Interspeech 2018
2-6 September 2018, Hyderabad

Deep Metric Learning for the Target Cost in Unit-Selection Speech Synthesizer

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

This paper describes a unified Deep Metric Learning (DML) framework to predict the target cost directly by supervised learning method. The conventional methods to calculate the target cost include two separate steps: feature extraction and standard distance measurement. The proposed DML framework aims to measure the similarity between the candidate units and the target units more reasonably and directly. Firstly, the symmetrical DML framework is pre-trained to learn the metric between pairs of candidate units and target units. The relabeling procedure is added to correct the initial designed labels of the target cost. Secondly, the acoustic features of the target units are removed, which fits the runtime of the unit-selection synthesizer. The asymmetrical DML is fine-tuned to learn the metric between candidate units and target units. Compared with the conventional methods, the proposed unified DML framework can avoid the accumulation of errors in separate steps and improve the accuracy in labeling and predicting the target cost. The evaluation results demonstrate that the naturalness of synthetic speech has been improved by adopting DML framework to predict target cost.

Index Terms: speech synthesis, unit-selection, target cost, deep metric learning

1. Introduction

The unit-selection speech synthesis [1] has been challenged by the statistical parametric speech synthesis (SPSS) [2] and advanced methods (WaveNet, Deep Voice, Tacotron) [3-9] recently. However, the above advanced methods still need more delicate works in computational efficiency and robustness. And the SPSS system tends to generate "average" speech which would defect the perception of sound. When the speech corpus is highly-curated or the studio-level quality of the synthetic speech is required, the unit-selection synthesizer is preferred.

One of the core problem for unit-selection synthesizer is the discontinuousness between the selected adjacent basic units. People would identify that the selected sequence of units are extracted from difference utterances when acoustic clues (such as the intonation, the speaking style, and the speed) are unmatched, which would defect the perception of sound.

The target cost and the concatenation cost are defined to decide the best candidate from the corpus database. The target cost is designed to select the proper candidates from the database. The concatenation cost is designed to select adjacent units sound more coherently. The target cost, the metric of the similarities between candidate units and target units, is the foundation of selecting the proper combination of candidate units. And the target cost is different to define and predict.

Hunt and Black first presented current form of unit-selection speech synthesis system. The linguistic features were applied for target cost calculation. Then the hybrid unit-selection method added the acoustic features, which was predicted by the statistical model, for target cost calculation [10]. And several improvements were proposed, such as using the Deep Neural Networks (DNN) to generate the guiding acoustic parameters [11-13]. To avoid the error in extracting features from generated acoustic parameters, recurrent mixture density networks (MDNs) were applied for predicting target and concatenation distributions [14]. These features included duration, Mel-Frequency Cepstral Coefficient (MFCC), fundamental frequency (f0) and their dynamic counterparts.

The features that the above methods applied were hand-crafted. The automatic features generating methods were proposed. The frame-level embedding was extracted from the intermediate layers of a DNN [15] or a long short-term memory (LSTM) [13] network. In both cases, the unit-level embedding was constructed heuristically rather than being extracted from the whole unit directly. Then a sequence to sequence LSTM-based auto-encoders method was proposed to encode variable-length audio to a fixed-length vector [16]. The metric of the trained embedding was designed to represent the acoustic similarities between units.

The above methods mainly included two separate steps: feature extraction and standard distance measurement. These methods focused on finding and improving the accuracy of the intermediate representations for acoustic similarity. The target cost was calculated by standard distance measurement (e.g. L_1, L_2 norm). The intermediate representations might not be good enough to reveal differences between units. On the contrary, this paper proposes a unified Deep Metric Learning framework that measures the candidate units and the target units directly. The input of the DML framework is the asymmetric pair from the candidate units and the target units. And the internal structure of DML framework can predict the acoustic features based on the linguistic features of target units. The output is the target cost for unit-selection synthesizer, which is predicted by the supervised metric learning method.

Metric learning methods [17-19] had achieved state-of-the-art results in some areas. Compared with standard distance measures, the learned metric is more discriminative for the task. The features of pairs that above metric learning methods extracted are symmetric. In the runtime of the unit-selection synthesizer, the input is only text, which is the target linguistic features. Meanwhile, the candidate units have linguistic features and acoustic features. The features of candidate units and target units are asymmetric.

The main idea of the DML is inspired by a "Siamese" neural network for person re-identification task [20]. Their work is to use a "Siamese" DNN [21] to assess two person

10.21437/Interspeech.2018-1305

images. In their work, the cosine layer is used as the last layer. For two persons images x and y, the similarity equation can be written as $s = Cosine(B(x), B(y))$, where B denotes the sub-networks. Different from their work, the sub-network adds module that can model the mapping between linguistic features and acoustic features. The features extraction module is modified for fusing two channels of features. And two stages training is applied to tackle the asymmetric pair of the input and the inaccuracy of the target cost labeling.

The main contribution of this paper can be summarized as followed:

- The DML framework can learn a similarity metric between the candidate units and the target units directly, which is more effective than conventional methods including feature extraction and standard distance measurement two separate steps.
- Compared with the initial hand-crafted designed label, the DML framework can improve the accuracy of the target cost labeling
- Transfer learning method is applied for the training of the DML framework. The knowledge about acoustic and linguistic features extraction, which is acquired from the pre-training, is transferred to the asymmetric DML framework for target cost predicting.

The rest of the paper is organized as follows: Section 2 proposes the DML frameworks with pre-training and fine-tuning stages. Section 3 presents the experiments. And the results and analysis are presented in Section 4. The conclusions and future work are discussed in Section 5.

2. Deep Metric Learning Framework

Most of the neural network works in a standalone mode. The input of neural network is a sample and the output is a predicted label. For the metric learning between candidate units and target units, the DML framework is constructed to allow the two sub-networks work in a “sample pair → label” mode. The DML framework has two steps including pre-training and fine-tuning. The main procedure and structure of the DML framework can be concluded as the Figure 1.

In the pre-training stage, we select a part of the database as the target units. Both candidate units and target units have linguistic and acoustic features as the input of the symmetric DML framework. And the target cost label for the training is calculated by acoustic similarity between candidate audio and target audio, which is defined as the initial designed label. After finishing the pre-training stage, the predicted output replaces the initial designed label for the fine-tuning stage, which is called the relabeling procedure. In the fine-tuning stage, the branch of the target acoustic features is removed. And the rest of the asymmetric DML framework is fine-tuned in this stage. The input of target units is only linguistic features and the output is the predicting target cost for unit-selection synthesizer.

2.1. Training unit pairs selection and labeling

We select the training unit pairs based on whether the two units have the same vowel or consonant. If two units have different pronunciation, they are marked as the negative pair. The positive pairs are same pronunciation of units, which is selected based on rank from the pre-selection procedure of the unit-selection synthesizer [22]. Different from the regular pair label for metric learning, the initial designed label for pre-training stage is designed as follow: all the negative pairs are all labeled as -1. For the positive pairs, the Kullback Leibler divergence (KLD) of the four sub-unit sections between units are computed. The positive pairs are labeled in the range of (0.5, 1] based on the weighted normalized sums of the KLDs.

2.2. Pre-training of the symmetric DML framework

The input for each step of the symmetric DML framework is a pair of candidate and target units, which is candidate linguistic and acoustic features x_1, x_2 and target linguistic and acoustic features y_1, y_2 respectively. They are first passed into the convolutional neural network (CNN) separately. The linguistic features x_1 and y_1 are processed by CNN-L, while the acoustic features x_2 and y_2 are processed by CNN-A. In the combined branch C, the output of the CNN-L and the CNN-A are fusing together before the fully connected layer. Then the fused CNN output is connected to the DNN1 and DNN2 separately. And the output of DNNs is connected to the cosine

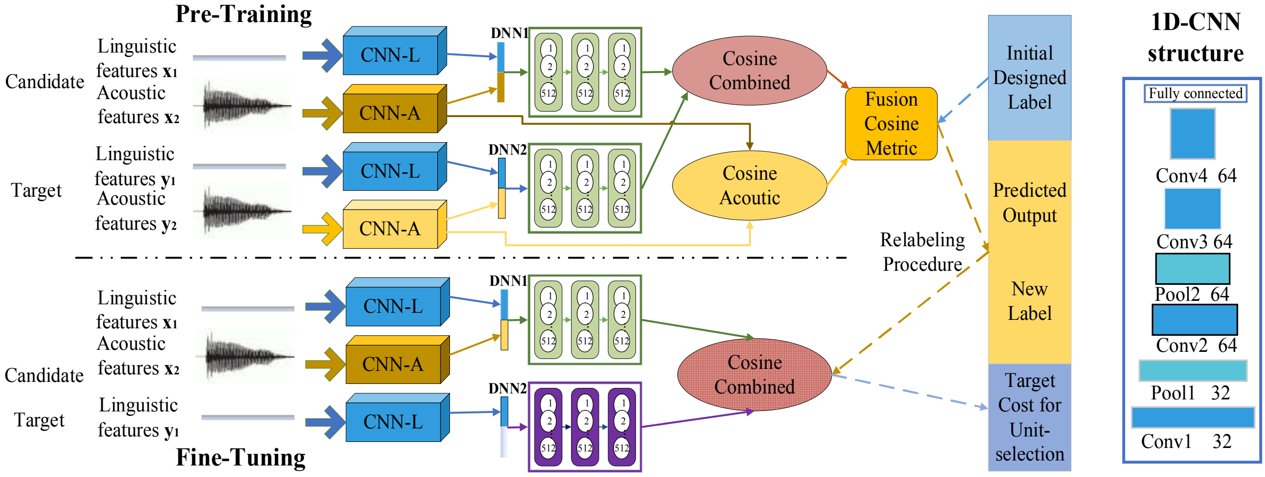

Figure 1: *Flowchart of the unified deep metric learning framework (Left) and the 1D-CNN structure (Right). Positive and negative unit pairs are feed into the symmetric DML framework for pre-training. After finishing the pre-training, the relabeling procedure replaces the initial designed label to the predicted similarity results. In the fine-tuning stage, positive pairs with target linguistic features are feed into the asymmetric DML framework, which accords with the runtime of unit-selection. The predicted output of the fine-tuning stage is the target cost for unit-selection synthesizer.*

layer that calculates the linguistic and acoustic combined similarity. In another acoustic branch A, the output of the CNN-A is connected to the acoustic cosine layer directly. The similarities are calculated by

$$s_A = \frac{A_1(x_2)^T A_2(y_2)}{\sqrt{A_1(x_2)^T A_1(x_2)}\sqrt{A_2(y_2)^T A_2(y_2)}} \quad (1)$$

$$s_C = \frac{C_1(x_1,x_2)^T C_2(y_1,y_2)}{\sqrt{C_1(x_1,x_2)^T C_1(x_1,x_2)}\sqrt{C_2(y_1,y_2)^T C_2(y_1,y_2)}} \quad (2)$$

where A_1, A_2, C_1 and C_2 are the functions of the two branches respectively.

Finally, the cosine fusion layer output the target cost s_p combined the s_A and s_C by the coefficient α ($0 \leq \propto \leq 1$)

$$s_p = (1 - \alpha)s_A + \alpha s_C \quad (3)$$

In the pre-training process, the two CNN-L for preprocessing the linguistic features of candidate and target share the same weights and bias. Besides, the two CNN-As and two DNNs also share the same weights and bias. After the pre-training stage is finished, the s_p predicted by the symmetric DML framework become the new label of the input pair.

2.3. Fine-tuning of the asymmetric DML framework

In the fine-tuning stage, the target units have only linguistic features y_1, which suits the runtime of the unit-selection synthesizer. The acoustic branch A and the CNN-A for processing the target acoustic features are removed. Only the positive pairs are feed to the asymmetric DML framework for fine-tuning. The label is the target cost predicted in the pre-training stage. In the training process, all the parameters are inherited from the pre-training stage. Only the weights and bias of DNN2 is fine-tuned. The weights and bias of CNN-L, CNN-A, and DNN1 are fixed. After the fine-tuning stage is finished, the s_C predicted by the asymmetric DML framework is the target cost for the unit-selection synthesizer.

3. Experiments

3.1. Database

A Mandarin database, which contains 30,000 phonetically rich sentences from a professional male broadcaster, is adopted in this paper. For the experiments described in this paper, the audio was down-sampled to 16 kHz. The 24,000 sentences of the database are chosen randomly as the candidates. The rest of the 6,000 sentences are chosen as the targets, in which 5,600 sentences as training set, 200 sentences as validation set, and the rest 200 sentences are reserved as test set. In the training set, each unit of sentences generates 50 pairs, which include 25 positive pairs and 25 negative pairs. Each utterance has around 15 units. To sum up, there are about 4 million unit pairs for pre-training and 2 million pairs for fine-tuning.

The linguistic features, which contain the phonetic and prosodic contexts of Mandarin in each unit, can be included as follows: 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 dimension of the acoustic features is 504.

To extract a fixed dimension of the acoustic features, each unit is equally divided into 4 sections. And the MFCC, fundamental frequency f_0 and duration are extracted in each section. The mean and variance of above are computed as the acoustic features of each unit. The input continuous features are normalized to the range of (0,1] and the input discrete features are encoded in One-Hot. The dimension of the acoustic features is 196.

3.2. Experimental setup

Backpropagation (BP) [23] is used to learn the parameters of the DML framework. Square loss is used as the cost function. Given a sample pair's similarity s $(-1 \leq s \leq 1)$ and their corresponding label l $(-1 \leq l \leq 1)$, the cost function is written as

$$J_{square} = (s - l)^2 \quad (4)$$

In the pre-training stage, the size of batch is 128 including 64 positive and 64 negative pairs. The number of negative pairs is far more than positive pairs. Therefore, we randomly select negative pairs from the whole negative sample pool for each batch. In the fine-tuning stage, the size of batch is 128 including 128 positive pairs. And the output dimension of the DNN is 256. The ReLU neuron [23] is used as activation function for each layer.

Two baselines about the calculation of target cost have been applied for comparison:

- **Baseline1**: It uses BLSTM based SPSS system to generate the acoustic parameters first and then calculate the KL divergence to get the final target cost [22].
- **Baseline2**: It uses BLSTM to predict the features, including mean and variance of the acoustic features in the 4 sub-units, to calculate the final target cost directly without generating the whole acoustic parameters sequence [14].

The basic unit in our experiment is vowels or consonants of the Mandarin, which resemble the syllables of the English. And a Viterbi search is used to find the best sequence that minimizes the combined cost. The combined cost is defined as:

$$C = \sigma C_{target} + C_{\text{concatenation}} \quad (5)$$

where σ denotes the target cost weight, C_{target} denotes the target cost, and $C_{\text{concatenation}}$ denotes the concatenation cost. Except for the calculation of the target cost, the other modules are described in [22].

3.3. Objective evaluation

To evaluate the accuracy of the predicted target cost, the root mean square error (RMSE) between the predicting target cost and the initial designed label of the test set is chosen as the objective metric. The target costs predicted by the two baselines are all normalized to the range of [0.5, 1] by the same scale used for the labeling the initial designed target cost. Different values of coefficient α is tested in the training of the DML.

3.4. Subjective evaluation

To evaluate the performance of unit-selection synthesizer with the modification of the target cost, 30 native speakers are arranged to evaluate the synthetic speech based on a 5-point discrete scale Mean Opinion Scores (MOSs) [24] labeled "Bad", "Poor", "Fair", "Good", and "Excellent". Each listener listens to 30 pairs random selecting sentences synthesized in each system. To investigate contribution that relabeling procedure for the training of the DML, a contrast experiment that uses the initial designed label to fine-tuning the framework is conducted, which is marked as "UR-DML".

Different target cost weight σ are tested to investigate the contribution of the target cost to the whole unit selecting procedure.

4. Results

As illustrated in the Figure 2, the proposed DML framework predicts the target cost more precisely than the two baselines. It illustrates that the DML can avoid the accumulation of errors in each module in the baseline methods. The cost function and structure let the DML framework to focus on the mission directly, which is predicting the acoustic similarity between the candidate units and the target speech. The knowledge about the extracting and preprocessing the linguistic and acoustic features is acquired from the pre-training of the symmetric DML framework. And the ability that let the partial asymmetrical DML framework predict the similarity with only linguistic features of targets is learned by fine-tuning the DNN in the branch of the target. The global optimal of DML achieves better performance than the sub-optimal of each module in the baselines.

The minimal RMSE is 0.055 when the coefficient α is set to 0.6 in the DML method. The following subjective evaluations are conducted on the same condition. Observing the fluctuation of RMSE in the DML method, the RMSE is decreasing while the coefficient α is decreasing from 1 to 0.6. It suggests that the introduction of acoustic metric module s_A in the pre-training stage can prevent the DML framework from learning the metric highly depending on the linguistic and ignoring the acoustic features. The attribute of many linguistic features is discrete, while all the attribute of acoustic features are continuous. Due to the subtle differences of acoustic features between units, the linguistic features are more sensitive to discriminate differences. If we only use the combined cosine layer alone in the pre-training procedure, the DML framework is prone to predict the target cost mainly based on the linguistic features, which deviates the motivation of the acoustic similarity metric.

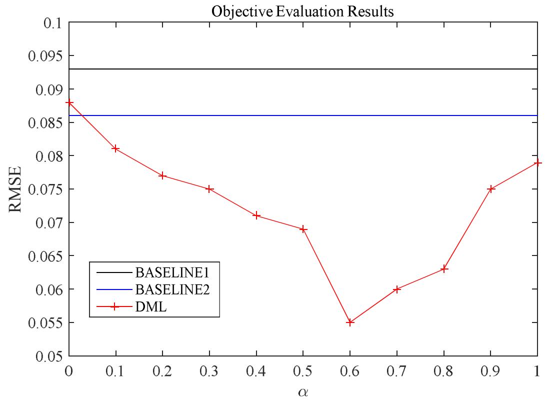

Figure 2: *RMSE of the predicted target cost using different coefficient α and systems. BASELINE 1 and BASELINE 2 systems have no α and are expanded to a line for comparison.*

Observing the MOS results illustrated in the Figure 3, the best MOS is 3.85 when the target cost weight is 1.5 in the DML experiment. The proposed DML with relabeling procedure achieves better performance than the baselines. And it also outperforms the UR-DML. These results indicate that the relabeling procedure corrects the initial designed label. The DML framework networks structure with "sample pairs→label" training mode can learn the knowledge about finding the similarity better. The motivation of the DML framework is to find the best positive pair rather than to judge whether the pair is similar. With the guidance of initial designed label, the DML structure is trained to extract better intermediate representations than hand-crafted features in acoustic similarity metric.

The MOS of baseline 1 and baseline 2 systems both decrease when the weight of the target cost increases. It indicates that target cost calculated by baselines could not reflect the similarity properly and the systems mainly depend on the concatenation cost to select the candidates. The target cost losses its designed function. Meanwhile, the systems with the DML method perform better when the weight of the target cost is increasing in certain range. It illustrates that the target cost of the DML is more precisely to measure the similarity between the candidates and the targets and can help the unit-selection to select the candidates better.

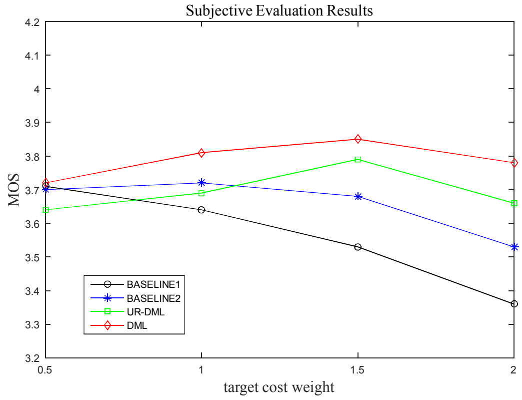

Figure 3: *MOS test for naturalness of synthetic speech using different target cost weight σ and systems.*

5. Conclusion

We present a unified Deep Metric Learning framework including pre-training and fine-tuning for the target cost in the unit-selection based synthesis system. The target cost is predicted directly, which avoids the accumulation of errors in each module. The "sample pair → label" training mode and the relabeling procedure improve the performance of the DML framework. We report an improvement of up to 0.13 MOS compared with the baseline using the BLSTM-guided hybrid method.

This paper focus on acquiring the target cost more directly and reasonably. However, the defined target cost in our method still involves human knowledge. In the future, the similarity between audio still need to be explored by using more delicate semi-supervised methods. Besides, the target cost is only part of the unit-selection synthesizer, the more direct method to choose a proper sequence of candidate units without calculating the target cost and the concatenation cost first is also our research focus.

6. 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. 2017YFB1002801) and Inria-CAS Joint Research Project (173211KYSB20170061).

7. References

[1] 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.

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

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

[4] A. V. D. Oord, Y. Li, I. Babuschkin, et al. ,"Parallel WaveNet: Fast High-Fidelity Speech Synthesis," arXiv preprint arXiv:1711.10433 ,2017.

[5] 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

[6] S. Arik, G. Diamos, A. Gibiansky, et al. “Deep Voice 2: Multi-Speaker Neural Text-to-Speech,” in NIPS- Annual Conference on Neural Information Processing Systems, 2017

[7] W. Ping, K. Peng, A. Gibiansky, et al, “Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning,” arXiv preprint arXiv:1710.07654 ,2017.

[8] 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.

[9] J. Shen, R. Pang, R. J. Weiss, et al, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions,” in ICASSP-2018- IEEE International Conference on Acoustics, Speech, and Signal Processing, 2018.

[10] Z.-H. Ling, L. Qin, H. Lu, et al, “The USTC and iflytek speech synthesis systems for Blizzard Challenge 2007,” in Blizzard Challenge Workshop, 2007.

[11] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, “Using deep bidirectional recurrent neural networks for prosodictarget prediction in a unit-selection text-to-speech system,” in INTERSPEECH 2015–Annual Conference of the International Speech Communication Association, pp. 1606–1610.

[12] T. Merritt, R. A. 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, pp. 5145–5149.

[13] L.-H. Chen, Y. Jiang, M. Zhou, et al, “The USTC system for Blizzard Challenge 2016,” in Blizzard Challenge Workshop, 2016.

[14] 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.

[15] 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, March 2016, pp. 5145–5149.

[16] 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.

[17] W.-S. Zheng, S. Gong, and T. Xiang, “Reidentification by relative distance comparison,” Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 3, pp. 653–668, 2013.

[18] M. Kostinger, M. Hirzer, P. Wohlhart, et al, “Large scale metric learning from equivalence constraints,” in CVPR2012-Computer Vision and Pattern Recognition, IEEE Conference on, 2012, pp.2288–2295.

[19] W. Li and X. Wang, “Locally aligned feature transforms across views,” in CVPR2013-Computer Vision and Pattern Recognition, IEEE Conference on, 2013, pp. 3594–3601.

[20] Y. D, L. Z, L. S, et al. “Deep Metric Learning for Person Re-identification,” in International Conference on Pattern Recognition. IEEE, 2014:34-39.

[21] J. Bromley, I. Guyon, Y. LeCun, et al. “Signature verification using a siamese time delay neural network,” in NIPS-Annual Conference on Neural Information Processing Systems,1993, pp. 737–744.

[22] J.Tao, R. Fu, Y. Zheng, et al, “The NLPR Speech Synthesis entry for Blizzard Challenge 2017 ” in Blizzard Challenge Workshop, 2017.

[23] 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.

[24] Y. Agiomyrgiannakis, “Vocaine the vocoder and applications in speech synthesis,” in ICASSP-2015- IEEE International Conference on Acoustics, Speech, and Signal Processing, April 2015, pp. 4230–4234.