the best objective evaluations. The testing speech are randomly shuffled to avoid preferential bias. 30 listeners are invited to take part in the evaluation of synthetic speech.

The proposed PDNN achieves best MOS results at 3.65, which indicates the ability of PDNN in improving the precision of acoustic model and naturalness of synthetic speech. Compared to the PDNN, the synthetic speech by the PDNN-FT system has a slight drop in the naturalness of speech.

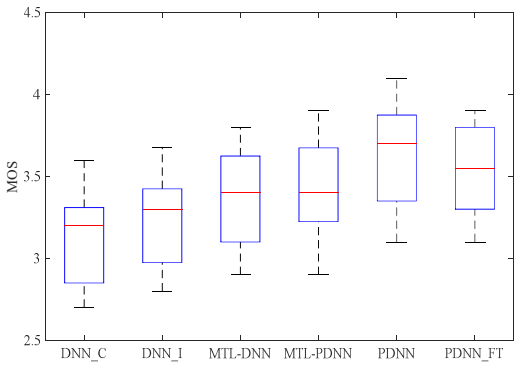

Figure 2 *Boxplot of Naturalness MOS results for six types of SPSS system*

3.5. Computational complexity analysis

The sharing weights strategy make the quantity of parameters in the lateral connections increase at linear growth with the size of networks rather than in factorial growth. In the training stage, each column has to be trained one by one. The PDNN method would increase 2 to 3 times of training time compared with the standard DNN methods. In the runtime of the SPSS synthesizer, the delay caused by the lateral connections can be ignored.

4. Conclusions

In this paper, we present a progressive deep neural networks framework for speech synthesis, which could predict different types of acoustic features one by one. The PDNN is immune to forget the previous memory on processing the linguistic features. Each training process has its own train criterion. So each type of acoustic features can be trained to its own optimum. PDNN with different topology, quantity of hierarchies, MTL using similar PDNN structure, and fine-tuning after PDNN3 by MTL are compared in the experiments. Compared to the standard DNN-based and MTL-DNN with methods, both objective and subjective experimental results demonstrated the better performance of the PDNN.

Our future research will try to use other networks, such as recurrent neural networks and BLSTM, to replace the DNN for acoustic model. Besides, the idea of progressive neural networks can be applied to other fields in speech synthesis, such as voice conversion.

5. Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61425017, No. 61773379, No. 61603390, No. 61771472), the National Key Research & Development Plan of China (No. 2018YFB1005003) and Inria-CAS Joint Research Project (173211KYSB20170061).

Table1: *Objective evaluation for the system DNN-C, DNN-I, MTL-DNN, MTL-PDNN, PDNN and PDNN-FT. RMSE of LogF0 is computed in Logarithm frequency. V/UV error means frame-level voiced/unvoiced error. LCD is Linear Cepstral Distortion.*

Model		Architecture	Parameter (million)	Coefficient α	LSD (dB)	V/U Err (%)	LogF0 RMSE (Hz)
DNN	DNN-C	5*1024	5.24	/	7.83	5.37	0.225
	DNN-I	5*1024*2	10.49	/	7.26	5.29	0.218
	MTL-DNN	5*1024	5.24	0.6	7.18	5.26	0.207
MTL -PDNN	2:LSP,U/VF0	5*1024*2	10.50	0.7	6.58	4.98	0.207
	2:U/VF0,LSP	5*1024*2	10.50	0.6	6.53	4.96	0.211
	3:LSP,U/V,F0	5*512*1 +5*1024*2	11.82	0.9	6.56	4.98	0.203
	3:U/V,F0,LSP	5*512*1 +5*1024*2	11.82	0.5	6.48	4.92	0.208
PDNN	2:LSP,U/VF0	5*1024*2	10.50	/	7.26	4.93	0.194
	2:U/VF0,LSP	5*1024*2	10.50	/	5.86	5.29	0.218
	3:LSP,U/V,F0	5*512*1 +5*1024*2	11.82	/	6.79	**4.63**	**0.191**
	3:U/V,F0,LSP	5*512*1 +5*1024*2	11.82	/	**5.83**	4.63	0.196
PDNN -FT	3:LSP,U/V,F0	5*512*1 +5*1024*2	11.82	0.7	5.98	4.78	0.193
	3:U/V,F0,LSP	5*512*1 +5*1024*2	11.82	0.4	5.86	4.72	0.194

6. References

[1] A. Black, N. Campbell, “Optimizing selection of units from speech database for concatenative synthesis,” in ICASSP-1996-IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. p. 373-376.

[2] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in ICASSP-1996- IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. p. 373-376.

[3] R. E. Donovan, P. C. Woodland, “A hidden Markov-model-based trainable speech synthesizer,” Computer Speech & Language, 1999, 13(3):223-241.

[4] T. Merritt, R. A. J. Clark, Z. Wu, et al. “Deep neural network-guided unit selection synthesis,” in ICASSP-2016-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2016:5145-5149.

[5] T. Capes, P. Coles, A. Conkie, et al, “Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System,” in INTERSPEECH 2017–Annual Conference of the International Speech Communication Association. 2017:4011-4015.

[6] W. Vincent, A. Yannis, S. Hanna, et al, “Google’s Next-Generation Real-Time Unit-Selection Synthesizer Using Sequence-to-Sequence LSTM-Based Autoencoders,” in NTERSPEECH 2017 –Annual Conference of the International Speech Communication Association, 2017:1143-1147.

[7] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” Speech Communication 51.11: 1039-1064,2009.

[8] A. V. D. Oord, S. Dieleman, H. Zen, et al, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499,2016.

[9] J. Sotelo, S. Mehri, K. Kumar, et al, “Char2Wav: End-to-end speech synthesis,”in ICLR2017 workshop submission, 2017.

[10] S. O. Arik, M. Chrzanowski, A. Coates, et al. “Deep Voice: Real-time Neural Text-to-Speech,” in ICML, International Conference on Machine Learning, 2017

[11] Y. Wang, R. Skerry-Ryan, D. Stanton, et al, “Tacotron: Towards End-to-End Speech Synthesis,” in INTERSPEECH 2017 –Annual Conference of the International Speech Communication Association,2017,4006-4010.

[12] T. Yoshimura, K. Tokuda, T. Masuko, et al, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in Sixth European Conference on Speech Communication and Technology. 1999.

[13] Z. H. Ling, L. Deng, D. Yu, “Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis,” IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(10): 2129-2139.

[14] S. Kang, X. Qian, H. Meng, “Multi-distribution deep belief network for speech synthesis,” in ICASSP-2013-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013: 8012-8016.

[15] H. Zen, A. Senior, M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in ICASSP-2013-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013: 7962-7966.

[16] H. Zen, H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” in ICASSP-2015-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2015: 4470-4474.

[17] Y. Fan, Y. Qian, F. L. Xie, et al, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in INTERSPEECH 2014–Annual Conference of the International Speech Communication Association, 2014.

[18] Z. Wu, C. Valentini-Botinhao, O. Watts, et al, “Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis”, in ICASSP-2015-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2015: 4460-4464.

[19] Z. Wen, K. Li, Z. Huang, et al, “Improving Deep Neural Network Based Speech Synthesis through Contextual Feature Parametrization and Multi-Task Learning,” Journal of Signal Processing Systems, 2017(4):1-13.

[20] R. Li, Z. Wu, X. Liu, et al, “Multi-task learning of structured output layer bidirectional LSTMS for speech synthesis,” in ICASSP-2017-IEEE International Conference on Acoustics, Speech, and Signal Processing, 2017:5510-5514.

[21] K.N. Stevens, G. Weismer, “Acoustic Phonetics,” Acoustic phonetics. Journal of the Acoustical Society of America, 2001, 109(1), 607-607.

[22] S. Arthi, and T. V. Sreenivas. "Influence of time-varying pitch on timbre: "Coherence and incoherence" based on spectral centroid," in ICASSP-2015-IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4240-4244, 2015.

[23] Karimian-Azari, Sam, Nasser Mohammadiha, Jesper R. Jensen, and Mads G. Christensen. "Pitch estimation and tracking with harmonic emphasis on the acoustic spectrum." in ICASSP-2015-IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 4330-4334, 2015.

[24] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, et al, “Progressive neural networks,” arXiv:1606.04671

[25] H. Kawahara, I. Masuda-Katsuse, D. Cheveign, and et al. “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, 1999, 27(3–4):187-207.

[26] M. Abadi, A. Agarwal, P. Barham, et al, “Tensorflow: large-scale machine learning on heterogeneous distributed systems”, 2016.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS- Annual Conference on Neural Information Processing Systems, 2012, pp. 1106–1114.