SPEECH-BASED EMOTION RECOGNITION WITH SELF-SUPERVISED MODELS USING ATTENTIVE CHANNEL-WISE CORRELATIONS AND LABEL SMOOTHING

Sofoklis Kakouros[1], Themos Stafylakis[2], Ladislav Mošner[3], Lukáš Burget[3]

[1]University of Helsinki, Finland
[2]Omilia - Conversational Intelligence, Athens, Greece
[3]Brno University of Technology, Faculty of Information Technology, Speech@FIT, Czechia

ABSTRACT

When recognizing emotions from speech, we encounter two common problems: how to optimally capture emotion-relevant information from the speech signal and how to best quantify or categorize the noisy subjective emotion labels. Self-supervised pre-trained representations can robustly capture information from speech enabling state-of-the-art results in many downstream tasks including emotion recognition. However, better ways of aggregating the information across time need to be considered as the relevant emotion information is likely to appear piecewise and not uniformly across the signal. For the labels, we need to take into account that there is a substantial degree of noise that comes from the subjective human annotations. In this paper, we propose a novel approach to attentive pooling based on correlations between the representations' coefficients combined with label smoothing, a method aiming to reduce the confidence of the classifier on the training labels. We evaluate our proposed approach on the benchmark dataset IEMOCAP, and demonstrate high performance surpassing that in the literature. The code to reproduce the results is available at `github.com/skakouros/s3prl_attentive_correlation`.

Index Terms— emotion recognition, self-supervised features, iemocap, hubert, wavlm, wav2vec 2.0

1. INTRODUCTION

Emotional expressions are a fundamental component of spoken interaction. When we communicate with other people, we are implicitly monitoring their emotional state and respond based on that emotional state [1]. Emotions fall in the realm of prosodic function. In recent years, the importance of prosodic qualities in speech has attracted increasing attention. This is also the case for speech emotion recognition (SER) which has seen a growing interest with the increasing role of spoken language interfaces in human-computer interaction (HCI) applications [2]. However, recognizing emotions in speech remains a challenging problem complicated by numerous factors including fundamental issues of how emotion is defined, elicited, expressed, and communicated [3] and extending to how we can capture this information from speech.

The challenges in SER can be split into three distinct problems. First, we encounter the issue of developing engineered representations that can robustly capture the acoustic information in speech that best describe the variation found across different emotions. This has been traditionally done using features such as mel-frequency cepstral coefficients (MFCCs), filterbanks, fundamental frequency, energy, zero-crossing rate, chroma-based features and their feature functionals [4] or through standardized feature collections and their functionals such as eGeMAPS [5]. More recently, self-supervised learning (SSL) has shown its effectiveness in various domains, including SER, and is becoming the new principle for extracting representations from speech. HuBERT [6], Wav2vec 2.0 [7], WavLM [8], are some of the self-supervised approaches for speech representation learning that have been used in the context of SER [9, 5].

The second issue that we face in SER is the effective modeling of the long temporal context over which emotions take place. Emotion specific information lies beyond segmental productions and in longer time scales. These may include parts of an utterance but can also span across one or more utterances. To appropriately model long-term dependencies we need to utilize suitable methods that are capable of capturing and connecting the relevant cues across time. These vary from approaches that simply take the first and second order statistics of self-supervised representations across time [9] to approaches that focus on complex sequence modeling tasks [10]. For example Sarma et al. [10] used a TDNN architecture combined with LSTM and self-attention to model the long-term temporal context and to capture the emotionally relevant portions of speech. In a recent work, Liu et al. [11] used a cascaded attention network to locate the relevant emotion regions from the input features. Other approaches also use different types of recurrent neural networks (RNNs) to explain the long temporal contexts of emotions in speech [12].

The third and final issue comes from the observation that human emotional expressions are often unclear and ambiguous, leading to disagreement and confusion among human

evaluators [13, 14]. This confusion might be partly attributed to the multimodal nature of emotion expression. Facial expressions, hand gestures, and speech with its prosodic and linguistic content all work together in eliciting different emotions. Perhaps the absence of multimodality may be one source of confusion that leads to overlaps in the clusters of the different emotion classes. However, speech alone holds much relevant information in its prosodic content that can be used for robust SER. Different ways to tackle the problem of noisy labels and consequently the uncertainty in predicting emotions have been suggested in the literature. These typically include custom loss functions [15, 11] and modifications to the target hard labels [16, 17]. For example Liu and Wang used a triplet loss to make anchor utterances more similar to all other positive utterances [15] while Tarantino et al. used regression targets instead of hard categorical targets by taking the proportion of the classes within the annotations [17].

This paper presents a framework for SER that uses pretrained speech models with a novel approach to attentive pooling based on channel-wise correlations on soft targets. We evaluate the framework with HuBERT [6], Wav2vec 2.0 [7], and WavLM [8] upstream models. We use the SUPERB [9] evaluation setup throughout our experiments. The effectiveness of our proposed framework is evaluated on the interactive emotional dyadic motion capture (IEMOCAP) dataset [18] and shows state-of-the-art performance in SER.

2. RELATION TO PRIOR WORK

The idea of taking the correlations between different filter responses over the spatial extent of the feature maps to obtain a representation of the style of an input image was introduced by Gatys et al. [19]. Their method was later adapted to speech where it has found applications in speech generation and voice conversion [20], pooling to obtain speaker embeddings [21], and sentence-level tasks such as speaker identification, speaker verification, and SER [22].

In this work, we extend the method for correlation pooling presented in [22]. In [22], it was shown that channel-wise correlations provide an alternative way of extracting speaker and emotion information from self-supervised models, providing also improvements over the standard mean and mean-std pooling (std stands for standard deviation). In the present work, we add an attention mechanism to pool representations before estimating the correlation matrix, while reducing the confidence on the target labels with label smoothing.

3. PROPOSED METHOD

We construct our SER framework based on the pipeline and principles of SUPERB [9]. As finetuning pretrained models has a high resource demand in terms of the computational power needed, we use a simple framework with a frozen pretrained model and lightweight classification heads. An
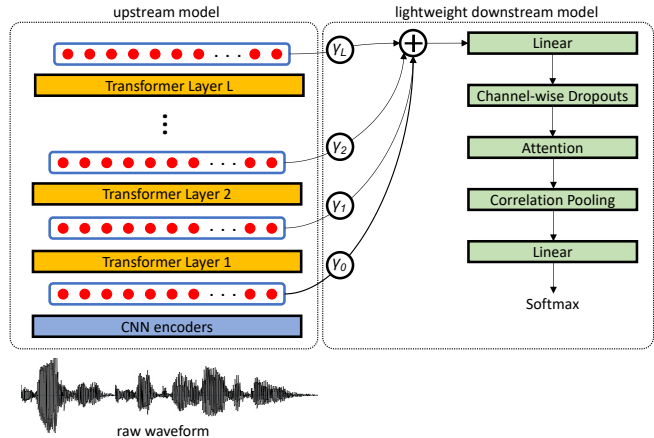

Fig. 1. Overview of the proposed architecture.

overview of our framework is presented in Fig. 1. In this section we describe details of the proposed approach.

3.1. Layer-wise pooling

We extract information relevant for our downstream task as in SUPERB ([9]), by taking representations from all transformer layers in the model and collapsing them to one via a weighted average (see Fig. 1). There is one weight for each layer (a total of $L + 1$) and all weights are trained jointly with the classification network. The weighted average representation is expressed as follows

$$\mathbf{h}_t = \sum_{l=0}^{L} \gamma_l \mathbf{h}_{t,l}, \qquad (1)$$

where the weights $\sum_{l=0}^{L} \gamma_l = 1$, $\gamma_l \geq 0$ are implemented with a learnable vector of size $L+1$, followed by the Softmax function, and $\mathbf{h}_{t,l}$ is the representation of the lth layer at time t ($\mathbf{h}_{t,0}$ is the output of the Convolutional Network – ConvNet).

3.2. Frame-wise pooling

Tasks requiring sentence-level classification typically employ a pooling method, such as mean, max or attentive pooling. Mean pooling, which is employed in SUPERB is defined as

$$\mathbf{r} = \bar{\mathbf{h}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t, \qquad (2)$$

where T is the number of acoustic features of an utterance extracted by the ConvNet, $\mathbf{r}$ is the resulting pooled representation, while $\mathbf{h}_t$ are the representations at time t after layer-wise pooling. Concatenating the pooled representations with std features is in general helpful in speaker recognition [23], and is implemented as

$$\mathbf{r} = \left[\bar{\mathbf{h}}; \left(\frac{1}{T} \sum_{t=1}^{T} (\mathbf{h}_t - \bar{\mathbf{h}})^2 \right)^{1/2} \right], \qquad (3)$$

where $[\cdot; \cdot]$ denotes vector concatenation and the exponents should be considered as element-wise operators.

3.3. Correlation pooling

Correlation pooling (introduced in [21]) is an alternative pooling method which has shown improvements in speaker recognition. The embedding dimension of SSL models (typically 768 or 1024) is too high to estimate correlations. We therefore project $\mathbf{h}$ onto a lower d_v-dimensional space $\mathbf{v}$ via a linear layer ($d_v = 256$). We then calculate the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{v}$ as follows

$$\boldsymbol{\mu} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{v}_t, \ \ \boldsymbol{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{v}_t - \boldsymbol{\mu})(\mathbf{v}_t - \boldsymbol{\mu})', \quad (4)$$

where $\mathbf{x}'$ denotes the transpose of $\mathbf{x}$. Finally, the correlation matrix is derived by normalizing with respect to the variances as follows

$$\mathbf{C} = \boldsymbol{\Sigma} \oslash \mathbf{S}, \quad (5)$$

where $\mathbf{S} = \mathbf{s}\mathbf{s}' + \epsilon\mathbf{1}$, $\mathbf{s} = \text{diag}(\boldsymbol{\Sigma})^{1/2}$ (i.e. the vector of std), $\oslash$ denotes element-wise division, while $\epsilon = 10^{-8}$. Since $\mathbf{C}$ is a symmetric matrix and its diagonal elements are equal to 1, we vectorize the elements above the diagonal, yielding a $(d_v \times (d_v - 1)/2)$-sized vector, which we project onto a linear layer followed by the Softmax over the emotion classes. For regularization, dropout is applied to $\mathbf{v}$, where whole channels are dropped with probability $p_d = 0.25$.

3.4. Attentive correlation pooling

We introduce here the attentive correlation pooling, by inserting weights in the estimates of the statistics, i.e.

$$\boldsymbol{\mu} = \sum_{t=1}^{T}w_t\mathbf{v}_t, \ \ \boldsymbol{\Sigma} = \sum_{t=1}^{T}w_t(\mathbf{v}_t - \boldsymbol{\mu})(\mathbf{v}_t - \boldsymbol{\mu})' \quad (6)$$

and we calculate $\mathbf{C}$ as in eq. (5). The weights $\{w_t\}_{t=1}^{T}$ (where $\sum_t w_t = 1$ and $w_t \geq 0$) are estimated using a new flavor of attention. Similarly to the single-head attention a single set of weights is estimated. Similarly to the multi-head attention, multiple heads are employed, however their similarities with $\mathbf{v}_t$ are aggregated prior to the Softmax function via the log-sum-exp function, as follows

$$\{w_t\}_{t=1}^{T} = \text{Softmax}\left(\{a_t\}_{t=1}^{T}\right), \quad (7)$$

where

$$a_t = \log\sum_{h}\exp\left(\mathbf{q}_h'\mathbf{o}_t + b_h\right), \quad (8)$$

the heads $\{\mathbf{q}_h, b_h\}_{h=1}^{H}$ are trainable d_v-dimensional vectors and biases, $\mathbf{o}_t = \text{ReLU}(\mathbf{W}_{att}\mathbf{v}_t)$ and $\mathbf{W}_{att}$ is a square matrix ($d_v \times d_v$). Note that an equivalent implementation is to use as input to the Softmax the $(H \times T)$-sized *vector* of dot-products $\mathbf{q}_h'\mathbf{o}_t + b_h$ and sum the outputs over heads to obtain $\{w_t\}_{t=1}^{T}$.

As we observe, the proposed attention resembles a mixture model with heads parametrizing the mixture components.

The log-sum-exp function is a soft version of the max operator, meaning that a_t is high when at least one of the H head-specific dot-products $\{\mathbf{q}_h'\mathbf{o}_t + b_h\}_{h=1}^{H}$ is high.

The rationale for proposing this kind of attention is two-hold. We desire to keep the multi-modality of multi-head attention since a single head is too weak to capture the phonetic, speaker, emotion and channel variability. On the other hand, the standard multi-head attention results in H context-vectors (in our case H correlation matrices), which can be hard to estimate robustly, especially when the utterances are short and the estimation involves second-order statistics.

3.5. Label smoothing

With label smoothing we soften the hard (one-hot) targets vectors $\mathbf{y}$ of the training set as follows

$$\mathbf{y}^{LS} = \mathbf{y}(1 - p_l) + p_l/K, \quad (9)$$

where p_l is the label smoothing parameter (i.e. the probability mass equally distributed to all classes) and $\mathbf{y}^{LS}$ is the smoothed target vector [24]. Cross-entropy is still employed as loss function, but with soft targets.

4. EXPERIMENTS

4.1. Datasets

The IEMOCAP database consists of multi-modal recordings (speech, video) by 10 actors in dyadic sessions in English ($\approx$ 12 hours) [18]. The dataset is split in 5 dialogue sessions (one female-male speaker pair per session). The emotions conveyed are happiness, anger, excitement, sadness, surprise, fear, frustration, and neutral state. As in other studies on IEMOCAP, we relabel excitement as happiness and use 4 balanced emotion classes, namely: anger, happiness, sadness, and neutral [9, 5, 25]. All other classes are discarded.

4.2. Experimental Setup

We use a 5-fold cross-validation setup where at each fold we leave out one session from the dataset. Each held-out session consists of two speakers that are not present in the train and validation sets. This approach leaves approximately 19% of the data for testing. Mean and standard deviation across folds is computed and presented as the aggregated result. Our SER framework is evaluated with WavLM, Wav2vec 2.0, and HuBERT speech representations.

5. RESULTS AND ANALYSIS

An overview of the results for the most common pooling methods and our proposed approach is shown in Table 1. The results are presented for the best configuration of our framework with p_d and p_l both equal to 0.25 and $H = 4$. The best overall performance was achieved for our proposed approach using WavLM (75.60%; see also Fig. 2) which is higher compared to other SSL approaches in the literature on the same data — 67.20% with frozen Wav2vec 2.0 [5], 67.62% with frozen HuBERT [9], 73.01% with fine-tuned HuBERT [26], and 69.08% with fine-tuned Wav2vec 2.0 model [26].

Table 1. Unweighted accuracy (% mean and std) between test sets for SER in IEMOCAP using HuBERT large, Wav2vec 2.0, and WavLM large self-supervised representations.

Pooling method	HuBERT	Wav2vec 2.0	WavLM
mean	65.73 (2.73)	66.86 (1.76)	69.44 (1.53)
mean-std	69.15 (1.61)	69.92 (1.17)	72.56 (1.67)
corr ($p_d = 0$)	69.82 (1.35)	68.44 (1.85)	72.34 (1.54)
corr ($p_d = 0.25$)	69.72 (1.19)	67.85 (1.84)	72.27 (1.45)
corr attentive	**73.86 (2.10)**	**70.01 (2.20)**	**75.60 (2.33)**

Our method also yields results that surpass those on the benchmark setup of SUPERB on the same dataset and splits; SUPERB is a leaderboard that benchmarks the performance of a shared model across a range of speech processing tasks [9]. In particular, SUPERB reports an accuracy of 70.62% with WavLM and 67.62% with HuBERT whereas we obtain 75.60% and 73.86% respectively. For comparison, using standard acoustic features, the reported performance is considerably lower: 52.4% using eGeMAPS [5], 49.8% with spectrograms [5], 57.6% using filter banks [27], and 54.6% using MFCCs [28].

5.1. Attention

To investigate the performance of the proposed attentive correlation pooling method we experimented with different numbers of attention heads (H). Specifically, we tested for $H = 1, 4, 16, 32$ heads. We obtained the best result with $H = 4$. Note that we did not observe any correlation between heads and emotion classes, meaning that there is no direct mapping between heads and emotions. In particular, for Session 1, HuBERT, $p_d = 0.25$, $p_l = 0.25$ accuracy was 70.32% ($H = 1$), 73.18% ($H = 4$), 72.53% ($H = 16$), and 71.34% ($H = 32$).

5.2. Label smoothing and dropout

To better understand the performance of our system, we probed our setup by varying the parameters for label smoothing and dropout rate. For a set number of attention heads ($H = 4$), label smoothing was varied for $p_l = 0, 0.15, 0.25, 0.3$ and dropout rate for $p_d = 0...0.5$. The effect was evaluated on Session 1 of the setup for attentive correlation pooling

Fig. 2. Confusion matrix for WavLM using attentive correlation (left) and meanstd pooling (right).
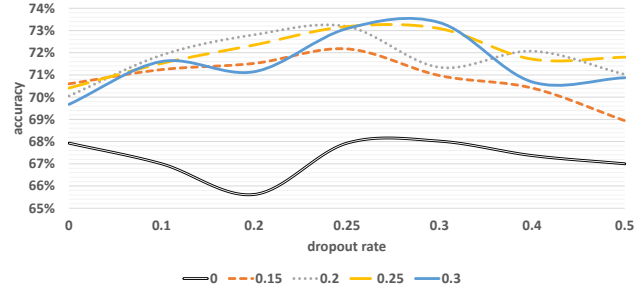

Fig. 3. Dropout-label smoothing interaction for Session 1 - HuBERT. Lines represent different label smoothing values.

using HuBERT and the results can be seen in Fig. 3. Both dropout and label smoothing have an impact on the performance with label smoothing having a greater positive effect; increasing performance from 67.93% ($p_d = 0, p_l = 0$) to 70,41% ($p_d = 0, p_l = 0.25$) and even further with increasing dropout to 73.18% ($p_d = 0.25, p_l = 0.25$).

The impact of dropout rate and label smoothing was also investigated for mean, mean-std, and correlation pooling. The impact in all was small to negligible. For example, for $p_d = 0, p_l = 0.25$ the performance remained unchanged compared to $p_d = 0, p_l = 0$ for mean, mean-std, and correlation pooling while for $p_d = 0.25, p_l = 0.25$ there was a small improvement for correlation pooling (from 68.20% to 70.60%).

6. CONCLUSIONS

In this work we presented an SER framework that uses self-supervised representations and is based on label smoothing and a novel approach to attention, attentive correlation pooling. Notably, our method does not require fine-tuning of the pre-trained SSL models but rather uses a light-weight classification head that attempts to capture all relevant emotion information from the pre-trained representations. We run several experiments using a 5-fold cross-validation setup and we have clearly demonstrated that our method reaches high performance in all pre-trained models tested surpassing that of the literature in similar tasks. In future work, we will extend the evaluation setup and validate the performance of our method on more datasets.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] Scott Brave and Cliff Nass, "Emotion in human-computer interaction," in *The human-computer interaction handbook*, pp. 103–118. CRC Press, 2007.

[2] Chul Min Lee and Shrikanth S Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE transactions on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[3] Rosalind W Picard, *Affective computing*, MIT press, 2000.

[4] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[5] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[6] Wei-Ning Hsu, Benjamin Bolte, et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] Alexei Baevski, Yuhao Zhou, et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[8] Sanyuan Chen, Chengyi Wang, et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *arXiv preprint arXiv:2110.13900*, 2021.

[9] Shu-wen Yang, Po-Han Chi, et al., "SUPERB: Speech processing universal performance benchmark," in *Proceedings of Interspeech*, 2021.

[10] Mousmita Sarma, Pegah Ghahremani, et al., "Emotion identification from raw speech signals using dnns.," in *Interspeech*, 2018, pp. 3097–3101.

[11] Yang Liu, Haoqin Sun, et al., "Discriminative feature representation based on cascaded attention network with adversarial joint loss for speech emotion recognition," *Proc. Interspeech 2022*, pp. 4750–4754, 2022.

[12] Jinkyu Lee and Ivan Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, 2015.

[13] Yelin Kim and Emily Mower Provost, "Leveraging inter-rater agreement for audio-visual emotion recognition," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 553–559.

[14] Emily Mower, Angeliki Metallinou, et al., "Interpreting ambiguous emotional expressions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–8.

[15] Jiawang Liu and Haoxiang Wang, "A speech emotion recognition framework for better discrimination of confusions," in *Interspeech*, 2021, pp. 4483–4487.

[16] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 566–570.

[17] Lorenzo Tarantino, Philip N Garner, et al., "Self-attention for speech emotion recognition.," in *Interspeech*, 2019, pp. 2578–2582.

[18] Carlos Busso, Murtaza Bulut, et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[19] Leon A Gatys, Alexander S Ecker, and Matthias Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[20] Jan Chorowski, Ron J Weiss, et al., "On using backpropagation for speech texture generation and voice conversion," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2256–2260.

[21] Themos Stafylakis, Johan Rohdin, and Lukas Burget, "Speaker embeddings by modeling channel-wise correlations," in *Interspeech*, 2021.

[22] Themos Stafylakis, Ladislav Mošner, et al., "Extracting speaker and emotion information from self-supervised speech models via channel-wise correlations," in *2023 IEEE Workshop on Spoken Language Technology (SLT)*, to appear, pp. 1–8.

[23] Shuai Wang, Yexin Yang, et al., "Revisiting the statistics pooling layer in deep speaker embedding learning," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.

[24] Christian Szegedy, Vincent Vanhoucke, et al., "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[25] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[26] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," 2021.

[27] Edmilson Morais, Ron Hoory, et al., "Speech emotion recognition using self-supervised features," in *ICASSP 2022*. IEEE, 2022, pp. 6922–6926.

[28] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.