Singing Voice Detection via Similarity-based Semi-supervised Learning

Xi Chen
chenxi19@fudan.edu.cn
School of Computer Science and Technology
Fudan University
Shanghai, China

Yongwei Gao
ywgao@suibe.edu.cn
School of Statistics and Information
Shanghai University of International Business and Economics
Shanghai, China

Wei Li*
weili-fudan@fudan.edu.cn
School of Computer Science and Technology
Shanghai Key Laboratory of Intelligent Information Processing
Fudan University
Shanghai, China

Abstract

Data-driven methods play an important role in Singing Voice Detection (SVD). However, datasets with precise annotations are scarce. In this paper, we propose an SVD method via similarity-based semi-supervised learning (SSSL_SVD). For one thing, we propose to enrich the diversity of training data using the self-training semi-supervised method (SSL). In SSL, pseudo labels of the unlabeled data are first generated by a pre-trained teacher model and are then used to train a student model. For another thing, we propose to measure the audio frame from a similarity-based perspective. Taking it into consideration, we could provide more appropriate learning targets. Finally, experiment results indicate that the proposed method achieved comparable results with state-of-the-art (SOTA) algorithms.

CCS Concepts: • **Information systems** → *Clustering and classification.*

Keywords: Singing Voice Detection, Voice Detection, Semi-supervised Learning, Music Information Retrieval

ACM Reference Format:
Xi Chen, Yongwei Gao, and Wei Li. 2022. Singing Voice Detection via Similarity-based Semi-supervised Learning. In *ACM Multimedia Asia (MMAsia '22), December 13–16, 2022, Tokyo, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3551626.3564963

*Corresponding author

This work was supported in part by the National Key R&D Program of China(2019YFC1711800), NSFC(62171138).

MMAsia '22, December 13–16, 2022, Tokyo, Japan

ACM ISBN 978-1-4503-9478-9/22/12…$15.00
https://doi.org/10.1145/3551626.3564963

1 Introduction

Singing voice detection (SVD) aims to detect singing voice containing segments in a polyphonic music signal. It is an important task in music information retrieval (MIR) [6] with many potential applications, such as singing voice separation [14], lyric transcription [9] and query by humming [3]. Many deep learning based SVD methods have been proposed and these methods generally require a large well-labeled dataset. However, the high temporal accuracy annotation required is time-consuming and skilled annotators demanding [2, 4, 11]. Therefore, how to efficiently mine the information hidden behind labeled and unlabeled data while saving human labor has become a hot research point in the field.

A series of approaches have been applied to alleviate this issue. Pitifully, they have their own limitations. The weakly supervised learning method in [11] and the active learning method in [8] could help to enrich the dataset with a small amount of human labor. However, additional human labor could be substantially saved but not entirely eliminated. Label regenerating method [4, 10] could help to enlarge the dataset without additional human labor. However, designing a label regenerating mechanism is experience-demanding and the corresponding prerequisites are always too harsh to meet, with either instrumental version recordings [4] or draft time-aligned lyrics and notes [10] needed. Transferring knowledge from an artificial speech-plus-music dataset could help to explore the timbre diversity with fewer prerequisites [2] and no additional human labor. However, the specific articulation mode in singing and the synchronized nature between singing and accompaniment, which make SVD a challenging task, could not be provided by the artificial speech-plus-music dataset.

We propose to use the self-training semi-supervised learning method (SSL) to alleviate the data-scarce issue, which has been applied and proven powerful in many other fields [15, 16, 18]. First, compared with [8, 11], it can eliminate additional human-labor entirely. Next, compared with [4, 10], it is prerequisite-free. Last, compared with [2], it mines information from real-world unlabeled data without missing

information about singing articulation and the synchronized nature between singing and accompaniment.

Besides enlarging the training data using SSL, we propose to measure the audio frame from the similarity-based perspective which can provide a more appropriate training target. The goal and learning target in SVD is simple, i.e. whether the given audio frame is 'singing', which means the given audio frame contains the singing voice, or 'non-singing', which means otherwise. However, sound sources and their combinations in polyphonic music are diverse. For example, a non-singing audio frame which contains time-continuous and pitch-fluctuating sounds, especially that from the electronic guitar, performs more singing than drums [7]; a singing audio frame with a low singing-to-accompaniment gain ratio (SAR) performs more non-singing. This similarity-based measurement is of high entropy, provides more timber-related information, and facilitates decision-making. In our work, we propose introducing a scaling parameter from the Knowledge Distillation (KD) [1] to help similarity measuring.

In summary, the contribution of this paper is three-fold: 1). We propose to use the self-training semi-supervised learning method. 2). We propose to measure the audio frame from a similarity-based perspective. 3). The experimental study demonstrates the effectiveness of our method, our method is comparable with SOTA ones.

2 Methodology

In this section, we will present the baseline model in Section 2.1, the similarity-based measurement in Section 2.2, the self-training semi-supervised learning method in 2.3, and the proposed in Section 2.4

2.1 The Baseline Model

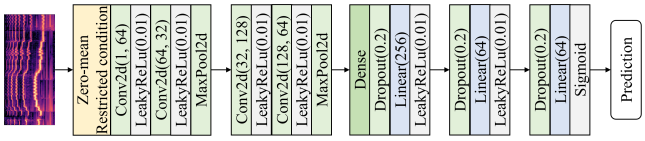

Figure 1. The baseline model, which includes a CNN architecture and the sound level invariant restricted condition.

As shown in Fig 1, we employ Schlüter's convolutional neural network (CNN) model [12] and the sound level invariant restricted condition [13] as the baseline model.

Since Schlüter's CNN has achieved SOTA performance, we employ it as one part of the baseline model. It is composed of 2 convolutional blocks and 3 fully connected (FC) blocks. For the convolutional block, there are 2 convolutional layers followed by a LeakyReLu activation function and a max-pooling layer; For the FC block, it is made up of a dropout layer, a linear layer, and an activation function with leakyReLu in the first two FC blocks and Sigmoid in the last FC block.

Since the log-melspectrogram, the representation we used, is sensitive to sound level which has no correlation with the existence of the singing voice, we employ the lightweight and parameter-free sound level invariant restricted condition as the other part of the baseline model. This sensitivity of the log-melspectrogram to loudness is that changing the sound level equals adding a constant number c to the log-melspectrogram X. Denoting W as the weight matrix for a 2D convolutional kernel, the restricted condition could be formulated as

$$\sum_{i,j} W_{i,j} = 0. \tag{1}$$

Using it on the first convolutional layer, Schlüter's CNN performs exactly the same under different sound levels

$$(X + c) \otimes W = X \otimes W + c \otimes W = X \otimes W, \tag{2}$$

where $\otimes$ denotes the convolution operation.

2.2 The Similarity-based Measurement

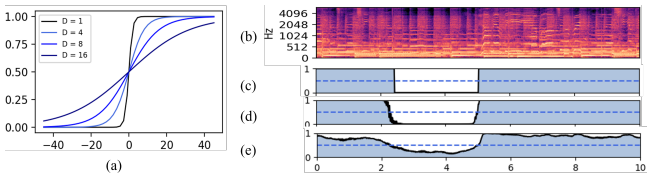

Figure 2. (a). The similarity-based measurement under different scaling parameters. (b). a 10s log-melspectrogram; (c). ground truth hard label for (b); (c). a model's probability prediction of the existence of singing voice for (b); (d). the similarity level of being 'singing' for (b) which is calculated using the scaling parameter in (a) and a model's probability prediction in (c).

We employ another model's probability prediction and a scaling parameter for the similarity-based measurement which measures an audio frame's similarity level of being 'singing'. As described in [1], the model's probability output could serve as an approximate similarity estimation. However, it is often too subtle to concentrate in the model training phase, for the sigmoid function is designed to be unit-step-like to serve as a differentiable substitute, which is steep where output is around 0.5 and nearly flat where output is around 0 and 1. Therefore, we employ the scaling parameter [1]. It could make the probability output softer, in other words, drawing the probability output closer to 0.5 which will be more suitable for the similarity-based measurement.

The similarity-based, i.e. the softened sigmoid function, is formulated as:

$$Sigmoid(x, S) = \frac{1}{1 + e^{-\frac{x}{S}}}, \tag{3}$$

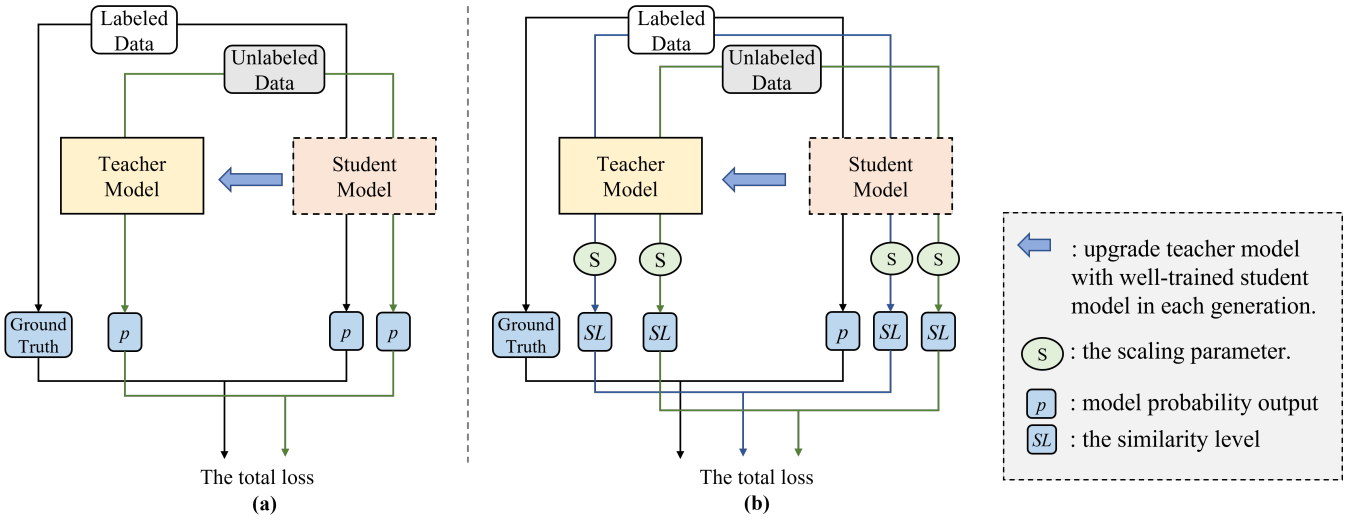

Figure 3. (a). the self-training semi-supervised learning (SSL) method; (b). the similarity-based semi-supervised learning (SSSL_SVD) method.

where x is a model's probability prediction. As shown in Figure 2(a), a larger S makes the sigmoid function softer and the similarity level more obvious; When S is equal to 1, the sigmoid function is identical to the original one. Therefore, by setting S to 1, we could obtain the probability output; while setting S to other values, we could obtain the similarity level without changing model parameters. Figure 2 shows an audio segment in (b), the ground truth label in (c), another model's probability prediction in (d), and the similarity level in (e). What's more, it has to mention that the gradient magnitude produced under the scaling parameter will be scaled to $1/S^2$ of the original, multiplying S^2 to the loss function is needed while training [1].

2.3 The Self-training Semi-supervised Learning Method

Algorithm 1 The framework of SSL method.

Require: D: the well-labeled dataset.
Require: U: the unlabeled dataset.
Require: T_i/S_i: the teacher/student model in the i-th generation.
Train the first-generation model using D, which will serve as T_2 .
do
T_i generates pseudo labels for S_i to learn.
$S_i \leftarrow D$, calculate loss L_D.
$S_i \leftarrow U_i$, calculate loss L_U.
Tune parameters in S_i using L_t (a weighted average of L_D and L_U).
Upgrade T_{i+1} with S_i.
while S_i outperforms T_i.

Two agents, the teacher model, and the student model make up the SSL method. As shown in Figure 3 (a), the teacher model generates pseudo labels for the student model to learn, then the student model serves as a next-generation teacher model and this process is iterated to obtain a model with better generalization power.

The detailed learning procedure is shown in Algorithm 1. In the first generation, a model is trained on the well-labeled data D with the cross-entropy loss. It will later serve as the next generation teacher model. In the next generations, T_i makes predictions on U_i, and the prediction thus serves as the pseudo soft label for S_i to learn. Denoting x, y as the audio sample and its corresponding ground truth label in D, x_u as the sample in U_i, $T_i(\cdot)$ and $S_i(\cdot)$ as model's probability output, λ_u as the weight on U_i, $CE(\cdot)$ as the cross-entropy loss function, the total loss L_t for S_i is defined as:

$$L_t = CE(y, S_i(x)) + \lambda_u * CE(T_i(x_u), S_i(x_u)). \quad (4)$$

When S_i is well-trained, it serves as T_{i+1} for the next generation. Finally, the training iteration stops when S_i meets its performance ceiling.

2.4 The Similarity-based Semi-supervised Learning Method

Two agents and the similarity-based measurement make up the SSSL_SVD method. As shown in Figure 3 (b), compared with SSL, the main difference is that the teacher model provides an additional similarity level instead of merely the probability output for the student model to learn.

The detailed difference in learning procedure is as follows. In the iterated generations, with the help of the scaling parameter, S, described in Section 2.2, T_i generates both the pseudo label and the similarity level for S_i to learn. Tuning

S to be 1, T_i and S_i generate probability output $T_i(\cdot, 1)$ and $S_i(\cdot, 1)$; Tuning it to be a constant number S, T_i and S_i generate the similarity level $T_i(\cdot, S)$ and $S_i(\cdot, S)$. As a result, on the labeled dataset D, the loss function of the student model L_D^s is defined as:

$$L_D^s = CE(y, S_i(x, 1)) + \lambda_s * S * S * CE(T_i(x, S), S_i(x, S)), \quad (5)$$

on the unlabeled dataset U_i, the loss function L_U^s is defined as:

$$L_U^s = S * S * CE(T_i(x_u, S), S_i(x_u, S)). \quad (6)$$

The total loss L_t^s is defined as:

$$L_t^s = L_D^s + \lambda_u * L_U^s. \quad (7)$$

In addition, it could be found that, when setting λ_s and S to be 0 and 1 separately, this SSSL_SVD method is identical to the SSL method in Section 2.3.

3 Experiments and Results

In this section, we will present the datasets and metrics in Section 3.1, the ablation study results in Section 3.2, and comparison results with previous methods in 3.3.

3.1 Datasets and Metrics

For the well-labeled data, we use Jamendo Corpus. For the unlabeled data, in order to ensure data diversity, we collected songs mainly from world-class competitions. In total, we take advantage of around 11 hours of well-labeled data and improve the model's generalization power using around 135 hours of unlabeled real-world data.

For audio signal processing, we follow the settings in [12]. First, we downsample the audio signal to 22050 Hz and perform a Short Time Fourier Transform (STFT) with the frame length of 1024 samples and the hop length of 315 samples. Further, we transform the spectrogram to the mel scale with 80 mel filters and we logarithmize the magnitudes to get the log-melspectrogram representation. Finally, for each audio frame, we represent it with its surrounding 115 frames.

For performance measurement, we use four typical metrics for general classification tasks, namely, Accuracy (Acccc), Precision, Recall, and F1 score (F1).

To foster reproducibility and comparison of the system, we make the code public in https://github.com/OzymandiasChen/SSSL_SVD.

3.2 Ablation Study

In this section, to validate each component in the proposed, we tested 4 models' performance. They are the baseline, the model trained using the SSL method (shorten as SSL_SVD), the model trained with the similarity level (shorten as similarty_SVD), and the proposed, i.e. SSSL_SVD. Regrettably, it should be noticed that all semi-supervised models used in the ablation part are ones retrained directly by the first-generation models and they only take into consideration of approximately 100 pieces of unlabeled music. The reason is that the training of semi-supervised models algorithms, which needs generations of iterations, is time-consuming and we merely want to concentrate on the degree of improvement under the same condition, rather than touching the method's performance ceiling.

Table 1. Ablation study results on Jamendo Corpus.

Method	Acc	F1	Precision	Recall
The baseline	0.897	0.891	0.88	0.903
SSL_SVD	0.894	0.890	0.859	**0.923**
similarity_SVD	0.899	0.893	0.882	0.904
SSSL_SVD	**0.901**	**0.895**	**0.881**	0.909

The ablation study results will be seen in Table 1. First, comparing the baseline and the similarty_SVD, the similarty_SVD surpasses the baseline on all 4 metrics, which demonstrates the effectiveness of measuring the audio frame from the similarity-based perspective. Second, though SSL_SVD performs slightly badly on Acc, F1, and Precision, it performs best on Recall among all four models. Third, compared with the similarty_SVD and the SSL_SVD, SSSL_SVD has a further promotion on Acc, F1, and Precision.

In the last, we could conclude that the self-training semi-supervised method could help the model generalize well to the large-scale real-world data; And the similarity-based measurement could provide a more proper learning target for the model; Combining the SSL method and the similarity-based measurement, SSS_SVD has a further improvement.

3.3 Comparisons with Previous Methods

Table 2. Comparison results on Jamendo Corpus

Method	Acc	F1	Precision	Recall
Leglaive's [5]	0.915	0.910	0.895	0.926
Schlüter's [12]	0.923	—-	—-	0.903
Schlüter's [11]	0.901	—-	—-	—-
Humphrey's [4]	0.878	—-	—-	—-
Lehner's [6]	0.881	—-	—-	—-
Meseguer-Brocal's [10]	0.860	—-	—-	—-
Zhang's [17]	0.924	0.927	0.926	0.924
The baseline	0.897	0.891	0.880	0.903
SSSL_SVD	**0.924**	0.918	0.924	0.911

We provide comparison results with other methods, especially those dedicated to overcome the data-scarce problem. Comparison results will be seen in Table 2. For one thing, our method could outperform those dedicated to overcome the data-scarce problem [4, 10, 11]. For another thing, our method is comparable with the SOTA one, Zhang's. We attribute the inferior performance to the noise in pseudo labels which are generated by the imperfect teacher model. In the

future, We would like to design some pseudo-label selection mechanism which can filter out the unreliable ones to ensure reliability.

4 Conclusion

In this paper, we propose to utilize the self-training semi-supervised method to improve the model's generalization power. What's more, we propose to measure the audio frame from the similarity-based perspective which is more appropriate for the learning target. Finally, the proposed could be on par with the SOTA on Jamendo Corpus. For future work, we would like to design some pseudo-label selection mechanisms to filter out the noise labels and further improve model performance.

References

[1] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the Knowledge in a Neural Network. *CoRR* abs/1503.02531 (2015).

[2] Yuanbo Hou, Frank K. Soong, Jian Luan, and Shengchen Li. 2020. Transfer Learning for Improving Singing-Voice Detection in Polyphonic Instrumental Music. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, 25-29 October 2020*, Helen Meng, Bo Xu, and Thomas Fang Zheng (Eds.). ISCA, Shanghai, China, 1236–1240.

[3] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang, and Ke Hu. 2012. A tandem algorithm for singing pitch extraction and voice separation from music accompaniment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 20, 5 (2012), 1482–1491.

[4] Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Andreas Jansson, and Tristan Jehan. 2017. Mining Labeled Data from Web-Scale Collections for Vocal Activity Detection in Music. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, October 23-27, 2017*, Sally Jo Cunningham, Zhiyao Duan, Xiao Hu, and Douglas Turnbull (Eds.). International Society for Music Information Retrieval, Suzhou, China, 709–715.

[5] Simon Leglaive, Romain Hennequin, and Roland Badeau. 2015. Singing voice detection with deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, April 19-24, 2015*. IEEE, South Brisbane, Queensland, Australia, 121–125.

[6] Bernhard Lehner, Jan Schlüter, and Gerhard Widmer. 2018. Online, loudness-invariant vocal detection in mixed music signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 8 (2018), 1369–1380.

[7] Bernhard Lehner, Gerhard Widmer, and Reinhard Sonnleitner. 2014. On the reduction of false positives in singing voice detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, May 4-9, 2014*. IEEE, Florence, Italy, 7480–7484.

[8] Wei Li, Xiangyi Feng, and Min Xue. 2016. Reducing manual labeling in singing voice detection: An active learning approach. In *IEEE International Conference on Multimedia and Expo, ICME 2016, July 11-15, 2016*. IEEE Computer Society, Seattle, WA, USA, 1–5.

[9] Matt McVicar, Daniel P. W. Ellis, and Masataka Goto. 2014. Leveraging repetition for improved automatic lyric transcription in popular music. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, May 4-9, 2014*. IEEE, Florence, Italy, 3117–3121.

[10] Gabriel Meseguer-Brocal, Alice Cohen-Hadria, and Geoffroy Peeters. 2018. DALI: A Large Dataset of Synchronized Audio, Lyrics and notes, Automatically Created using Teacher-student Machine Learning Paradigm.. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, September 23-27, 2018*, Emilia Gómez, Xiao Hu, Eric Humphrey, and Emmanouil Benetos (Eds.). Proceedings of the 19th International Society for Music Information Retrieval Conference, Paris, France, 431–437.

[11] Jan Schlüter. 2016. Learning to Pinpoint Singing Voice from Weakly Labeled Examples. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, August 7-11, 2016*, Michael I. Mandel, Johanna Devaney, Douglas Turnbull, and George Tzanetakis (Eds.). International Society for Music Information Retrieval, New York City, United States, 44–50.

[12] Jan Schlüter and Thomas Grill. 2015. Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, October, 26-30, 2015*, Meinard Müller and Frans Wiering (Eds.). International Society for Music Information Retrieval, Málaga, Spain, 121–126.

[13] Jan Schlüter and Bernhard Lehner. 2018. Zero-Mean Convolutions for Level-Invariant Singing Voice Detection. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, September 23-27, 2018*, Emilia Gómez, Xiao Hu, Eric Humphrey, and Emmanouil Benetos (Eds.). International Society for Music Information Retrieval, Paris, France, 321–326.

[14] Daniel Stoller, Sebastian Ewert, and Simon Dixon. 2018. Jointly Detecting and Separating Singing Voice: A Multi-Task Approach. In *Latent Variable Analysis and Signal Separation - 14th International Conference, LVA/ICA 2018, July 2-5, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 10891)*, Yannick Deville, Sharon Gannot, Russell Mason, Mark D. Plumbley, and Dominic Ward (Eds.). Springer, Guildford, UK, 329–339.

[15] Hui Tang and Kui Jia. 2022. Towards Discovering the Effectiveness of Moderately Confident Samples for Semi-Supervised Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montréal, Québec, 14658–14667.

[16] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 993–1003.

[17] Xulong Zhang, Yi Yu, Yongwei Gao, Xi Chen, and Wei Li. 2020. Research on Singing Voice Detection Based on a Long-Term Recurrent Convolutional Network with Vocal Separation and Temporal Smoothing. *Electronics* 9, 9 (2020), 1458.

[18] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al. 2022. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing* (2022).