INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Non-verbal Vocalisation and Laughter Detection using Sequence-to-sequence Models and Multi-label Training

Scott Condron, Georgia Clarke, Anita Klementiev, Daniela Morse-Kopp, Jack Parry and Dimitri Palaz

Speech Graphics Ltd, Edinburgh, United Kingdom

{scott.condron, g.clarke, anita.klementiev, daniela, jack.parry, dpalaz}@speech-graphics.com

Abstract

Non-verbal vocalisations (NVVs) such as laughter are an important part of communication in social interactions and carry important information about a speaker's state or intention. There remains no clear definition of NVVs and there is no clearly defined protocol for transcribing or detecting NVVs. As such, the standard approach has been to focus on detecting a single NVV such as laughter and map all other NVVs to an "other" class. In this paper we hypothesise that for this task such an approach hurts performance, and that giving more information by using more classes is beneficial. To address this, we present studies using sequence-to-sequence deep neural networks where we include multiple NVV classes rather than mapping them to "other" and allow more than one label per sample. We show that this approach yields better performance than the standard approach on NVV detection. We also evaluate the same model on laughter detection using frame-based and utterance-based metrics and show that the proposed approach yields state-of-the-art performance on the ICSI corpus.

Index Terms: computational paralinguistics, non-verbal vocalisation detection, laughter detection, multi-label training.

1. Introduction

As well as speech, human conversation comprises non-verbal vocalisations such as laughter, fillers, sneezes, and breaths. NVVs carry information about a speaker's physiological state, emotional state or even their intentions, and can be characterised by any combination of vocalisation, facial expression or body movement. While we acknowledge that expressions and body movements are an important part of human communication, our focus in this paper is exclusively on vocalisations.

A system to detect NVVs has a wide range of potential applications, for example, in security [1] and healthcare [2], including recent work pertaining to COVID-19 detection and care [3]. NVVs are commonly present in spontaneous conversational speech as opposed to perfectly read speech [4]. For this reason, NVV detection is an important factor in tasks dealing with spontaneous speech, such as in automatic speech recognition (ASR) systems to reduce the number of errors [5], and emotion recognition systems where it can be used as an additional feature [6].

There remains no clear definition of NVVs, nor are there standard protocols for transcription and annotation [4]. Thus one of the major challenges of NVV detection is that it is almost impossible to establish an exhaustive list covering all possible vocalisations. However, to use supervised learning, classifiers such as neural networks need a list of classes in order to be trained so an approach to incorporate this data into the training set must be selected. The standard approach is to aggregate all non-relevant classes into a single class [7]. For example, in the case of laughter detection, the "laughter" class comprises all segments labelled as "laughter", while the "non-laughter" class comprises all other segments, which results in a very heterogeneous collection of sounds, such as speech, sneezes and whistles.

We hypothesise that this approach can hurt the training of the model. Our intuition is that the model learns each class by extracting relevant patterns from the input: when learning the "other" class, the model will see lots of very different patterns which may cause confusion. We argue that using all available information instead of collapsing classes will improve performance.

In this paper, we present a study on the ICSI meeting corpus [8] where we use all available classes during model training and evaluate on both binary ("laughter"/"non-laughter") and multi-class (NVV) tasks. More specifically, we investigate two training approaches: *multi-class*, where the model is trained on an set of classes using the standard cross-entropy loss function, and *multi-label*, which consists of using a loss function composed of a sum of binary classifiers, allowing the model to output more than one label per frame. We also investigate using a sequence-to-sequence (Seq2Seq) model, where the input is a sequence of feature frames and the output is a sequence of labels. Lastly, we propose to use utterance-based metrics borrowed from the sound event detection literature [9] to measure boundary placement correctness as well as the rate of insertions, deletions and substitutions rather than evaluating using frame-based metrics only.

Firstly, for NVV detection we show that the multi-class and multi-label approaches improve performance compared to a frame-based baseline. We then present a study on laughter detection, where we compare to the literature and a model trained using the standard "laughter"/"non-laughter" approach and show that the proposed approaches yield state-of-the-art performance. In both tasks, multi-label approach outperforms the multi-class approach. To summarise, the contributions of this paper are as follows:

- We show that using multi-class and multi-label training improve performance for laughter and NVV detection.
- We achieve state-of-the-art performance for laughter detection on the ICSI meeting corpus using Seq2Seq models.
- We introduce a new baseline for NVV detection.

The remainder of the paper is organised as follows: a literature review is first presented in Section 2; the methodology is then described in Section 3. Section 4 presents the experimental setup; Section 5 discusses the results of the studies; and Section 6 concludes the paper.

http://dx.doi.org/10.21437/Interspeech.2021-1159

2. Related Work

2.1. Non-verbal vocalisation detection

Previous works on NVV detection usually focus on one NVV and frame it as a binary classification problem. This has shown promising results for breath [10, 11]; cough [12] and scream [13] detection. As laughter is the most frequently annotated NVV class [4], laughter detection comprises a large part of NVV detection research. Early work in this field concentrated on detecting the presence of laughter in fixed segments of audio with support-vector machines (SVM) [14] and Gaussian mixture models (GMM) [15, 16]. [17] took a frame-level binary classification approach to the same problem, using multilayer perceptron (MLP) models, the results of which were later improved upon with the use of a hybrid MLP/hidden Markov model using posteriors [7]. More recent work has continued with a frame-level approach using deep rectifier neural networks [18] and long short-term memory (LSTM) [2].

In addition to binary classifiers, multi-class classifiers have been implemented using 3 [19, 20], 4 [21, 22], and 7 [5] classes, all of which include the presence of an "other" class. Some of these classifiers also made use of GMMs and SVMs [20], while others have employed multivariate adaptive regression splines (MARS) [22].

2.2. Multi-label training

Datasets with multiple labels per sample are common in many fields such as image recognition [23], protein classification [24] and document classification [25]. Three traditional approaches to classifying these datasets discussed in [26] include: discarding classes; considering any combination of classes a unique class; and training one independent binary classifier for each class. Using independent classifiers does not require discarding any data or lead to additional problems of data sparsity; however, these models do not take advantage of class correlations, which [27] addressed with the use of classifier chains.

More recent approaches have used deep neural networks (DNNs) trained in an end-to-end manner to encode class dependencies such as the CNN-LSTM model in [28]. Although multi-label loss functions usually treat each class prediction as independent, [29] explained that the shared hidden layers of the model allow for class dependencies to be represented in the model parameters. More explicit modelling of class dependencies in DNNs has also been explored in [30].

3. Methodology

The task of non-verbal vocalisation detection has no clear definition or standard annotation protocol [4], hence defining the set of classes is non-trivial. In order to use supervised learning on this task, a set of classes must first be defined. This step is often overlooked in the literature so in this section we formalise it and propose different approaches towards incorporating NVV annotations into model training.

3.1. Training approaches

Let's assume that we have an annotated dataset $\mathcal{D}$, where the annotation y_a belongs to a set of classes $y_a \in \mathcal{C}_a$. This set has N_a different classes that can overlap (i.e. there is possibly more than one label per sample). We denote this set the *annotation set*. Let's then assume that we want to perform a task $\mathcal{T}$, which consists of detecting another set of classes $\mathcal{C}_\mathcal{T}$, composed of $N_\mathcal{T}$ classes, which we denote the *task set*. We finally assume that $N_\mathcal{T} \leq N_a$ and that a mapping function $f_{map}(\cdot)$ exists, which can map every class from the annotation set $\mathcal{C}_a$ to the task set $\mathcal{C}_\mathcal{T}$. The question is then: how can we use the mapping function to prepare the training set with $\mathcal{C}_\mathcal{T}$, so models can be trained to perform the task $\mathcal{T}$?

The standard approach to solve this problem consists of mapping all classes y_a before training, i.e. apply the mapping function: $y_\mathcal{T} = f_{map}(y_a)$. Hence, the dataset is now labelled with $y_\mathcal{T} \in \mathcal{C}_\mathcal{T}$ and supervised training can be used. The most commonly used mapping function is the *binary* function: it outputs two classes, "in-class" and "out-of-class". For example, as in the laughter detection task in [17], the "laughter" class is preserved while the "speech" class and all other non-verbal vocalisations are mapped to "non-laughter". Other mapping functions are sometimes used, as in [21] where the task set is composed of "laughter", "filler", "speech" and "silence".

In this study, we hypothesise that combining classes in this way hurts model performance in the case of NVV detection. Intuitively, one may assume that grouping some classes in the dataset into an "other" class would help the model ignore unrelated information; however, we propose that the task of modelling "other" is actually extremely difficult because it is comprised of a very diverse collection of sounds. We suggest that using a multi-class training approach and providing the model with a more granular set of classes will improve its performance at NVV detection. Also, inspired by the knowledge that NVVs often co-occur we explore using a multi-label training procedure to remove the assumption of one label prediction per sample.

3.1.1. Multi-class training

In this approach, we use the annotation set $\mathcal{C}_a$ for training and apply the mapping function only at inference time. For training, we assume that there is only one class prediction per sample so that the model can be trained using a softmax layer at the output of the model and the standard multi-class cross-entropy criterion. At inference time, the standard `argmax` operation can be applied on the output of model, which will give a predicted class $\hat{y}_a \in \mathcal{C}_a$. The mapping function is then used at this stage to obtain the prediction for the task: $\hat{y}_\mathcal{T} = f_{map}(\hat{y}_a)$.

3.1.2. Multi-label training

This approach is similar to the previous approach, as the mapping function is only applied at inference time. The main difference is that we relax the hypothesis that each sample must have only one class. To achieve this, the standard cross-entropy loss function cannot be used, so instead we use a loss function which is the sum of binary logistic regressions:

$$\mathcal{L} = \sum_{c=1}^{N} \log(1 + e^{-y_c \cdot f_c(x)}) \tag{1}$$

where N denotes the number of classes, $y_c \in \{-1, 1\}$ denotes the presence or absence of class c and $f_c(x)$ is the output of the model for class c. During inference, a sigmoid can be used to obtain the probability $p(c|x)$ for each class c from the output of the model $f(x)$:

$$p(c|x) = \frac{1}{1 + e^{-f_c(x)}} \tag{2}$$

To be able to evaluate a model trained using this approach, we need to compare it to binary and multi-class approaches, which can only output one class per frame. Hence, methods need to

be defined in order to obtain a single class per frame $\hat{y}$ from the output probabilities $p(c|x)$. The proposed methods are the following:

1. Binary classification: the decision is made by looking only at the relevant class probability and applying a threshold: if $p(c_1|x) > \tau \implies \hat{y}_a = c_1, \quad \hat{y}_a = c_2$ otherwise. In this paper, we used $\tau = 0.5$.
2. Multi-class classification: the decision is made by selecting the class with the highest probability: $\hat{y}_a = \underset{c}{\operatorname{argmax}}(p(c|x))$

The class prediction for the task can then be derived as before: $\hat{y}_\mathcal{T} = f_{map}(\hat{y}_a)$

3.2. Sequence-to-Sequence Models

As with other speech-related tasks, NVV detection can be considered a sequence problem. The traditional approach consists of separating the problem into two steps: the acoustic model, which predicts a class from a sample with context, and the sequence model, which yields a class prediction sequence based on the acoustic model outputs. Seq2Seq models refer to a family of neural networks able to take a sequence of samples as input and able to output an arbitrary length sequence of predictions. In this approach, the acoustic model and the sequence model are learned jointly by the same neural networks. These models clearly outperform traditional models in speech-based tasks such as automatic speech recognition [31, 32], speech emotion recognition [33] and voice conversion [34]. However in non-verbal vocalisation, the literature still used traditional approaches (see [17, 7] for example).

In this paper, the Seq2seq model used takes feature frame sequences of arbitrary length L as input and outputs L labels, i.e. one label per input frame, where the label belongs to either the annotation set or the task set as presented above. The architecture is composed of N_l bidirectional long short-term memory (BLSTM) layers with N_{hu} hidden units, followed by several feed-forward layers with a rectified linear unit (ReLU) non-linearity.

4. Experimental setup

4.1. Dataset

The ICSI meeting corpus [8] is composed of 75 meeting recordings, where each participant is recorded separately using a close-talking microphone. The corpus provides detailed hand-labelled annotations, which contain the orthographic transcription as well as non-lexical annotations separated into two categories: "vocal" containing mainly non-verbal vocalisations and "non-vocal", which contains mostly noise.

In this study, in line with previous works [17, 7], we trained and tested on the 'Bmr' subset of the ICSI meeting corpus: the first 21 meetings for training, the next 5 for validation and the last 3 for testing. Also in line with the previous studies, we discarded the segments where speech and non-verbal vocalisations are used together, as there is no timestamp to separate them, and we discarded silence between annotated segments.

This corpus was created with an unbounded set of possible annotations for non-verbal vocalisations, therefore there was no viable annotation set "out-of-the-box". To obtain a usable annotation set, while keeping as many classes as possible, we manually included all annotations which matched any of the classes in Table 1, while discarding any labels that did not match[1]. Note also that while there were some instances of class co-occurrences, where two or more NVVs occur simultaneously, we removed these so as to ensure all models were trained on the same set of data. The resulting dataset is composed of 82,959 utterances, with an average duration of 2.7 sec. Regarding the label distribution for the multi-class and multi-label training approaches: "speech" is the most represented class as expected (91%). The most represented NVV classes are "laugh" (5.7%) and "breath" (2.2%). All other classes are below 1%.

Table 1: *The annotation set used in this paper.*

Classes
speech, laugh, breath, sniff, mouth, cough, whistle, sip, click, blowing nose, sneeze, squeak, sigh, swallow, yawn, clear throat, closure, hiccup, gasp, tongue taps.

4.2. Baseline

We use a simple fully-connected neural network as a baseline. It is composed of N_l feed-forward layers of N_{hu} hidden units. Each feed-forward layer is followed by a ReLU non-linearity. Similarly to previous works [17], this model is trained frame-by-frame, where each frame is given as input to the model with some left and right context. In this paper we used a context of 37 frames as in [17]. At inference time, each frame of the sequence is given one after another to the model to obtain the sequence of prediction.

4.3. Feature and model hyper-parameters

We use log Mel filterbank coefficients as input features. These features consist of 40 coefficients computed on a 25ms window with a 10ms shift and no speed or acceleration coefficients. Features are normalised to zero-mean and unit variance by utterance.

The BLSTM architecture has 2 BLSTM layers of 256 hidden units, the output is composed of 2 feed-forward layers of 2048 hidden units. The baseline architecture has 3 layers of 1024 hidden units. The neural networks were implemented in PyTorch [35] and trained using the `fastai` library [36]. We use the "1cycle" learning rate policy introduced in [37] with a maximum learning rate of 10^{-3}.

4.4. Metrics

We use two different kinds of metrics in this study: frame-based and utterance-based. The frame-based category aggregates individual frame-level statistics, which can be misleading when the detection task involves events spanning more than one frame and the start and end times have a significant impact. For this reason, we also use utterance-based metrics.

4.4.1. Frame-based metrics

We use two frame-based metrics in this paper: (1) the frame accuracy (FA), which measures the amount of correctly predicted frames across all classes; and (2) the equal error rate (EER), which is defined as the point where the false negative rate is equal to the false positive rate. This metric can only be used for binary tasks.

[1]The scripts are available at `https://github.com/SpeechGraphics/NVV-laughter-multilabel`

4.4.2. Utterance-based metrics

Instead of considering frames only as presented above, utterance-based metrics consider the whole utterance and aim at evaluating the position of the events in the sequence. To this aim, several aspects must be taken into account such as: the correctness of the starting and ending time of the event and the number of insertions, substitutions and deletions. In this paper, we use the `sed_eval` toolbox[2], which introduces event-based metrics for sound event detection [9]. The goal of these metrics is to evaluate the position of sound events in an utterance and is well-suited to NVV detection. In this study, we use two metrics:

- F-1 score: for each class, an event is considered as a true positive if the start and end boundaries are close to the ground truth within a given time threshold. In this study we use 50ms.
- Error Rate (ER): Similar to the Word Error Rate (WER) used in ASR, this metric is computed as the number of insertions, deletions and substitutions divided by the number of events in the ground-truth. Hence, it can easily go above 100% if there are more insertions than events in an utterance.

5. Results

In this section, we present the results of the study. We train two models using the approaches described in Section 3: one using the multi-class approach and the other using the multi-label approach. They both use the annotation set presented in Table 1. We use these two models for both experiments presented in this section.

5.1. Non-verbal vocalisation detection

In this experiment, we evaluate the two models and compare them with the baseline. Table 2 presents the results using the metrics introduced in Section 4.4.

Table 2: *Performance on NVV detection using the ICSI test set.*

System	FA [%]	F-1 [%]	ER [%]
Baseline	95.2	15.2	192.3
Seq2Seq, multi-class	97.2	79.0	41.2
Seq2Seq, multi-label	**97.4**	**79.9**	**39.8**

We can see that the two Seq2Seq models outperform the baseline. The frame accuracy is quite close for all three systems, its very high value can be explained by the label distribution, as most of the frames are labeled "speech" (see Section 4.1). The utterance-based metrics show a clear difference: the baseline only achieves a 15% F-1 score and more than 100% error rate. This is expected, as the baseline does not have a sequence model to help smooth its output. This clearly shows that simply using the frame accuracy is not sufficient to evaluate a system for this task. It is also worth noting that the multi-label approach slightly outperforms the multi-class approach.

5.2. Laughter detection

In this experiment, we focus on the task of laughter detection, where the task set is binary: "laughter" or "non-laughter". We compare the two models with a model that has the same architecture but is trained on the task set directly as in [7]. We also compare them with the literature. Table 3 presents the results. On the frame-level EER, it clearly shows that the Seq2Seq models outperform the literature, yielding a new state-of-the-art result for this task. Also, the multi-class approach outperforms the binary approach, and the multi-label approach outperforms both, showing that training with multiple classes and then mapping to the binary "laughter" or "non-laughter" at inference time helps for this task.

Table 3: *Performance on laughter detection using the ICSI test set.*

System	EER [%]	F-1 [%]	ER [%]
MLP [17]	8.15	na	na
MLP MF [7]	5.4	na	na
Seq2Seq, binary	2.04	68.6	62.8
Seq2Seq, multi-class	1.84	78.3	45.6
Seq2Seq, multi-label	**1.58**	**79.3**	**42.7**

5.3. Discussion

These results support our hypothesis: training on more classes, rather than mapping to a smaller set of classes improves model performance for the task of NVV detection. This is in agreement with our assertion that mapping before training and including an "other" class leads to more ambiguity, making it difficult to model.

The fact that the multi-label approach seems to bring an improvement over the multi-class approach is promising. The multi-label approach can be seen as generalisation of the multi-class approach and brings the natural benefit of being able to model data with class co-occurrences. With regards to the cause of this improvement, a possible explanation could be due to ambiguous boundary placement between NVVs or unannotated NVV co-occurrence; however, further analysis is required. This will be part of our future work along with analysing in more detail the performance of the NVV detection models, especially focusing on the under-represented classes. We will also study the effect of the multi-label loss function on training.

6. Conclusions

In this paper, we presented the hypothesis that training with more classes improves performance for non-verbal vocalisation and laughter detection. We proposed two training approaches: (1) *multi-class* training, where the mapping from the annotation set to the task set is performed at inference time, and (2) the *multi-label* approach, which used a similar mapping strategy while additionally being able to predict more than one class per frame. We showed that both approaches outperform the standard approach on non-verbal vocalisation and laughter detection, and that the multi-label approach yields better performance than the multi-class approach. Finally, we showed that using utterance-based metrics for evaluation gives a better picture of the performance of the models, highlighting the benefits of the sequence-to-sequence approach. For future work, we will focus on a more detailed analysis of the multi-label approach, especially on data with label co-occurrences.

[2] `https://tut-arg.github.io/sed_eval/index.html`

7. References

[1] W. Huang, T. K. Chiew, H. Li, T. S. Kok, and J. Biswas, "Scream detection for home applications," in *Proc. of IEEE Conference on Industrial Electronics and Applications*. IEEE, 2010, pp. 2115–2120.

[2] G. Hagerer, N. Cummins, F. Eyben, and B. Schuller, "'Did you laugh enough today?' - deep neural networks for mobile and wearable laughter trackers," in *Proc. of Interspeech*. ISCA, 2017, pp. 2044–2045.

[3] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for covid-19," *arXiv preprint arXiv:2005.08579*, 2020.

[4] J. Trouvain and K. Truong, "Comparing non-verbal vocalisations in conversational speech corpora," in *"Proc. of Workshop on Corpora for Research on Emotion Sentiment and Social Signals*. ELRA, 2012, pp. 36–39.

[5] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Perception and Interactive Technologies for Speech-based Systems*, 06 2008, pp. 99–110.

[6] K. Huang, C. Wu, Q. Hong, M. Su, and Y. Chen, "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds," in *Proc. of ICASSP*. IEEE, 2019, pp. 5866–5870.

[7] M. T. Knox, N. Morgan, and N. Mirghafori, "Getting the last laugh: Automatic laughter segmentation in meetings," in *Proc. of Interspeech*. ISCA, 2008, pp. 797–800.

[8] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," in *Proc. of ICASSP*. IEEE, 2003, pp. 364–367.

[9] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.

[10] V. Nallanthighal, A. Härmä, and H. Strik, "Deep sensing of breathing signal during conversational speech," in *Proc. of Interspeech*. ISCA, 09 2019, pp. 4110–4114.

[11] M. Yasar Arafath K. and A. Routray, "Automatic detection of breath using voice activity detection and svm classifier with application on news reports," in *Proc. of Interspeech*. ISCA, 09 2019, pp. 609–613.

[12] J. Liu, M. You, Z. Wang, G. Li, X. Xu, and Z. Qiu, "Cough detection using deep neural networks," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014, pp. 560–563.

[13] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *Proc. of ICASSP*. IEEE, 2015, pp. 161–165.

[14] L. Kennedy and D. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop*. National Institute of Standards and Technology, 2004.

[15] K. P. Truong and D. V. Leeuwen, "Automatic detection of laughter," in *Proc. of Interspeech*. ISCA, 2005, pp. 485–488.

[16] K. Truong and D. Van Leeuwen, "Evaluating automatic laughter segmentation in meetings using acoustic and acousticphonetic features," in *Proc. of the Interdisciplinary Workshop on the Phonetics of Laughter*, 2007.

[17] M. T. Knox and N. Mirghafori, "Automatic laughter detection using neural networks," in *Proc. of Interspeech*. ISCA, 2007, pp. 2973–2976.

[18] G. Gosztolya, A. Beke, T. Neuberger, and L. Tóth, "Laughter classification using deep rectifier neural networks with a minimal feature subset," *Archives of Acoustics*, vol. 41, no. 4, 2016.

[19] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Laughter and filler detection in naturalistic audio," in *Proc. of Interspeech*. ISCA, 2015, pp. 2509–2513.

[20] A. Janicki, "Non-linguistic vocalisation recognition based on hybrid gmm-svm approach," in *Proc. of Interspeech*. ISCA, 2013.

[21] H. Salamin, A. Polychroniou, and A. Vinciarelli, "Automatic detection of laughter and fillers in spontaneous mobile phone conversations," in *Proc. of IEEE International Conference on Systems, Man, and Cybernetics*, 2013.

[22] W.-H. Liao and Y.-K. Lin, "Classification of non-speech human sounds: Feature selection and snoring sound analysis," in *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, 10 2009, pp. 2695–2700.

[23] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *Proc. of CVPR*. IEEE, 2017, pp. 5513–5522.

[24] R. Cerri, R. C. Barros, A. C. de Carvalho, and Y. Jin, "Reduction strategies for hierarchical multi-label classification in protein function prediction," *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–24, 2016.

[25] N. Ghamrawi and A. McCallum, "Collective multi-label classification," in *Proc. of International Conference on Information and Knowledge Management*, 2005, pp. 195–200.

[26] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[27] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, p. 333, 2011.

[28] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. of CVPR*. IEEE, 2016, pp. 2285–2294.

[29] D. Palaz, G. Synnaeve, and R. Collobert, "Jointly learning to locate and classify words using convolutional networks," *Proc. of Interspeech*, pp. 2741–2745, 2016.

[30] M. Cisse, M. Al-Shedivat, and S. Bengio, "Adios: Architectures deep in output space," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proc. of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. PMLR, 20–22 Jun 2016, pp. 2770–2779.

[31] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. of ICASSP*, IEEE. IEEE, 2018, pp. 4774–4778.

[32] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *Proc. of ICASSP*, IEEE. IEEE, 2021, pp. 5874–5878.

[33] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. of ICASSP*. IEEE, 03 2017.

[34] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 3, pp. 631–644, 2019.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. of NIPS*, 2019, pp. 8024–8035.

[36] J. Howard and S. Gugger, "Fastai: A layered API for deep learning," *Information*, vol. 11, pp. 108–134, 2020.

[37] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100612.