Music Removal by Convolutional Denoising Autoencoder in Speech Recognition

Mengyuan Zhao, Dong Wang*, Zhiyong Zhang, Xuewei Zhang
Center for Speech and Language Technology (CSLT)
Research Institute of Information Technology, Tsinghua University
Tsinghua National Lab for Information Science and Technology
{zhaomy,zhangzy,zxw}@cslt.riit.tsinghua.edu.cn;
wangdong99@mails.tsinghua.edu.cn

***Abstract*—Music embedding often causes significant performance degradation in automatic speech recognition (ASR). This paper proposes a music-removal method based on denoising autoencoder (DAE) that learns and removes music from music-embedded speech signals. Particularly, we focus on convolutional denoising autoencoder (CDAE) that can learn local musical patterns by convolutional feature extraction. Our study shows that the CDAE model can learn patterns of music in different genres and the CDAE-based music removal offers significant performance improvement for ASR. Additionally, we demonstrate that this music-removal approach is largely language independent, which means that a model trained with data in one language can be applied to remove music from speech in another language, and models trained with multilingual data may lead to better performance.**

Index Terms: speech recognition, music removal, noisy training, denoising autoencoder

I. Introduction

Music embedding is often observed in speech recordings. For example in movies and broadcast news, speech signals are often mixed with background music to improve affection of the expression. Unfortunately, mixing music in speech usually causes significant performance reduction in automatic speech recognition(ASR) [1][2].

Some research has been conducted to boost music-embedded ASR and most of the existing methods focus on music/voice separation. The basic idea is to separate music and speech signals according to some human-discovered patterns or properties of music. The music patterns that have been studied include entropy [1], repeating patterns [3], F0 and harmonic structures [4][5][6]. A large body of research is based on low-rank models, e.g., robust PCA [7], non-negative matrix factorization (NMF) [8][9][10] and robust NMF [11]. The spectral sparseness and temporal continuity have been employed as well [12][10].

It should be noticed that most of the above separation-based approaches have not been applied to music-embedded ASR, although performance gains are expected if they were. A potential problem of the separation-based approaches is that the music patterns and properties (F0, harmonic structure, etc.) that these methods rely on are human-designed, which may lead to suboptimal music removal and difficulty in dealing with the complexity of music signals in different genres. For this reason, the effectiveness of these methods for ASR is limited, and the performance on music-removed speech is still far from being satisfactory. This can be seen from the relative small performance improvement reported in [1].

This paper proposes a learning-based approach. Instead of relying on human discovery, our method learns music patterns from data directly. Specifically, we promote to learn from music signals a model that represents music patterns, and use this model to recover clean speech from music-embedded speech. In this study, the denoising autoencoder (DAE) model is selected to conduct the learning. DAE is a special implementation of autoencoder (AE), by introducing random corruptions to the input features in model training. It has been shown that this model is very powerful in learning low-dimensional representations and can be used to recover noise-corrupted input [13]. In [14], DAE is extended to a deep recurrent structure and has been employed to recover clean speech in noisy conditions for ASR. A recent study employs DAE in de-reverberation [15].

In this study, DAE is used to remove the music component from music-embedded speech signals. Particularly, we focus on the convolutional DAE (CDAE), which involves one or several convolutional layers to extract repeating patterns of music in the spectral and temporal domains. Several researchers have studied the repeating patterns of music and employed them to separate music from speech, e.g., [3]. Convolutional networks are powerful in learning repeating patterns, and so CDAE is assumed to be more suitable for music learning compared to DAE.

Besides the capability in learning spectral and temporal repeating patterns, CDAE possesses several other merits: first, it involves a deep neural network (DNN) structure, which allows learning high-level music patterns layer by layer, e.g. the harmonic structures. Second, the high freedom associated with the CDAE structure enables it to learn music with multiple melodies, instruments and genres. Finally, the corruption-injection training (see Section III) allows the model being trained with a small amount of data.

The rest of the paper is organized as follows: Section II discusses some related works, and Section III presents the music-removal DAE and CDAE. The experiments are reported in Section IV and the paper is concluded in Section V.

II. Related work

This work is closely related to various speech/voice separation approaches. Some representative works have been discussed in the previous section. A big advantage of the separation-based approach is that it relies on 'general patterns' of music such as F0, harmonic structures and repeating patterns. These patterns are common for most music and so can be well generalized to new music. This advantage, on the other side, is also a disadvantage since for some music (rap for example), these patterns are not so clear. The DAE approach does not rely on human-discovered patterns but learns the patterns from data, and so the generalizability heavily depends on what music signals have been learned. Thanks to the power of DAE in learning multiple conditions with the deep structure, our method is able to learn music patterns of any melody, any instrument and any genre in a single model, provided that sufficient data of the target music are available.

This work is also closely related to the DAE research and its applications in speech signal enhancement, e.g., the study in noise robustness [14] and the study in de-reverberation [15]. Finally, our work is related to the research on multilingual ASR where the DNN model has been demonstrated to be effective in learning language-independent speech features from multilingual training data [16][17][18].

It also deserves to mention that music pattern learning has been proposed in the separation-based framework as well. For example in [2], a model-based separation approach was proposed where music and speech signals are modelled by two Gaussian mixture models (GMM). The music GMM is trained on the audio prior to the start of speech, and the clean speech is estimated from the two GMMs. The authors showed over 8% relative improvement in word error rate (WER) for a real world voice search ASR system. This improvement is rather marginal, partly attributed to the limited power of GMMs in modeling music patterns. The DAE-based approach proposed in this paper is supposed to be much more powerful in pattern learning, however the disadvantage is that the training requires more data and so can not be adapted online for each utterance as [2] does.

III. Music removal with CDAE

DAE was first proposed to learn robust low-dimensional features [14], and later was extended to recover the original 'clean' signal from a noise-corrupted signal [14][15]. This can be simply employed to remove music from music-corrupted speech signals.

A potential shortage of the DAE-based music removal is that the music patterns are learned 'blindly', which means no prior knowledge of music signals has been utilized. In fact, most of music signals possess unique properties in both the spectral domain and the temporal domain. These properties have been extensively employed in the separation-based approach as discussed already in Section II. In the learning-based approach, these properties should be employed as well.

For example, it is well-known that music signals involve strong harmonic structures, which suggests clear repeating patterns in the spectral domain. These repeating patterns can be learned with shared parameters across the frequency axis. Similar repeating patterns also exist in the temporal domain, and can be learned by shared parameters across the time axis.

We therefore propose to combine the convolutional neural network (CNN) and DAE. The CNN model involves a convolutional layer that is powerful to learn local patterns by shared weights of connections, and a pooling layer which can deal with spectral and temporal variations of the input. These two properties are desirable for learning music signals, and can deal with frequency harmonics and temporal repeating patterns.

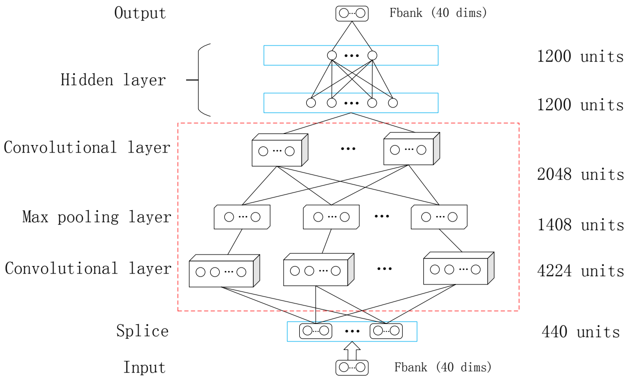

Fig. 1. The network structure of CDAE that involves two convolutional layers and one pooling layer. The LDA transform in DAE is replaced by a CNN structure, as shown in the red-line box.

A CNN-DAE hybrid structure is promoted in this work, as shown in Figure 1, which involves two convolutional layers and a pooling layer in the DAE network. This hybrid structure is denoted by CDAE in this study. We found that CDAE can improve performance of music removal significantly, as will be seen shortly.

A particular concern of the DAE/CDAE training is how to generate the random corruption, i.e., how to select the music segment for each speech signal. We follow the noisy training strategy proposed in [19]. The main steps are as follows: for each clean speech signal, sample a particular music recording following a Dirichlet distribution, and sample the intensity of the corruption in terms of signal to noise ratio (SNR) following a Gaussian distribution. The selected music recording is then amplified to meet the required SNR, and mixed with the clean speech signal to generate a training sample.

IV. Experiments and Results

This section presents the experiments. We first describe the databases and the baseline system, and then present the results with the DAE-based and CDAE-based music removal. The potential of this approach in the multilingual scenario will be finally demonstrated.

A. Data and baseline

The Aurora4 database will be used in most of the experiments in this section. It is an English database and involves about 18 hours of speech in total. Following the standard setup, the training set and the cross-validation (CV) set consist of 15.1 hours (7138 utterances) and 2.2 hours (1206 utterances) of speech signals, respectively. The test set consists of 0.65 hours of speech (324 utterances).

Another database used in the experiments is a subset of the 863 database, which is in Chinese and contains about 50 hours of speech in total. The training set consists of 43 hours of speech signals (33279 utterances), and the CV set consists of 2.5 hours of speech (2080 utterances). Another 5 hours of speech signals (4160 utterances) are used as the test set.

The music repository involves 4 music signals, which are 'Pa' (Piano, Betthoven Moonlight Sonata Chapter 3), 'Vi' (Violin, Theme from Schindler's List), 'Sy'(Symphony, Radetzky March), and 'Ra' (Rap, Nunchaku Jay Chow). These music signals are mixed with the training and test data to examine the capability of DAE/CDAE in learning 'known' music. Additionally, we select another piano signal 'Pa2' (Chopin Nocturne No.2 Op.9) to examine the performance of the model in learning 'out-of-repository' music.

The baseline ASR system is based on the DNN-HMM hybrid architecture [20]. The acoustic feature is the 40-dimensional Fbanks. For each frame, the central frame is concatenated with the left and right 5 frames, forming a 440-dimensional feature vector, and then an LDA transform is employed to reduce the feature dimension to 200.

The acoustic model is a DNN that contains 4 hidden layers, each involving 1200 units. For the English model, the number of output units is 3356, corresponding to the number of context-dependent states. For the Chinese model, the number of output units is 3361. The Kaldi toolkit[1] is used to train the system, and the training process largely follows the WSJ S5 GPU recipe, with the training criterion set to cross entropy.

B. DAE-based music removal

The input feature of the DAE is totally the same as the input of the DNN acoustic model, i.e., 200-dimensional LDA features that are derived from 11-frame-concatenated Fbanks. The output is simply the 40-dimensional Fbank feature corresponding to the central frame of the input. The training process follows a similar procedure as the DNN training, except that the training criterion is set to the mean square error. It should be noted that the DAE output is the music-removed Fbank feature, and therefore can be simply fed into the DNN acoustic model. In this sense, the DAE-based music removal can be regarded as a special pre-processing and can be easily integrated in the pipe-line of acoustic feature extraction.

In this experiment, the four music signals are mixed with the training data following the strategy described in Section III, where the SNR of the corruption is sampled from a Gaussian distribution $N(5, 10)$. Meanwhile, the music signals are mixed with the test data at a fixed SNR, which we set to 5 dB. These configurations are chosen according to [19]. Additional, 'Pa2' is used as the 'out-of-repository' music and is mixed with the test data only.

Three scenarios are tested. In the first scenario, only one music signal is used to corrupt the data in DAE training; in the second scenario (mixA), all the 4 music signals are involved; in the third scenario (mixB), the training is the same as in mixA, but the special type 'no-music' is involved.

We conduct the experiment with the Auroa4 database, and the performance in terms WER is reported in Table I, where each row represents a training condition and each column represents a test condition.

From the baseline results, it can be seen that music embedding seriously degrades the ASR performance. For example, with the piano music embedded, the WER increases from 6% to 49%. When the DAE-based music removal is applied, significant performance improvement is obtained in all the tested scenarios.

TABLE I
WER WITH DAE-BASED MUSIC REMOVAL

	WER%					
	$Clean$	Pa	Vi	Sy	Ra	$Pa2$
$Baseline$	5.98	49.41	59.31	57.22	54.61	45.60
Pa	6.70	11.10	26.64	30.20	52.36	12.13
Vi	6.70	21.42	11.39	31.25	48.08	16.58
Sy	7.03	24.43	27.36	15.46	50.04	21.84
Ra	6.66	29.00	29.09	33.61	13.73	25.86
$MixA$	6.85	12.89	13.59	18.24	18.05	12.72
$MixB$	6.49	13.77	14.03	18.49	18.41	12.89

Interestingly, involving a particular music in the DAE training improves performance on speech embedded by other music, particularly music of the same type. This can be clearly seen from the row 'Pa' in Table I, where the DAE trained with music Pa improves test with all other music embedding, especially Pa2 that is played by piano as well. This observation suggests that different types of music share some common properties and these properties can be learned by DAE.

Finally, the results in the row mixA and mixB suggest that DAE can learn multiple music signals in a single model, especially when some training data remain uncorrupted (mixB). This is highly interesting since it means that a general music removal DAE is possible by training with multiple music signals. This general model can deal with any music, provided that some music signals of the same type have been involved in the DAE training. For instance, the result on Pa2 has demonstrated the performance of the general model on out-of-repository (new) music.

C. CDAE-based music removal

The second experiment tests the CDAE model-based music-removal. As shown in Fig. 1, the network involves two convolutional layers and one pooling layer. The two convolutional layers consist of 4224 and 2048 units respectively, and the pooling layer consists of 1408 units.

The experiment is conducted on the Aurora4 database, and the configurations are all the same as in the DAE experiment. The results have been presented in Table II. Compared to the results with DAE in Table I, consistent performance improvement is observed with the CDAE-based music removal. This confirms our conjecture that CNN can better learn the music-specific patterns and variations, and so it is more powerful for music removal.

TABLE II
WER WITH CDAE-BASED MUSIC REMOVAL

	WER%					
	$Clean$	Pa	Vi	Sy	Ra	$Pa2$
$Baseline$	5.98	49.41	59.31	57.22	54.61	45.60
Pa	6.72	9.35	23.80	30.75	47.68	10.03
Vi	6.66	21.65	9.90	31.70	46.29	15.25
Sy	6.76	24.50	25.15	12.22	49.58	19.46
Ra	6.36	25.27	26.66	32.04	11.21	19.54
$MixA$	6.51	11.08	11.16	15.23	14.19	10.49
$MixB$	6.30	10.43	11.16	15.67	14.30	10.30

[1] http://kaldi.sourceforge.net/about.html

D. Music removal across languages

Music is assumed to be language-independent, which suggests that the DAE/CDAE music-removal model can be trained and applied across languages. To test this conjecture, we learn two monolingual CDAE models based on the Aurora4 database (in English) and the 863 database (in Chinese) respectively, where only the piano music 'Pa' is embedded. Each of the two models is tested on *both* the Aurora4 task and the 863 task, where the test utterances are corrupted by the two piano music signals 'Pa' and 'Pa2'. In another experiment, a multilingual DAE is trained by pooling the data of the two databases. Again, the resulting multilingual model is tested on the two databases respectively.

TABLE III
WER WITH MULTILINGUAL CDAE

	Aurora4 (WER%)			863 (CER%)		
	$clean$	Pa	$Pa2$	$clean$	Pa	$Pa2$
$Baseline$	5.98	49.41	45.60	15.93	67.36	61.69
$Auraro4$	6.72	9.35	10.03	19.84	33.67	32.35
863	7.54	12.66	15.44	17.74	23.78	24.31
$Auraro4 + 863$	6.68	8.76	10.97	17.49	23.33	24.32

The results on the Auroa4 database and the 863 database are presented in Table III. Note that the results on the 863 database are reported in Chinese character error rate (CER). The results on the two databases show the same trend, that cross-lingual application of the music-removal DAE is possible, although the performance with the cross-lingual application is still worse than that with the monolingual application. We tend to believe that this performance gap is more likely caused by the different acoustic conditions of the two database, rather than by languages.

It is also interesting to see that the multilingual CDAE delivers rather good performance on both the two databases, and the performance is even better than that with the monolingual CDAEs in some cases. Combined with the findings in the previous section, this suggests that a general music-removal CDAE is possible by training the model with multilingual speech data embedded with multiple music.

V. CONCLUSIONS

This paper presented a new music-removal approach based on CDAE. The experimental results on ASR tasks demonstrated that CDAE can learn music patterns and remove them from music-embedded speech signals, and the CDAE model outperforms the DAE model. We also found that the CDAE model can be applied across languages, and a general music-removal CDAE is possible by learning with multilingual data embedded with multiple music. The future work will investigate more complex music types, and study the multiple music embedding which involves several music signals in the same speech segment.

VI. ACKNOWLEDGEMENTS

This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No.20130002120011. It was also supported by Sinovoice and Huilan Ltd.

REFERENCES

[1] P. Vanroose, "Blind source separation of speech and background music for improved speech recognition," in *The 24th Symposium on Information Theory*, 2003, pp. 103–108.

[2] T. Hughes and T. Kristjansson, "Music models for music-speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4917–4920.

[3] Z. Rafii and B. Pardo, "Repeating pattern extraction technique (repet): A simple method for music/voice separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 73–84, 2013.

[4] M. Ryynanen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 1417–1420.

[5] J.-L. Durrieu and J.-P. Thiran, "Musical audio source separation based on user-selected f0 track," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 438–445.

[6] Y.-G. Zhang and C.-S. Zhang, "Separation of music signals by harmonic structure modeling," in *Advances in Neural Information Processing Systems*, 2005, pp. 1617–1624.

[7] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 57–60.

[8] B. Zhu, W. Li, R. Li, and X. Xue, "Multi-stage non-negative matrix factorization for monaural singing voice separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 10, pp. 2096–2107, 2013.

[9] I.-Y. Jeong and K. Lee, "Vocal separation using extended robust principal component analysis with schatten p/lp-norm and scale compression," in *2014 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING*. IEEE, 2014.

[10] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 1, pp. 228–237, 2014.

[11] P. Sprechmann, A. M. Bronstein, and G. Sapiro, "Real-time online singing voice separation from monaural recordings using robust low-rank modeling." in *ISMIR 2012*, 2012.

[12] T. Ming, X. Xiang, and J. Yishan, "Nmf based speech and music separation in monaural speech recordings with sparseness and temporal continuity constraints," in *3rd International Conference on Multimedia Technology (ICMT-13)*. Atlantis Press, 2013.

[13] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[14] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr." in *INTERSPEECH*. Citeseer, 2012.

[15] T. Ishii, H. Komiyama, T. Shinozaki, Y. Horiuchi, and S. Kuroiwa, "Reverberant speech recognition based on denoising autoencoder." in *INTERSPEECH*, 2013, pp. 3512–3516.

[16] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[17] N. T. Vu, F. Metze, and T. Schultz, "Multilingual bottle-neck features and its application for under-resourced languages." in *SLTU*, 2012, pp. 90–93.

[18] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual mlp features for low-resource lvcsr systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4269–4272.

[19] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng, and Y. Li, "Noisy training for deep neural networks in speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 2, 2015.

[20] L. Deng and D. Yu, *DEEP LEARNING: Methods and Applications*. NOW Publishers, January 2014.