EMOTION NEURAL TRANSDUCER FOR FINE-GRAINED SPEECH EMOTION RECOGNITION

Siyuan Shen[1,2,3†] *Yu Gao*[3*] *Feng Liu*[1,2] *Hanyang Wang*[1,2] *Aimin Zhou*[1,2*]

[1]Shanghai Institute of AI for Education, East China Normal University, Shanghai, China
[2]School of Computer Science and Technology, East China Normal University, Shanghai, China
[3]AI Innovation Center, Midea Group, Shanghai, China

ABSTRACT

The mainstream paradigm of speech emotion recognition (SER) is identifying the single emotion label of the entire utterance. This line of works neglect the emotion dynamics at fine temporal granularity and mostly fail to leverage linguistic information of speech signal explicitly. In this paper, we propose Emotion Neural Transducer for fine-grained speech emotion recognition with automatic speech recognition (ASR) joint training. We first extend typical neural transducer with emotion joint network to construct emotion lattice for fine-grained SER. Then we propose lattice max pooling on the alignment lattice to facilitate distinguishing emotional and non-emotional frames. To adapt fine-grained SER to transducer inference manner, we further make blank, the special symbol of ASR, serve as underlying emotion indicator as well, yielding Factorized Emotion Neural Transducer. For typical utterance-level SER, our ENT models outperform state-of-the-art methods on IEMOCAP in low word error rate. Experiments on IEMOCAP and the latest speech emotion diarization dataset ZED also demonstrate the superiority of fine-grained emotion modeling. Our code is available at https://github.com/ECNU-Cross-Innovation-Lab/ENT.

***Index Terms*—** Speech emotion recognition, speech emotion diarization, automatic speech recognition

1. INTRODUCTION

Speech emotion recognition (SER) aims to identify emotional states of human speech signals. Many works follow the recipe of classifying the whole utterance into single emotion category [1, 2, 3, 4, 5, 6]. Fueled by various datasets containing emotional labels at utterance level [7, 8], such sequence-to-label methods have made great progress in recent years. However, emotional states inherently exhibit diverse temporal dynamics, often leading to alternating shifts between emotional and non-emotional states within a single utterance [9]. Consequently, recognizing emotions at a fine temporal granularity is desirable for better emotion understanding [10, 11].

For fine-grained SER, another line of previous works consider this task as a sequence-to-sequence problem. These methods can be thought of as aligning frames and emotional labels in a weakly supervised manner. Frame-wise methods simply assign overall emotional label to each frame [12] while segment-wise methods identify emotional regions according to contribution of salient parts with attention mechanism [13]. Connectionist temporal classification (CTC) methods [14] first construct emotion sequence heuristically and then align emotionally relevant segments within the utterance automatically [9]. Though driven by the common motivation for fine-grained SER, these approaches only consider acoustic information of speech signals and evaluate performance at utterance level. Thanks to the latest proposed benchmark of speech emotion diarization [10], the frontier of distinguishing emotions at a fine temporal granularity is to be uncovered.

Motivated by the nature of neural transducer for sequence alignment conditioning on both linguistic tokens and acoustic units [15, 16], we explore SER and fine-grained SER based on transducer models with automatic speech recognition (ASR) joint training. To date, recent paradigms for joint SER and ASR at utterance level include cascading off-the-shelf ASR model [17] as well as adopting multi-task learning framework, where intermediate [18, 19] or task-specific output layers [20] are supervised by CTC loss. Despite the huge success of transducer family in the field of ASR [15, 16, 21], existing extension on RNN-T for additional SER functionality solely focuses on modifying target transcriptions with emotion tags [22]. Moreover, these ASR-based methods view SER as a typical utterance-level classification problem, disregarding the temporal granularity of emotion.

In this paper, we aim to bridge the gap between ASR-based SER and fine-grained SER, allowing generating rich transcripts along with emotion synchronously. We propose Emotion Neural Transducer (dubbed ENT) for fine-grained speech emotion recognition with ASR joint training. We first build the emotion joint network upon the typical acoustic

This paper is funded by the Science and Technology Commission of Shanghai Municipality (Grant No. 22511105901) and the Beijing Key Laboratory of Behavior and Mental Health, Peking University.

† Work done while intern at Midea Group.

* Corresponding author. gaoyu11@midea.com, amzhou@cs.ecnu.edu.cn

encoder and vocabulary predictor and thus enable modeling emotion categorical distribution through the alignment lattice as standard neural transducer [15]. Since fine-grained SER operates under a a weakly-supervised learning paradigm, we propose lattice max-pooling loss for the emotion lattice to distinguish emotional and non-emotional timestamps automatically. Motivated by the inference manner of neural transducer, we further extend emotion neural transducer to the factorized variant (called FENT). The key concept behind FENT is to utilize the blank symbol as both a time separator and an underlying indicator of emotion. Specifically, we disentangle emotion and blank prediction from vocabulary prediction with separate predictors and share the predictor for both blank and emotion prediction. Our proposed ENT models outperform previous state-of-the-arts on the benchmark IEMOCAP dataset with low word error rate. Moreover, we validate fine-grained emotion modeling with ASR on the newly proposed emotion diarization dataset ZED.

2. NEURAL TRANSDUCER

Standard neural transducer consists of three components, the acoustic encoder, prediction network and joint network [15, 16]. Considering the acoustic input x with duration T and target label sequence y with length U, the acoustic encoder takes acoustic features $x_{\leq t}$ as input and produces hidden features h_t for each timestamp. The prediction network generates label representations g_u conditioning on previous tokens $y_{\leq u}$. The joint network integrates the outputs of acoustic encoder and prediction network as $z_{t,u}$ to compute vocabulary label distribution. The procedure can be formulated as

$$\begin{aligned} h_t &= \text{Encoder}(x_{\leq t}) \\ g_u &= \text{Predictor}(y_{\leq u}) \\ z_{t,u} &= \text{Joint}(h_t, g_u) \end{aligned} \tag{1}$$

Then the probability of next token can be computed as

$$P(y_{u+1} \mid x_{\leq t}, y_{\leq u}) = \text{softmax}(z_{t,u}). \tag{2}$$

To address the length difference between acoustic features x and token sequences y, transducer models add a special blank symbol to the vocabulary for alignment and optimize log probability over all possible alignments as

$$\mathcal{L}_{trans} = -\log \sum_{\alpha \in \beta^{-1}(y)} (P(\alpha \mid x)), \tag{3}$$

where α denotes the alignment, each containing $T+U$ tokens and β is the mapping from alignment to target sequence by removing blank symbols.

3. EMOTION NEURAL TRANSDUCER

In this section, we present two key components of ENT and subsequently extend it to its factorized variant FENT. First, we construct the emotion joint network to integrate representations from the encoder and predictor to yield emotion lattice. To further enhance emotional and non-emotional awareness at temporal granularity, we then devise lattice max pooling loss to the generated emotion lattice. Next, we make the blank symbol work as an emotion indicator for FENT by disentangling blank from vocabulary and meanwhile sharing the same predictor for both blank and emotion prediction.

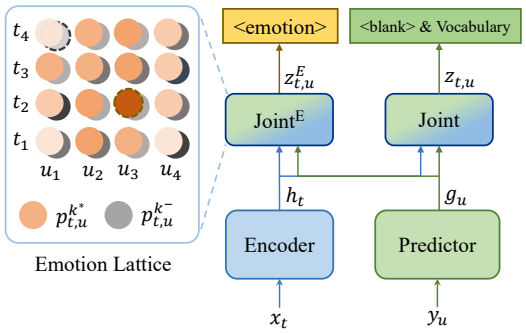


Fig. 1. Emotion Neural Transducer.

3.1. Joint Emotion Prediction

To integrate both acoustic and linguistic tokens for emotion recognition at a fine temporal granularity, we build emotion joint network upon the typical acoustic encoder and vocabulary predictor (Figure 1). Formally, the emotion representation $z^E_{t,u}$ given speech and text history can be obtained by substituting joint$^{\text{E}}$ into Equation 1. Similar to standard neural transducer modeling sequence alignment via lattice [15], our emotion joint network models emotion emission probability through the $T \times U$ alignment lattice. As shown in Figure 1, each node $p_{t,u}$ denotes the emotion probability distribution having output u tokens by frame t, where $p^k_{t,u}$ is the probability of emotion k with darker color indicating higher probability, orange/grey denoting emotional/non-emotional.

3.2. Lattice Max Pooling

Given the utterance-level emotion label k^*, our goal is to identify the emotional and non-emotional frames automatically through the lattice. Inspired by max pooling loss used in keyword spotting [23, 24], we extend the frame-level max pooling loss on the emotion lattice, thus leveraging acoustic and linguistic alignment. For each utterance, we select the node with the highest predicted posterior probability of target emotion $p^{k^*}_{t,u}$ and the node with the minimum non-emotional or neutral category probability $p^{k^-}_{t,u}$. In Figure 1, the selected nodes are indicated by dashed borderline. Our proposed lattice max pooling loss can be formulated as

$$\mathcal{L}_{lattice} = -\log \max_{t,u}(p^{k^*}_{t,u}) - \log \min_{t,u}(p^{k^-}_{t,u}). \tag{4}$$

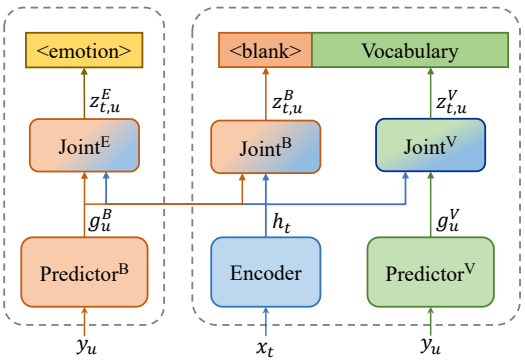


Fig. 2. Factorized Emotion Neural Transducer.

From the view of positive and negative samples as max pooling loss, the first term optimizes the most positive frame while the second term selects the hardest negative sample through the emotion lattice. See Appendix A for variants.

To maintain the capability of conventional SER at utterance level, we do mean pooling for the representations from acoustic encoder and predictor, optimized by cross entropy loss $\mathcal{L}_{emotion}$ at utterance level.

3.3. Factorized Emotion Neural Transducer

Our intuition for the factorized variant is based on the natural inference manner of neural transducer. At each timestamp, the standard transducer model consumes one frame and then outputs multiple non-blank tokens until the blank is emitted. We assume the blank symbol as accumulation of both acoustic and linguistic information and thus we allocate temporal emotion awareness to blank representations. Inspired by recent advances in language model adaptation [21, 25], we first disentangle blank from vocabulary prediction by using two separate predictors. The overall architecture of FENT is described in Figure 2. Specifically, the vocabulary $\text{predictor}^{\text{V}}$ is dedicated to predicting label vocabulary representations g^V_u excluding blank while the blank predictor $\text{predictor}^{\text{B}}$ produces blank representations g^B_u as the right part of Figure 2. Then the corresponding joint network fuses acoustic features h_t with predictor outputs similar to Equation 1, yielding $z^V_{t,u}$ and $z^B_{t,u}$ respectively. The whole vocabulary label distribution can be computed by softmax and concatenation as

$$P(y_{u+1} \mid x_{\leq t}, y_{\leq u}) = \text{softmax}([z^B_{t,u}; z^V_{t,u}]). \quad (5)$$

To bias the blank symbol towards emotion, we employ a shared predictor for emotion and blank prediction and adopt aforementioned emotion joint network as shown in left part of Figure 2. For each time step during inference, the acoustic encoder takes one frame as input and the vocabulary predictor outputs the most probable tokens iteratively until the blank is emitted. At this point, the emotion joint network fuses acoustic and blank representation to predict current emotion.

Method	Year	WA (%)	UA (%)
Wav2vec2-PT [1]	2021	67.90	-
Corr Attentive [2]	2023	-	70.01
DCW+TsPA [3]	2023	72.08	72.17
Shiftformer [4]	2023	72.10	72.70
MSTR [5]	2023	70.60	71.60
EmotionNAS [6]	2023	69.10	72.10
ENT (ours)	2023	**72.43**	**73.88**
FENT (ours)	2023	71.84	72.37

Table 1. Comparison with utterance-level SER methods using wav2vec 2.0 as feature extractor on IEMOCAP.

4. EXPERIMENTS

In this section, we first demonstrate the superiority of ENT models on the benchmark dataset IEMOCAP for utterance-level SER. Next we validate the capability of fine-grained speech emotion recognition on the speech emotion diarization dataset ZED and meanwhile ablate key components.

4.1. Experimental Setup

Dataset and evaluation. Interactive emotional dyadic motion capture database (IEMOCAP) [7] is a widely-used benchmark SER dataset, where each utterance is annotated with the transcript and single emotion category label. We adopt leave-one-session-out 5-fold cross-validation, following the typical evaluation protocol. The unweighted accuracy (UA) and weighted accuracy (WA) for utterance-level SER are computed by averaging the results obtained from the 5 folds. The average word error rate (WER) across the 5 folds is reported to measure ASR performance.

Zaion Emotion Dataset (ZED) [10] is a recently proposed dataset for fine-grained SER, named as speech emotion diarization, including 180 utterances annotated with emotional boundaries for each. It is worth noting that due to its limited scale, ZED is primarily suitable for evaluating the fine-grained SER capability rather than serving as a comprehensive training set in a fully supervised manner. Thus we train our ENT models on IEMOCAP and validate on ZED. We adopt emotion diarization error rate (EDER) for fine-grained SER, which assesses the temporal alignment between predicted emotion intervals and the actual emotion intervals. Lower EDER indicates better fine-grained SER ability.

Implementation Details. We take wav2vec 2.0 Base [26] as feature extractor for input speech signals, where the pre-trained model is frozen for training efficiency and the features from different layers are performed weighted sum in line with SUPERB [27]. The acoustic encoder and the predictors are one-layer LSTM with a hidden dimension of 640. The joint network combines features of encoder and predictor by addition operation, followed by a linear layer.

4.2. Utterance-level Speech Emotion Recognition

Comparison with state-of-the-arts. We compare our proposed models with recent state-of-the-art methods in Table 1. ENT outperforms all the strong baselines, showing the effectiveness of leveraging linguistic information and fine-grained temporal modeling. While FENT achieves competitive results as well, the factorization technique degrades its utterance-level SER performance slightly compared with ENT just as its counterpart in language adaptation [21]. This phenomenon indicates that factorization of predictor partially compromises the integrity of whole vocabulary modeling, resulting in inferior representation for utterance-level discrimination.

Comparison with ASR-based methods. We evaluate the SER and ASR performance in Table 2. For fair comparison, all the methods take features from self-supervised or ASR pre-trained models. Although ASR joint training enables the model to predict emotions along with transcriptions, previous attempts fail to balance the mutual influence between ASR and SER. Taking RNN-T method as an example, appending a special emotion tag to the target text is conductive to the original ASR output manner, yet deteriorating SER ability (only 58.2% WA). In contrast, the family of ENT attains better performance in both ASR and SER (+3% UA and meanwhile -0.7% WER). Notably, the top performance of FENT in WER validates the effectiveness of factorization of emotion from vocabulary, preserving the modularity of the transducer for ASR capability [28, 29] while endowing SER capability.

Type	Method	WA (%)	WER (%)
CTC	e2e-ASR [18]	68.60	35.70
	wav2vec 2.0+co-attention [19]	63.40	32.70
RNN-T	Emotion tag [22]	58.20	26.70
ENT	ENT (ours)	**72.43**	26.47
	FENT (ours)	71.84	**25.99**

Table 2. Comparison with ASR-based utterance-level SER methods on IEMOCAP.

4.3. Fine-Grained Speech Emotion Recognition

Table 3 is split into 3 parts to compare with frame-wise methods and ENT variants. It is noteworthy that the weak supervision paradigm based on utterance-level annotation and absence of an appropriate training set makes fine-grained SER validation on SED benchmark extremely challenging. Interestingly, standard ENT without lattice loss, though utilizing text information explicitly, lags behind frame-wise baseline by nearly 3% EDER, suffering from degraded ASR capability as well as imperfect transcripts. Thanks to disentangling emotion and blank from vocabulary prediction, our FENT reaches much lower EDER (about -4.6%) while enjoying speech transcription functionality along with fine-grained emotion.

Method	IEMOCAP		ZED	
	UA ↑	WER ↓	EDER ↓	WER ↓
Frame-wise	68.43	-	59.73	-
ENT	**73.88**	26.47	56.60	39.37
-w/o $\mathcal{L}_{lattice}$	71.76	26.06	62.47	39.19
-w. $\mathcal{L}^{T}_{lattice}$	73.11	26.42	61.88	39.39
-w. $\mathcal{L}^{U}_{lattice}$	69.86	26.14	61.40	**38.82**
-w. $\mathcal{L}^{all}_{lattice}$	71.85	26.19	61.12	39.28
-w. mixing	-	-	52.76*	42.54*
-w. BPE	71.37	30.13	67.68	47.42
FENT	72.37	**25.99**	**55.07**	39.34
-w/o $\mathcal{L}_{lattice}$	71.84	26.69	60.86	39.14
-w. $\mathcal{L}^{T}_{lattice}$	72.52	26.28	59.18	39.48
-w. $\mathcal{L}^{U}_{lattice}$	69.67	26.23	60.86	39.17
-w. $\mathcal{L}^{all}_{lattice}$	69.61	26.18	59.38	39.11
-w. mixing	-	-	54.41*	39.93*
-w. BPE	70.33	30.96	65.63	47.26

Table 3. Comparison of ENT varaints performance on IEMOCAP and ZED. * denotes training models with concatenated IEMOCAP audio segments like [10] and $\mathcal{L}^{all}_{lattice}$.

4.4. Ablation Studies

We investigate key components of ENT models in Table 3. Overall, FENT architecture excels at fine-grained SER on SED regardless of $\mathcal{L}_{lattice}$ while ENT obtains better UA in typical utterance-level SER. Compared with character units, text encoded with byte-pair encoding (BPE) degrades WER as well as emotion recognition performance significantly, which may be attributed to vocabulary sparsity for relatively small SER dataset, further yielding negative mutual impact of speech and emotion recognition. We then compare our lattice max pooling to some straightforward variants, where $\mathcal{L}^{T}_{lattice}$ selects the entire timestamp (target row of emotion lattice in Figure 1) while $\mathcal{L}^{U}_{lattice}$ selects the target token column. And $\mathcal{L}^{all}_{lattice}$ applies supervision on the whole emotion lattice. We can observe that $\mathcal{L}^{T}_{lattice}$ achieves on-par performance as original $\mathcal{L}_{lattice}$, signifying the importance of temporal localization. More importantly, the improvement of models with lattice max pooling on IEMOCAP also verifies that fine-grained emotion modeling helps utterance-level SER. Moreover, improvement by mixing different audio segments shows compatibility of our lattice loss to supervised data.

5. CONCLUSION

In this paper, we present Emotion Neural Transducer models for fine-grained speech emotion recognition, with a favorable capability of predicting transcripts along with emotion at fine temporal granularity for practice. We hope our work will draw more attention from the community toward more comprehensive fine-grained emotion benchmarks.

6. REFERENCES

[1] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[2] Ke Liu, Dekui Wang, Dongya Wu, and Jun Feng, "Speech emotion recognition via two-stream pooling attention with discriminative channel weighting," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] Sofoklis Kakouros, Themos Stafylakis, Ladislav Mošner, and Lukáš Burget, "Speech-based emotion recognition with self-supervised models using attentive channel-wise correlations and label smoothing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] Siyuan Shen, Feng Liu, and Aimin Zhou, "Mingling or misalignment? temporal shift for speech emotion recognition with pre-trained representations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] Zhipeng Li, Xiaofen Xing, Yuanbo Fang, Weibin Zhang, Hengsheng Fan, and Xiangmin Xu, "Multi-Scale Temporal Transformer For Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 3652–3656.

[6] Haiyang Sun, Zheng Lian, Bin Liu, Ying Li, Jianhua Tao, Licai Sun, Cong Cai, Meng Wang, and Yuan Cheng, "EmotionNAS: Two-stream Neural Architecture Search for Speech Emotion Recognition," in *Proc. INTERSPEECH 2023*, 2023, pp. 3597–3601.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[8] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 527–536.

[9] Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, and Björn Schuller, "Towards temporal modelling of categorical speech emotion recognition," *Interspeech 2018*, 2018.

[10] Yingzhi Wang, Mirco Ravanelli, Alaa Nfissi, and Alya Yacoubi, "Speech emotion diarization: Which emotion appears when?," *arXiv preprint arXiv:2306.12991*, 2023.

[11] Juncheng Li, Junlin Xie, Linchao Zhu, Long Qian, Siliang Tang, Wenqiao Zhang, Haochen Shi, Shengyu Zhang, Longhui Wei, Qi Tian, et al., "Dilated context integrated network with cross-modal consensus for temporal emotion localization in videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5083–5092.

[12] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[13] Shuiyang Mao, PC Ching, C-C Jay Kuo, and Tan Lee, "Advancing multiple instance learning with attention modeling for categorical speech emotion recognition," *Proc. Interspeech 2020*, pp. 2357–2361, 2020.

[14] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[15] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[16] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[17] Chengxin Chen and Pengyuan Zhang, "Cta-rnn: Channel and temporal-wise attention rnn leveraging pre-trained asr embeddings for speech emotion recognition," in *Interspeech*, 2022.

[18] Han Feng, Sei Ueno, and Tatsuya Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model.," in *Interspeech*, 2020, pp. 501–505.

[19] Yuanchao Li, Peter Bell, and Catherine Lai, "Fusing asr outputs in joint training for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7362–7366.

[20] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church, "Speech emotion recognition with multi-task learning.," in *Interspeech*, 2021, vol. 2021, pp. 4508–4512.

[21] Xie Chen, Zhong Meng, Sarangarajan Parthasarathy, and Jinyu Li, "Factorized neural transducer for efficient language model adaptation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8132–8136.

[22] Zvi Kons, Hagai Aronowitz, Edmilson Morais, Matheus Damasceno, Hong-Kwang Kuo, Samuel Thomas, and George Saon, "Extending RNN-T-based speech recognition systems with emotion and language classification," in *Proc. Interspeech 2022*, 2022, pp. 546–549.

[23] Jingyong Hou, Yangyang Shi, Mari Ostendorf, Mei-Yuh Hwang, and Lei Xie, "Mining effective negative training samples for keyword spotting," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7444–7448.

[24] Jie Wang, Menglong Xu, Jingyong Hou, Binbin Zhang, Xiao-Lei Zhang, Lei Xie, and Fuping Pan, "Wekws: A production first small-footprint end-to-end keyword spotting toolkit," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley, "Hybrid autoregressive transducer (hat)," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6139–6143.

[26] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[27] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

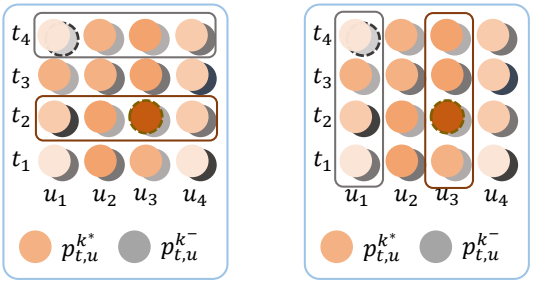
[28] Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein, "Rnn-transducer with stateless prediction network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7049–7053.

[29] Zhong Meng, Sarangarajan Parthasarathy, Eric Sun, Yashesh Gaur, Naoyuki Kanda, Liang Lu, Xie Chen, Rui Zhao, Jinyu Li, and Yifan Gong, "Internal language model estimation for domain-adaptive end-to-end speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 243–250.

A. VARIANTS OF LATTICE MAX POOLING

As mentioned in our experiment, the lattice max pooling loss can be extended to some variants based on the groups of selected node and the supervision manner. We define the indices of the nodes with the highest predicted probability of the target emotion as t^* and u^*, and the indices of the nodes with the minimum non-emotional probability as t^- and u^-.

$$t^*, u^* = \arg\max_{t,u}(p_{t,u}^{k^*}),$$
$$t^-, u^- = \arg\min_{t,u}(p_{t,u}^{k^-}). \tag{6}$$



(a) Temporal Lattice Max Pooling $\mathcal{L}_{lattice}^{T}$.

(b) Token Lattice Max Pooling $\mathcal{L}_{lattice}^{U}$.

Fig. 3. Temporal and Token Lattice Max Pooling Loss.

Temporal Lattice Max Pooling $\mathcal{L}_{lattice}^{T}$ (see Figure 3(a)) first selects the nodes within the entire timestamp row instead of a single node and then calculates the loss as follows

$$\mathcal{L}_{lattice}^{T} = -\sum_{u}\log(p_{t^*,u}^{k^*}) - \sum_{u}\log(p_{t^-,u}^{k^-}). \tag{7}$$

Token Lattice Max Pooling $\mathcal{L}_{lattice}^{U}$ (see Figure 3(b)) first selects the nodes within the entire token column and then calculates the loss as follows

$$\mathcal{L}_{lattice}^{U} = -\sum_{t}\log(p_{t,u^*}^{k^*}) - \sum_{t}\log(p_{t,u^-}^{k^-}). \tag{8}$$

Mixing method (see Figure 4) concatenates neutral speech recordings with other emotional speech recordings to create training samples that contain emotional intervals. Subsequently, we can apply supervision to each interval.

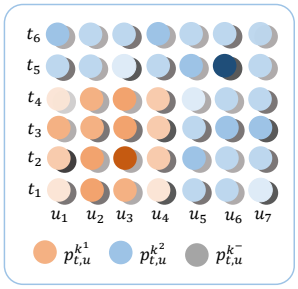


Fig. 4. Mixing on Emotion lattice.