Automatic Lyrics Transcription using Dilated Convolutional Neural Networks with Self-Attention

*

Emir Demirel
Centre for Digital Music
Queen Mary University of London
London, UK
e.demirel@qmul.ac.uk

Sven Ahlbäck
Doremir Music Research AB
Stockholm, Sweden
sven.ahlback@doremir.com

Simon Dixon
Centre for Digital Music
Queen Mary University of London
London, UK
s.e.dixon@qmul.ac.uk

***Abstract*—Speech recognition is a well developed research field so that the current state of the art systems are being used in many applications in the software industry, yet as by today, there still does not exist such robust system for the recognition of words and sentences from singing voice. This paper proposes a complete pipeline for this task which may commonly be referred as automatic lyrics transcription (ALT). We have trained convolutional time-delay neural networks with self-attention on monophonic karaoke recordings using a sequence classification objective for building the acoustic model. The dataset used in this study, *DAMP - Sing! 300x30x2* [1] is filtered to have songs with only English lyrics. Different language models are tested including MaxEnt and Recurrent Neural Networks based methods which are trained on the lyrics of pop songs in English. An in-depth analysis of the self-attention mechanism is held while tuning its context width and the number of attention heads. Using the best settings, our system achieves notable improvement to the state-of-the-art in ALT and provides a new baseline for the task.**

***Index Terms*—automatic speech recognition, machine learning, deep learning, music information retrieval, automatic lyrics transcription, language modeling**

I. Introduction

In contemporary pop music, the linguistic content in the singing voice is generally referred as *lyrics* and the process of automatic retrieval this lyrics content from singing voice can then be defined as Automatic Lyrics Transcription [2] [3]. The automatic retrieval of pronounced words from speech signals is a widely developed research field and the state of the art systems by today can be successfully applied to industrial applications. However, the same level of robustness has not yet been reached when the input is singing voice. According to prior research, there are several domain specific reasons word recognition performance reduces in singing including domain specific acoustic characteristics [2], [4] and the alterations of word pronunciations [5]. Specifically from a machine learning perspective, the main bottleneck for achieving a robust system is the availability of training data with fine-grained annotations, to be used in a supervised learning framework.

In this study, we exploit such large-scale singing voice dataset, *DAMP - Sing! 300x30x2* [1] - released by Smule [1], where prompt-level[2] (and occasionally word-level) annotations are provided. The dataset consists of monophonic Karaoke recordings of pop songs by multiple performers providing near 150 hours of trainable audio data. However, this dataset has not been widely utilized for the purpose of training a word recognition system. Through our proposed framework, we aim to highlight one way of utilization of this dataset in a complete ALT framework and further conduct an in-depth self-attention analysis via fine-tuning experiments.

A robust system for the retrieval of sung lyrics has a variety of potential applications in music information retrieval (MIR) related tasks and the music tech industry. In karaoke and music education apps, the recognition of sung words is essential for tracking a performance and providing feedback to the user. In

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowsa-Curie grant agreement No. 765068

[1]Smule is a commercial Karaoke singing application. More info at https://www.smule.com/

[2]In Smule app, lyrics are prompted to the users as words or sentences depending on the song arrangement. Each prompted sentence or word annotation is referred as *prompt-level* annotation.

combination with techniques like query-by-humming, ALT can be utilized for the song identification and metadata retrieval tasks.

Our system uses deep neural networks for building the final acoustic model that is composed of 2D convolutional layers at the front end for extracting more robust features followed by time-delay layers due to their capability of modeling long-term context information. A self-attention layer is added before the final projection layer for weighting the time context when computing the output activations for classification.

Overall, this paper targets at making the following contributions:

- Presenting a complete pipeline for ALT in monophonic singing
- Testing different language modeling strategies for lyrics
- Proposing a neural network architecture combining Self-Attention with CNN and TDNN layers
- Performing an in-depth analysis for Self-Attention,
- Reporting best results for ALT on a public dataset
- Providing the code for open-science and reproducibility

This paper is structured as follows: literature on ALT on monophonic singing recordings is reviewed (II). Then the details of the data used in training and evaluation are given (III). Then the proposed system (IV) and the basis of our experiments (V) are explained. The results for each of experimental steps are shown and an in-depth analysis of the self attention parameters is performed (VI). Finally, potential improvements to the proposed system are discussed (VII).

II. Related Work

Since the early days of ALT, researchers have shown a tendency to adapting ASR techniques for the retrieval of words or phonemes from the singing voice. Mesaros et al. [4] adapted a pretrained GMM-HMM based speech recognizer to singing domain using Maximum Linear Likelihood Regression (MLLR) transformation. In [7], the vocal parts were separated, and their proposed method was able to achieve 70% Phoneme Error Rate (PER) and 88 % Word Error Rate (WER) on clean and monophonic singing.

The performance of the ALT systems proposed in recent years has had a notable improvement due to the availability of new open datasets for singing voice. Within the *DAMP* repository[3] there are a few separate open-source datasets for singing made publicly available for research by Smule. The repository consists of Karaoke recordings performed by real-world users of the Smule app. Due to its nature, the audio recordings in the datasets are mostly monophonic accapella singing voice recordings. On one of the earlier releases within the repository, the *DAMP (multiple songs)* [8] dataset where no line-level lyrics annotations are provided, Kruspe [2] trained a DNN-HMM system and reported 77% PER on subset of the aforementioned dataset as the evaluation set. Tsai et. al [10] use a speech-pretrained TDNN-BiLSTM neural network and retrains on a smaller dataset of 110 monophonic singing recordings obtained from Youtube and reports a WER of 73.09%. Gupta et al. [11] adapted a pretrained DNN-HMM based speech recognizer on the *DAMP* dataset and reported 36.32% WER on a carefully selected subset of this dataset. Further in their work [5], they achieved 29.65% WER by extending the length of pronounced vowels in the pronunciation lexicon. This idea is plausible not only due to the observed performance boost, but also inherently increasing the probability of a frame of a voiced phoneme (i.e phonemes with vowels pronounced) to be followed by another frame with the same phoneme, which is a frequent case in singing.

One of the main challenges of ALT research is the lack of benchmark evaluation sets. Most prior work reports results on different datasets, which makes it harder to compare and evaluate different systems. To overcome this problem, Gerardo et al. [6] curated a new training and evaluation set, namely the '*DSing Corpus*' based on the *Sing! 300x30x2* dataset [1] within the *DAMP* repository. On this new evaluation set, the authors reported 19.60% WER using a factorized Time-Delay Neural Network (TDNN-F) setting and a 4G Language Modeling which is trained on lyrics pop songs in English. To our knowledge, this is the best result reported for ALT from monophonic singing.

III. Dataset

All of the experiments of this study are performed on *Sing! 300x30x2* [3]. The dataset consists of 18,767 real-world karaoke recordings performed by 13,154 Smule app users, and includes the lyrics prompt timings that are shown to users for the purpose of singing along with the original versions of songs. The karaoke performances in the dataset are unique interpretations of the 300 most popular songs from 30 countries, where the popularity is determined by user votes via the Smule app.

[3]Can be accessed from *https://ccrma.stanford.edu/damp/*

Even though the dataset is curated in a balanced manner and time-level annotations are provided alongside the audio recordings, the data still needs preprocessing to be used for training and evaluation. For instance, the prompt timings are not always in the same granularity, i.e there are both word and sentence level annotations. Dabike et al. [6] curated a cleaner version of this dataset based on heuristics, with the goal of removing noisy data. The dataset is also filtered out to keep recordings with lyrics only in English providing nearly 150 hours training data.

From these recordings, 70 of them with 480 utterances from 43 singers in total are chosen for the *Test* set and 66 songs with 482 utterances from 40 singers for the validation (*Dev*) set, where these recordings from both sets are originating from the UK. In order to meet 'the gold standard' for evaluation, the annotations in the *Dev* and *Test* sets are corrected and validated by human experts.

IV. System Details

Our workflow is based on Kaldi [12] which is a WFST-based ASR framework that implements the state-of-the-art in research[4]. WFST-based speech recognizers [13] build a decoding graph composing separate lexicon, acoustic and language models. We use the CMU English Pronunciation Dictionary[5] as the lexicon, and train the acoustic and language models via separate neural networks using in-domain data. In this section, we give the details of different components in our automatic lyrics transcription pipeline.

A. Dilated Convolutional Neural Networks with Self-Attention

For the acoustic model, we propose a deep convolutional neural neural network structure that has three major components: six 2-D convolutional layers at the front-end followed by a stack of sub-sampled factorized Time-Delay Neural Network (TDNN-f) [14], and finally, a time-restricted self-attention layer [15]. The TDNN-f layers are essentially equivalent to dilated 1-D convolutions on the time axis. For this reason, we refer our architecture as a 'Dilated Convolutional Neural Network'.

The CNN part consists of six convolutional layers with 3x3 filters. Subsampling (i.e MaxPooling) with a factor of 2 is applied after layers CNN_3, CNN_5, CNN_6, to extract more robust features for TDNN-F layers and to reduce the size of the feature vectors. The rest of the CNN parameters are shown in Table 1.

	CNN_1,2	CNN_3	CNN_4	CNN_5	CNN_6
Height	40	40	20	20	10
Num_filters	48	64	64	64	128
Subsamp. factor	N/A	2	N/A	2	2

TABLE I: Settings of the 2D convolutional layers

Time Delay Neural Networks (TDNN) are widely used in ASR systems due to their capability of successfully modelling long-term context and makes parallelization possible unlike RNNs. TDNN-F part of the network consists of nine hidden layers with dimension 1024 and a bottleneck dimension of 128. We use the same factorization setting in [14] for the TDNN-F layers. Each layer in the network is followed by a ReLU and batch normalization. This architecture without the self-attention layer is set as the baseline NN model in our experiments. We refer this part of the architecture as CTDNN in this paper.

Since its introduction in the seminal paper [16], the self-attention mechanism has been used in many state-of-the-art systems in Natural Language Processing (NLP) [17]. Recently, this mechanism is also successfully applied in speech recognition [18] [19]. Motivated by its success in ASR, we further extend our architecture design by adding a multi-head time-restricted self-attention layer [15] on top of the network, right before the final linear projection layer. In this attention design, multiple heads learn weights in parallel, focusing on different parts of the time context. Multi-head mechanism is shown to be useful in [15]. We have set a fixed size for key and query dimensions to 60, value to 40 and tune the context width and the number of heads in the experiments. The models that include the self-attention layer will be referred as CTDNN_SA.

After the attention layer the weights are projected to one dimensional vectors via `Linear Layer` continued by softmax. There are two separate output layers that are jointly trained on different objective functions: Maximum Mutual Information (**Output - chain**) and Cross-Entropy (**Output - xent**). This joint training is done for regularization as explained in Section V.

B. Language Models

In our lyrics transcription framework, we use Language Models (LM) that are constructed using in-domain text data

[4]Kaldi-ASR is an open-source toolkit and can be accessed from *kaldi_url*=https://github.com/kaldi-asr/kaldi/blob/master/

[5]Can be accessed from http://www.speech.cs.cmu.edu/cgi-bin/cmudict

that consists of the lyrics of all songs by all artists from UK, USA and Australia in the *DSing30* corpus and the artists from *the Billboard 'The Hot 100'* for the years from 2015 to 2018, which is the same training corpus as in [6]. The training of language models are validated on a set of lyrics from the songs in the validation (*Dev*) set of the DSing30 corpus. The lyrics of the songs in the *Test* set are excluded from the training text corpus when building the LMs. Some statistics of the training text corpus are shown in Table II.

	Training Corpus	Validation Corpus
Words	7,931,215	4018
Sentences	1,117,152	482
Songs	4,324	66

TABLE II: Some statistics of the LM training & validation corpora

In this study, we test three different language models, 3-gram (3G) and 4-gram (4G) LMs train with a maximum entropy objective (MaxENT) built using SRILM [20] and an LM built using Recurrent Neural Networks (RNNLM) [21].

Lyrics may contain made-up words that are not necessary included in the dictionary used in the decoding graph. To handle such words that may exist in the evaluation set, we followed the approach at [6]. This procedure helps with the word insertion penalty in scoring when ambigious phone sequences are observed. By ambigious phone sequences, we mean phone sequences that do not exist in the pronunciation dictionary provided. The resulting language model used in decoding will be referred as *n*-G_*unk* in the experimental results.

V. Experimental Setup

We adapt a similar procedure with the '*chain*' recipes[7] in Kaldi for training the neural networks. First a GMM-HMM triphone model is trained to generate phone-level alignments. Based on this model and the alignments, a data cleanup strategy is applied. Using the alignments on the cleaned data, a sequence classifier is trained with the NN architecture proposed in Section IV.

A. GMM-HMM Training

For training the GMM-HMM system, we use 13-band MFCCs plus delta and delta-delta features with a frame length of 25 milliseconds and a hop size of 15 milliseconds. ± 4 neighbouring frames are concatenated on the feature vectors (9 consecutive frames in total) to obtain context dependent features. Zero-mean normalization per singer is applied to MFCCs. The spliced frames are then projected to 40 dimensional feature vectors using Linear Discriminant Analysis (LDA). Maximum Likelihood Linear Transformation (MLLT) is applied on the LDA features. Further, 'speaker adaptive training (SAT)' is applied by transforming the feature space (LDA + MLLT) using feature-space maximum likelihood linear regression (fMLLR) per singer [22].

A data cleanup strategy is applied with the goal of removing the noise and bad portions of transcripts from the training data using the script at [8]. We then retrain another GMM-HMM triphone model on the clean training set and generate alignments using this model to be used in neural network training.

B. Neural Network Training

The acoustic model is trained using Convolutional Time-Delay Neural Networks (CTDNN) with a lattice-free sequence discriminative training strategy that uses Maximum Mutual Information (MMI) as the objective criterion (LF-MMI) [23]. A 3-way *speed-perturbation* [24] is applied for data augmentation, changing the speed of the original signal with the factors of 0.9 and 1.1. Speed perturbation helps to increase the generalizability by modeling different durations for phonemes. Then, we generate lattice alignments as the soft target phone boundaries using the triphone GMM-HMM model prior to training the network. As opposed to GMM-HMM training, we use mel-spectrogram features with 40 filter banks as the input to the neural network. Our framework performs SAT for training the neural networks based on i-vector extraction [25] as the input speaker (singer) representations with the height of 200. The i-vector speaker representations are projected to smaller matrices with the height of 40 and then combined with filter bank features as separate feature maps but sharing the same trainable weight matrix. The convolutional layers are preceded by a fully connected layer that applies a linear transformation to the input with a trainable matrix to feed to the convolutional layers.

One of the challenges in sequence-level training is that it is prone to overfitting. To overcome this problem we use the combination of three regularization techniques during training [26]: Cross-entropy regularization, *l-2* regularization and leaky

[6] *kaldi_url*/egs/wsj/s5/utils/lang/make_unk_lm.sh

[7] *kaldi_url*egs/wsj/s5/local/chain/

[8] *kaldi_url*/egs/wsj/steps/cleanup/clean_and_segment_data.sh

HMMs. In addition, dropout is used at each layer to alleviate the risk of overfitting. Dropout scheduling is used as suggested in [27]. Training is done on minibatches having sizes of 128 where the data chunks of variable sizes of 140, 100, 160 processed in each minibatch. We use preconditioned Stochastic Gradient Descent (SGD) as the optimizer. The initial and final learning rates are set to 0.0005 and 0.00005 respectively, where the learning rate is shrunk after each iteration (mini-batch processing). Half of the global learning rate is applied to the last and second-to-last layer weights. We train the network for 8 epochs. After the final iteration of training, we combine models from the last 10 iterations into a final single model using a weighted-average operation [28].

VI. Results and Discussion

In this section, we provide transcription performances of various models studied in this paper. These models include the baseline GMM-HMM model and neural networks with different language models. The effect of number of heads in the attention layer is tested. The training and validation losses of best performing neural network based models are illustrated.

A. GMM-HMM Baseline

First, we observe how much performance gain can be achieved using a larger dataset by comparing the WERs of GMM-HMM triphone models trained on *DSing1* and *DSing30* corpuses. In Table III, the scores are obtained using the 3G MaxEnt language model. Around 10% performance gain is observed when using the larger dataset. This improvement is due not only to the larger train set but also because the *DSing30* corpus has more variety in singing accents resulting in learning a more generalizable model.

Train Set	**Dev WER (%)**	**Test WER (%)**
DSing1	63.51	63.12
DSing30	52.69	50.80

TABLE III: WERs of Triphone GMM-HMM models on Dev / Test Sets using 3-gram LM.

B. Language Models

At this step, our goal is to find the most suitable LM for our task. First, the scoring is done by decoding the FST built using the 3G LM. Then generated lattices are rescored using a graph with 4G LM. Scoring is redone using LMs with 'unknown' language modeling. The second-pass scoring is done on RNNLMs for both n-gram models using a pruned rescoring method [29]. The scores in Table IV are obtained using the baseline CTDNN architecture.

LM model	MaxEnt		RNNLM	
	Dev	*Test*	*Dev*	*Test*
3G	24.84	20.84	20.93	17.44
3G_unk	24.56	20.84	20.98	17.47
4G	21.20	18.59	18.09	16.23
4G_unk	**21.08**	**18.57**	**17.70**	**15.65**

TABLE IV: WERs (%) of different language models using the CTDNN model

It is seen that RNNLM's outperform the MaxEnt LMs for both n-grams having 2-3 % WER improvement. Composing the n-gram_unk language model generally helps with the final WER score.

C. Neural Networks

In general, the proposed NN architectures in this study outperform the previously reported score of 23.33 on *Dev* set and 19.60 WER (%) on *Test* set using a 4-gram (4G) MaxEnt language model [6]. Using the same settings with the aforementioned study, the baseline CTDNN model obtained 18.57% on the *Test* set showing around 1% improvement.

The train and validation log-probability losses (vs. iterations) of CTDNN and CTDNN_SA with different contexts are shown in Figure 2. The losses in models with SA have lower values than the baseline CTDNN model. No explicit sign of overfitting is observed from the train/validation loss plots for any of the models. Early-stopping is not used, yet we have not observed any improvement with further training.

D. Self-Attention Analysis

At this stage, we aimed at tuning the self-attention layer by testing the number of heads and the context width by testing different configurations and performed an in-depth analysis of what the self-attention layer has learned. Rows # 1 and #3 in Figure 2 show the attention weight vectors for each head in the self-attention layer and below are the averaged values over all heads.

From (a,b,c,d), it is seen that the distribution of the attention weights gets closer to a uniform distribution as the number of heads gets higher, which implies that the self-attention

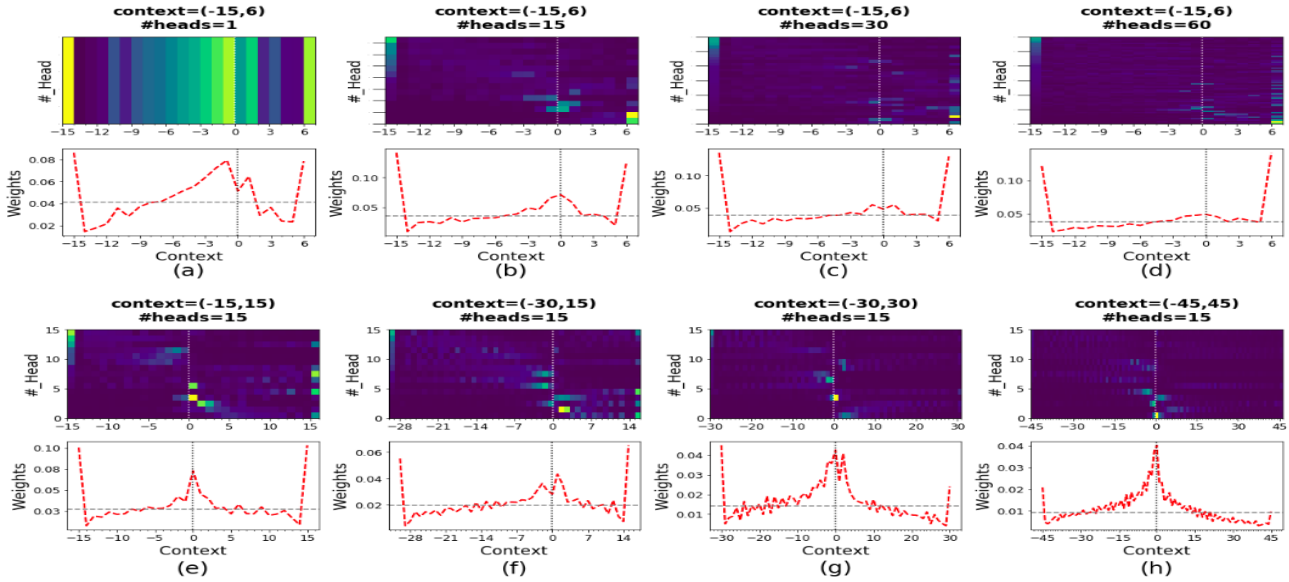


Fig. 1: Illustration of the weight vectors in the self-attention layer. (a,b,c,d) shows the weights for 1, 15, 30, 60 number of heads respectively with a context of (-6,15). Context of (e,f,g,h) are (-15,15), (-30,15), (-30,30), (-45,45) respectively with 15 heads. The figures on top are the the weights for all the heads sorted bottom-to-top w.r.t the weight of leftmost context. The dashed horizontal line in the bottom figures indicates the median of the averaged weights.

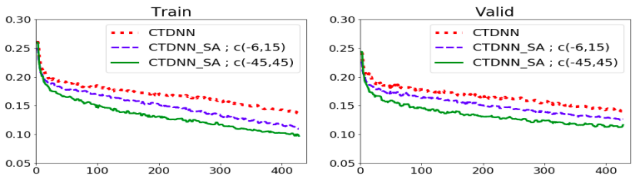


Fig. 2: Log-probability losses of neural network models. 'cw' in legends stand for 'context'. Both models with SA have 15 heads.'

NN Model	**#heads**	MaxEnt		RNNLM	
		Dev	*Test*	*Dev*	*Test*
TDNN [6]	N/A	23.33	19.60	N/A	N/A
CTDNN	N/A	21.08	18.57	**17.70**	15.65
CTDNN_SA	1	21.48	17.99	18.24	15.56
CTDNN_SA	15	**20.38**	**17.01**	18.74	**14.96**
CTDNN_SA	30	20.46	17.79	19.49	16.13
CTDNN_SA	60	21.06	17.75	18.82	16.21

TABLE V: WERs (%) using different number of heads in the self-attention layer. TDNN-f is given in the table to provide a comparison with the previous best result on the same dataset

assigns similar weights to context, thus achieving less non-linearity, and reducing the performance. Table V shows the scores of different number of heads in the self-attention layer. These results confirms with the above observation as 15 heads show the best performance. Moreover, the number of trainable parameters increase in the order of millions as the number of heads gets higher (Figure 4).

It is observed that the attention score peaks at the centering bin ($t = 0$) and the weights gradually decrease as getting further from the center. Notice the lower amplitude of the context around the centering bin as the context width gets larger. The attention weights are normalized values due to the softmax layer within the self-attention structure. As the number of context bins increase, the weights assigned to centering bins decrease, while keeping the position of the peaking region in the distribution. The smaller attention weights cause the centering bins have a lesser effect on the final classification. In both of the cases where the context is symmetrical (i.e. the same number of context bins from the past and the future), the decrease of attention scores is slightly sharper for the future bins, implying more attention paid to the past. Finally, Figure 4 shows that the number of trainable parameters increase in the order of 100K for each step we increase the context width. As a conclusion, even though larger context size helps with the training loss (see Figure 3), the WER performance does not improve accordingly as shown in Table VI.

Overall, setting the context to (-15,6) and number of heads to 15 shows the best WER performance which sets the best

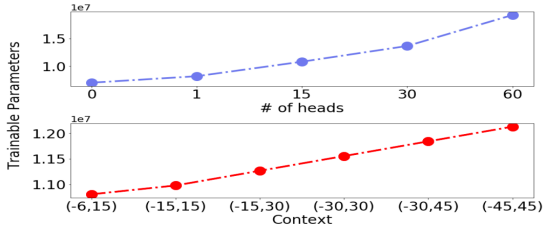


Fig. 3: Number of trainable parameters for each NN model

score reported in this paper.

Context	(-15,6)	(-15,15)	(-30,15)	(-30,30)	(-45,30)	(-45,45)
Dev	18.74	18.24	18.62	18.42	18.37	**17.30**
Test	**14.96**	15.61	15.93	16.26	15.13	15.26

TABLE VI: CTDNN_SA scores with different context widths. The results are reported using 4G RNNLM

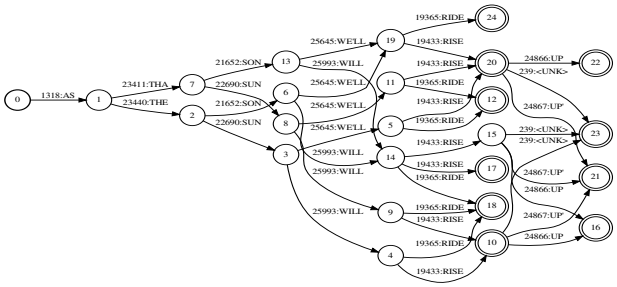
E. Decoding Lattices

Figure 5 illustrates the lattices generated during the decoding stage. The lattices in (a) are obtained using the baseline CTDNN acoustic model with 4G - MaxEnt LM. Adding the self-attention layer to the baseline model results in a much simpler lattice graph (b) reducing the computational complexity and potential confusions during decoding. The graph in (c) is generated by rescoring the lattices using RNNLM where there is only one possible word sequence prediction, boosting the decoding performance even further.

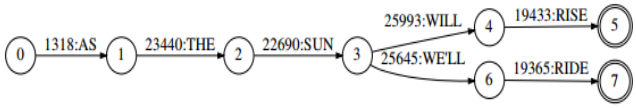
Compared to the state-of-the-art in ASR, there is still room for improvement for achieving a robust lyrics transcription system. One of the potential reasons for the high WERs in the experiments might be due to the cleanness of both the training and test data. By inspection, we have observed that the clean-up process described in Section IV does not perfectly clean the dataset from utterances with bad portions or imperfect annotations. In some utterances, the intelligibility of the sung lyrics is not even sufficient for human listeners. For further improvement, the alterations in word pronunciations have to be modified. The lexicon can be modified as suggested in [5] or learned from data which requires pronunciation annotations in singing recordings. The neural networks can be further tuned and more training data can be acquired through the combination of different datasets.

VII. Conclusion

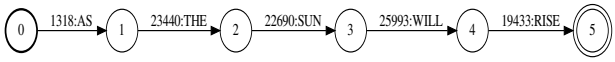
Though achieving an approximately 5% WER improvement to the previously reported best result on the same dataset [6],



(a) CTDNN - 4G MaxEnt



(b) CTDNN_SA - 4G MaxEnt



(c) CTDNN_SA - 4G RNNLM

Fig. 4: Decoding lattices generated with different acoustic and language models for the utterance: '*AS THE SUN WILL RISE*'. The lattice graph in (a) got much simplified in (b) where Self-Attention is used in the neural network. In (c), using RNNLM results in the simplest lattice graph.

.

the performance and the generalizability of the proposed system needs further evaluation on different benchmark datasets. Moreover, an evaluation is needed to compare the WER scores of human listeners in order to obtain a performance measure relative to the human level. Considering the current state of the art in ASR, further improvement is necessary to reach a similar robustness level. For reproducibility, we share our code to reproduce the results reported in this paper[9].

The next steps of this research will focus on handling variances in word pronunciations in singing, building a better language model and curating a larger and cleaner training dataset. Unlike spontaneous speech, sung lyrics tend to be pronounced with longer vowels and voiced phonemes. The lexicon used in the decoding graphs needs to be modified in consideration of these alterations in word pronunciations. This new lexicon can be created either be using heuristics

[9]https://github.com/emirdemirel/AutomaticLyricsTranscription-with-Self-Attention

[5] or by learning a pronunciation dictionary from singing data. Language modeling plays an important role in achieving an improved ALT system as shown in our experiments. To achieve this, one might focus on extending/refining the training data and more advanced methods for building the LM. For the former, a larger dataset of lyrics could be constructed and for the latter style-specific LMs can be trained as suggested in [3]. Finally, the size and variance of the training data for the acoustic model are crucial for performance boost. Training data can be cleaned and new annotations could be generated from weakly or unannotated data using musical heuristics.

The task of ALT for monophonic singing voice is a task far from being solved, yet recent research indicates a promising future to achieve a robust system. Through this study, we hope to draw more attention to ALT from researchers in both ASR and Music Information Retrieval (MIR) communities, potentially bridging the gap between these two research fields.

REFERENCES

[1] Smule Sing! 300x30x2 Dataset, “https://ccrma.stanford.edu/damp/“, accessed October 2019.

[2] A. Kruspe, “Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing.“ International Society for Music Information Retrieval Conference, 2016.

[3] C. Gupta, E. Yılmaz and H. Li. “Automatic lyrics transcription in polyphonic music: Does background music help?“ arXiv preprint arXiv:1909.10200, unpublished.

[4] A. Mesaros and T. Virtanen, "Adaptation of a speech recognizer for singing voice," in European Signal Processing Conference, 2009.

[5] C. Gupta, H. Li, and Y. Wang, “Automatic Pronunciation Evaluation of Singing.“ Interspeech. 2018.

[6] G. R. Dabike and J. Barker. “Automatic lyric transcription from Karaoke vocal tracks: Resources and a Baseline System“ Interspeech, 2019.

[7] A. Mesaros, “Singing voice identification and lyrics transcription for music information retrieval invited paper.“ 2013 7th Conference on Speech Technology and Human-Computer Dialogue, 2013.

[8] Smule, “Digital Archive Mobile Performances (DAMP), ”https: //ccrma.stanford.edu/damp/“, accessed October 2019.

[9] A. Kruspe, “Retrieval of textual song lyrics from sung inputs.“ Interspeech, 2016.

[10] C.-P. Tsai, Y.-L. Tuan and L.-S. Lee, “Transcribing lyrics from commercial song audio: The first step towards singing content processing,” in 43th International Conference on Acoustics, Speech and Signal Processing, 2018.

[11] C. Gupta and H. L. R. Tong, “Semi-supervised lyrics and solo-singing alignment“ International Society for Music Information Retrieval Conference, 2018.

[12] D. Povey, et. al, “The Kaldi speech recognition toolkit", IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

[13] M.Mohri, Fernando Pereira and M Riley, “Weighted finite-state transducers in speech recognition“ Computer Speech & Language 16 No:1, 2002.

[14] V. Peddinti, D. Povey and S. Khudanpur. “A time delay neural network architecture for efficient modeling of long temporal“, International Speech Communication Association, 2015.

[15] D. Povey, H. Hadian, P. Ghahremani1, K. Li and S. Khudanpur, “A time-restricted self-attention layer for ASR“, International Conference on Acoustics, Speech and Signal Processing, 2018.

[16] A. Vaswani, et al. “Attention is all you need“, Advances in neural information processing systems, 2017.

[17] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional Transformers for language understanding”, North American Chapter of the Association for Computational Linguistics, 2019. https://arxiv.org/abs/1810.04805

[18] M. Sperber, J. Niehues, G. Neubig, S. Stüker and A. Waibel, “Self-attentional acoustic models“, Interspeech,2018. arXiv preprint arXiv:1803.09519.

[19] W. Chan, N. Jaitly, Q. Le and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition“, International Conference on Acoustics, Speech and Signal Processing, 2016.

[20] A. Stolcke, “SRILM — An extensible language modeling toolkit”, International Conference on Spoken Language Processing, 2002.

[21] K. Li, H. Xu, Y. Wang, D. Povey and S. Khudanpur “Recurrent neural network language model adaptation for conversational Speech Recognition“, Interspeech, 2018.

[22] M.J.F. Gales "Maximum likelihood linear transformations for HMM-based speech recognition", Computer speech and language 12.2, 1998.

[23] D. Povey, et. al, “Purely sequence-trained neural networks for ASR based on lattice-free MMI”, International Speech Communication Association – Interspeech, 2016.

[24] T. Ko, V. Peddinti, D. Povey and S. Khudanpur, “Audio augmentation for speech recognition”, Interspeech, 2015

[25] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors”, Automatic Speech Recognition and Understanding (ASRU), 2013.

[26] K. Veselý, A. Ghoshal, L. Burget and D. Povey, "Sequence-discriminative training of deep neural networks", Interspeech Vol. 2013, 2013.

[27] G. Cheng, et al. “An exploration of dropout with LSTMs“, Interspeech, 2017.

[28] X. Zhang, Xiaohui, J. Trmal, D. Povey, S. Khudanpur, “Improving deep neural network acoustic models using generalized maxout networks“, International Conference on Acoustics, Speech and Signal Processing, 2014.

[29] H. Xu, et al. "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition", International Conference on Acoustics, Speech and Signal Processing, 2018.