INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Detection of Emotional Hotspots in Meetings Using a Cross-Corpus Approach

*Georg Stemmer*[1], *Paulo Lopez Meyer*[2], *Juan Del Hoyo Ontiveros*[3], *Jose A. Lopez*[4], *Hector A. Cordourier Maruri*[5], *Tobias Bocklet*[6]

[1,6]Intel Corp., Germany
[2,3,5]Intel Corp., Mexico
[4]Intel Corp., USA

firstname.lastname@intel.com

## Abstract

Speech emotion recognition for natural human-to-human conversations has many useful applications, including generating comprehensive meeting transcripts or detecting communication problems. We investigate the detection of emotional hotspots, i.e., regions of increased speaker involvement in technical meetings. As there is a scarcity of annotated, not-acted corpora, and to avoid introducing unwanted biases to our models, we follow a cross-corpus approach where models are trained on data from domains unrelated to the test data. In this work we propose a model ensemble trained on spontaneous phone conversations, political discussions and acted emotions. Evaluation is performed on the natural ICSI and AMI meeting corpora, where we used existing hotspot annotations for ICSI and created labels for the AMI corpus. A semi-supervised fine-tuning procedure is introduced to adapt the model. We show that an equal error rate of below 21% can be achieved using the proposed cross-corpus approach.

**Index Terms**: emotion recognition, human-computer interaction, computational paralinguistics.

## 1. Introduction

Speech does not only convey words but also paralinguistic information, e.g., a speaker's emotional state. This information typically gets lost in the transcript of a conversation. A computer's ability to understand a human conversation, or to communicate with humans, could be significantly improved by enriching transcripts using speech emotion recognition (SER). Various practical applications can be imagined, e.g., more comprehensive meeting summaries, detecting communication problems, or improved human-computer interaction. Even though SER has been studied for decades [1], very few such applications exist today. We believe that one of the reasons is a scarcity of research on SER for natural, not-acted data, where the experimental setup has been designed to match an actual application.

The goal of the work described in this paper was to create a baseline system for enriching transcripts of technical meetings with emotional information extracted from the speech signal. More specifically, the system detects emotional hotspots which are defined as regions of increased speaker involvement [2].

We evaluate the accuracy of the proposed system on two popular corpora, the ICSI Meeting Corpus [3], and the AMI Meeting Corpus [4], created by recording and annotating project meetings at research institutes in US and Europe. Due to the realistic setup when collecting the datasets, emotional events, including hotspots, occur relatively rarely. The sparseness of events, combined with the limited size of the corpora, increases the risk of overfitting a model if trained on the meeting corpora themselves. The models may learn invisible biases, e.g., caused by the correlation of speakers or discussion topics with emotional events. To ensure that evaluation results reflect a model's ability to detect emotional events, and nothing else, we follow a cross-corpus approach where the test corpora are not used for training. According to [2], hotspots are related to the emotional dimension of 'activation', or 'arousal'. Therefore, our approach is based on a set of models trained for different types of labels but generate scores that are expected to correlate with the arousal of a speaker. Hotspot detection is performed by combining models developed and published by others for conflict intensity estimation [5], laughter detection [6], and emotion classification [7]. These models have been trained on political discussions from the SSPNet Conflict Corpus [8], telephony conversations from the Switchboard corpus [9], and finetuned for acted emotions contained in the IEMOCAP data set [10]. Models are combined using a linear interpolation of scores, where the weights are optimized on a development partition of the ICSI meeting corpus. We also investigate the adaptation of the conflict intensity estimation model.

### 1.1. Related work

For a more detailed acoustic-prosodic analysis of the hotspots in the ICSI meeting corpus, we refer the reader to [2]. An early work for automatic hotspot detection in the ICSI meeting corpus has been described by Laskowski [11]. The author leveraged the available reference speech and laughter segmentations to extract low-level vocal activity features, which are fed into a classifier. More recently, automatic hotspot detection for the ICSI meeting corpus has been investigated by Makhervaks et al. in [12]. Similar to our work, a score combination is proposed to detect the hotspots. Prosodic features are extracted from the speech signal and combined with speaker activity, lexical, and laughter count features. This is in contrast to our work where we do not use any manual transcription, lexical or prosodic feature extraction, or even use the ICSI meeting corpus for training. The authors [12] mention that manual transcriptions and the lack of validation on an independent meeting corpus make it difficult to judge whether the performance figures reported could be achieved in a real application.

Applying a cross-corpus approach to address the issue of generalization in SER has already been proposed by others, see, e.g., [13] for one of the earliest works in this area. A recent literature review on this subject has been published by Zhang et al. in [14]. Most of the results are for acted emotions, which are difficult to transfer to real-world applications [1]. To best of our knowledge, hotspot detection in meetings has not been explored using a cross-corpus approach. In cross-corpus SER, different methods have been proposed to reduce the mismatch between training and test corpora. One popular research di-

10.21437/Interspeech.2023-1023

rection is to improve the generalization of the models, e.g., by multi-task learning with large amounts of unlabeled data [15]. Other approaches leverage transfer learning to improve models pre-trained on mismatched data, see e.g., [16]. The semi-supervised adaptation procedure described in this paper is comparable to the latter, except that we had to adjust the method to deal with the reference labels' sparsity by including the scores of the pre-trained model.

### 1.2. Contributions

The main contributions of the paper are the following: To our knowledge, this paper is the first to report hotspot detection results for ICSI and AMI meeting corpora using a cross-corpus approach. It demonstrates that robust hotspot detection on natural, realistic data can be performed with a simple shallow ensemble of various models trained on highly mismatched data. In addition, we introduce an adaptation method for the conflict estimation model that is robust to the highly imbalanced label distribution. Finally, we provide a new labeling of a subset of the AMI meeting corpus.

## 2. Data

### 2.1. ICSI Meeting Corpus

The ICSI meeting corpus [3] is a collection of 75 natural, i.e., not-acted, meetings at the ICSI research institute, totaling about 72 hours (including silence). All meetings were in the English language. A significant proportion of speakers is non-native. Further details on speaker demographics, recordings setup, transcription can be found in [3]. Here we focus on how the data has been prepared for our experiments. The meetings have been recorded with desktop and headset microphones; each headset channel corresponds to a speaker. For each channel, an utterance segmentation is provided. Hotspot labels are provided for each speaker and each segment. Hotspots can have different intensity levels (hotness): the three primary levels are 'lukewarm', 'warm', and 'hot'. Each label contains additional information, e.g., valence, which is ignored here. Details can be found in [2, 17].

As we found the audio quality of the desktop microphone channels unacceptable, we mixed all headset microphone channels into a single channel. The following procedure ensures no contradicting labels when mapping the speaker-specific hotspot labels to the mixed channel: All hotspot segments that overlap in time get merged. If the intensity level of those segments differs, the segment is assigned the 'hotter' label. Segments that overlap with a hotspot, but are not a hotspot, are removed. Overlap is defined as having overlapping start/end times or a gap of less than 0.5 seconds. As concise segments often corresponded to noises, we removed all segments with a duration of $< 1$ second. The average duration of the remaining segments is 10 seconds.

Note that this data preparation procedure differs from the one described in [12]. Makhervaks et al. mapped the hotspot labels to fixed-length segments, which introduces partial overlaps with the original utterance-level segmentation, and may result in segments being labeled as hotspots, but missing their acoustic realization (e.g., short laughter).

The merged segments are split into two partitions, a development set containing all segments from 24 meetings and a test set containing the remaining 51 meetings. There is no train set as we do not train models on the corpora used for evaluation. Tab. 1 provides all details on the label distributions in the resulting database. Clearly, the hotspot distribution is highly skewed - less than 9% of all segments have been labeled as hotspots, and just 1% have been marked as 'warm' or 'hot'.

Table 1: *Label distributions for the prepared ICSI data.*

| | **Dev.** | **Test** | **All** |
|---:|---:|---:|---:|
| # Meetings | 24 | 51 | 75 |
| # Segments | 6593 | 14304 | 20897 |
| # 'hot' hotspots | 0 | 8 | 8 |
| # 'warm' hotspots | 44 | 199 | 243 |
| # 'lukewarm' hotspots | 464 | 1135 | 1599 |
| Total duration [h] | 18.9 | 39.0 | 57.9 |

Listening to various segment files marked as hotspots led us to conclude that we could not distinguish 'lukewarm' hotspots from segments not marked as hotspots. We focused only on detecting segments labeled as either 'warm' or 'hot' and removed the 'lukewarm' labels from the test set. We kept the 'lukewarm' in the development set, however, for robust parameter estimation.

### 2.2. AMI Meeting Corpus

The AMI meeting corpus [4] consists of 100 hours of transcribed meetings in the English language recorded at different research institutes. The original recording has been multi-modal and with various scenarios; for the experiments in this paper we only use the audio signals of the 'headsetmix' signals of 26 natural meetings recorded in Edinburgh and Idiap. As no hotspot labels were available for this corpus, we cut the 26 meetings, which total 23 hours, into 2735 non-overlapping 30-second long segments. The segments were distributed to seven raters, each assigned to three different raters. The raters were asked to listen to each segment at least once and then classify it as either containing 'no hotspot', 'neutral hotspot (clarification, suggestion)', 'negative hotspot (conflict, disagreement)' or 'positive hotspot (amusement, agreement)'. This effort is still ongoing. At the time of writing, we had obtained more than 5000 ratings. For 433 segments, we had three ratings from five raters that each had labeled more than 100 segments, allowing us to do a majority voting. If two or three raters labeled a segment as a hotspot of any type, it was labeled as 'hotspot'; otherwise, it received a 'no hotspot' label. The pairwise inter-rater agreements for the five raters, measured using Krippendorff's alpha [18], are shown in Tab. 2. The average agreement is 0.48, which is an acceptable range for such a highly subjective task.

Table 2: *Pairwise inter-rater agreement for five raters of the AMI corpus. 'n/a' indicates that the number of the segments labeled by both raters was below 100.*

| **Rater ID** | **6** | **1** | **3** | **2** | **4** |
|---:|---:|---:|---:|---:|---:|
| **6** | 1.00 | 0.61 | $n/a$ | $n/a$ | 0.42 |
| **1** | 0.61 | 1.00 | 0.52 | 0.31 | 0.51 |
| **3** | $n/a$ | 0.52 | 1.00 | 0.49 | 0.49 |
| **2** | $n/a$ | 0.31 | 0.49 | 1.00 | $n/a$ |
| **4** | 0.42 | 0.51 | 0.49 | $n/a$ | 1.00 |

Because the labeled subset of the AMI corpus is quite small, we did not split it into development and test partitions. Consequently, no parameters, like interpolation weights or other model parameters, are optimized on the AMI corpus. However, the same values optimized for the development partition of ICSI are applied when testing with the AMI corpus. Thus, AMI is used as an independent test corpus. The corresponding label distributions are shown in Tab. 3. Interestingly, we have a much higher percentage of hotspots in the AMI corpus than in the ICSI corpus; 39% of all segments receive a hotspot label. This may be due to the raters having a higher sensitivity for hotspots or other reasons, like the AMI meeting participants being generally more 'involved'. We plan to investigate this as soon as we have completed labeling a more significant proportion of the corpus.

Table 3: *Label distributions for the prepared AMI data.*

| | **Test** |
|---:|---:|
| # Meetings | 26 |
| # Segments | 433 |
| # hotspots | 169 |
| Total duration [h] | 3.6 |

### 2.3. Training Corpora

The conflict intensity estimation model has been trained on the SSPNET Conflict Corpus [8]. The corpus comprises 12 hours of French political debates, recorded on television. The recordings were cut into 30-second long segments and assigned a conflict intensity score by averaging the estimates provided by ten different raters who were not speaking French.

The laughter detection model has been trained on the Switchboard Corpus [9], a collection of 240 hours of transcribed, spontaneous telephony conversations in English. Each conversation is about five minutes long. The transcripts contain laughter annotations including detailed start and end times.

The emotion classification model has been pretrained on the Libri-Light [19] corpus, a collection of 60k hours of audiobooks in English. For finetuning, the IEMOCAP dataset [10] has been employed. IEMOCAP contains about 12 hours of read and spontaneous English speech from ten actors. The acted emotions have been annotated with ten different categorial labels. For creating the emotion classification model the popular setup of just four emotion classes as targets has been deployed: sadness, happiness+excitement, anger, and neutral.

## 3. Experimental Setup

ICSI and AMI meeting corpora are not used for model training. Instead, we re-use three different models developed and published by others for other data: A conflict intensity estimator, a laughter detector, and an emotion classifier. The following sections provide a short overview of these models, focusing on how we applied them to the meeting datasets.

### 3.1. Conflict Intensity Estimation

Conflict detection in human communication involves identifying speech with opposite or negative verbal cues [20]. Rajan et al. [5] introduced a convolutional-recurrent architecture with an attention mechanism for estimating the intensity of conflict in a conversation. The work achieved state-of-the-art results for the SSPNet Conflict Corpus at the time of publication. The end-to-end network, which they call ConflictNET, directly processes a 30-second segment of speech samples. Feature extraction is replaced by several convolutional layers which learn the relevant speech properties from the training data. An attention layer enables the network to learn acoustic events which do not span the whole input segment. The network is trained to maximize the Pearson Correlation Coefficient [21], i.e., a linear correlation between the single output neuron and a conflict intensity measure that, after scaling, ranges between $-1$ and $+1$. We leverage the code published by the authors[1]. The following changes were applied to the code: First, the input signal is not downsampled but the full 16 kHz sampling rate is used, doubling the input segment size. Second, the input signal's Root Mean Square (RMS) amplitude is not normalized anymore. Third, when applying ConflictNet to the ICSI corpus, segments with a length of less than 30 seconds are not zero-padded but repeated until the whole input segment is filled with audio samples.

### 3.2. Laughter Detection

For laughter detection we applied a model published by Gillick et al. [6]. It is based on a ResNet-18 architecture which processes 128-dimensional Mel spectrograms as features. The model has been trained on data from the Switchboard corpus, which contains annotations for laughter. We directly downloaded the model trained and shared by the authors[2]. We processed all segments of the meeting corpora using the default parameter settings, i.e., a detection threshold of $0.5$ and a minimum laughter length of $0.2$ seconds. Preliminary experiments indicated that there is some potential in tuning these parameters for our test corpora, which is something we are going to investigate in the future.

### 3.3. Emotion Classification

For the emotion classification we downloaded a pretrained large model with HuBERT topology [22] from Huggingface[3] [23]. The model has been fine-tuned over the 4-class (neutral, sad, angry, happy) IEMOCAP dataset with a leave-one-session-out training strategy. The development of this component leveraged the s3prl open source toolkit[4] [7].

### 3.4. Evaluation Measures

As we deal with a detection problem, we evaluate a model by computing False Accept Rates (FAR) and False Rejection Rates (FRR) for a reasonable range of detection thresholds. FRR represents the proportion of false negatives, i.e., segments in a test set that are scored below the threshold but are a hotspot. FAR is the proportion of false positives, i.e., segments scored equal to or above the threshold but are not a hotspot. Equal Error Rate (EER) is the error rate where FAR and FRR are equal. The lower the EER, the better the system. We also report the Unweighted Average Recall (UAR), which is the average of the proportions of the true positives, i.e., hotspot segments that are scored above the threshold, and true negatives, i.e., segments that are scored below the threshold and which are not a hotspot. The higher the UAR, the better the system. The UAR reported for a given model is always the highest UAR that could be achieved for all detection thresholds.

[1] https://github.com/smartcameras/ConflictNET

[2] https://github.com/jrgillick/laughter-detection

[3] https://huggingface.co/facebook/hubert-large-ll60k

[4] https://github.com/s3prl/s3prl/tree/main/s3prl/downstream/emotion

### 3.5. Model Ensemble

If we consider each of the models for conflict intensity estimation, laughter detection, and emotion classification as source of useful information about the arousal level of an utterance, it makes sense to try to combine them to get improved results. In preliminary experiments we tried different ways to combine the models. Ultimately, it worked best to normalize each model's scores by subtracting a model-specific mean and dividing by a model-specific standard deviation. The normalized scores of the different models are then combined by linear interpolation. The optimal interpolation weights are determined on the ICSI development set using an exhaustive search that tries out all interpolation weights between $-1.0$ and $1.0$ in steps of $0.25$. For six different scores $9^6 \approx 0.5$M combinations are evaluated, and the one that leads to the highest UAR is selected. Our experiments showed that many weight combinations lead to very similar UARs. We also tried minimizing EER instead of maximizing the UAR, which made no difference.

### 3.6. Semi-Supervised Model Adaptation

We investigated whether adapting the ConflictNET model to the ICSI development set can improve results. As the hotspots rarely occur, simply re-training or finetuning using existing labels impairs the model. Therefore we first scored the ICSI development set with the pre-trained ConflictNET model. Then we heuristically modified the scores depending on the labels. If a segment was labeled as 'lukewarm', we increased the score by 1. For segments labeled 'warm', we increased the score by 2. Finally, for segments labeled 'hot', we increased by 3. We then finetuned the ConflictNET model on the modified scores. This semi-supervised method ensures the model does not deviate significantly from the pre-training, but can adapt to the hotspots in the ICSI development set.

## 4. Results and Discussion

The results for the different models, individually and in combination, are shown in Tab. 4.

Table 4: *Results on ICSI and AMI testsets.*

| | ICSI Test | | AMI Test | |
|---|---|---|---|---|
| | EER | UAR | EER | UAR |
| Model comb. | [%] | [%] | [%] | [%] |
| Conflict | 25.6 | 75.1 | 30.7 | 70.4 |
| Laughter | 35.8 | 64.2 | 31.3 | 68.7 |
| Happy | 31.9 | 69.0 | 43.2 | 60.6 |
| Sad | 59.8 | 50.0 | 52.8 | 50.0 |
| Angry | 45.5 | 55.2 | 50.9 | 52.2 |
| Neutral | 65.1 | 50.0 | 56.1 | 50.0 |
| All | 23.4 | 77.3 | 29.0 | 74.8 |
| Confl. adapt. to ICSI Dev. | 22.0 | 79.2 | 31.4 | 69.1 |
| All with adapt. Confl. | 20.9 | 79.5 | 29.6 | 74.1 |

The conflict intensity score has the lowest EER from all individual scores, with laughter being the second best. From the emotion classes, only 'happy' seems to provide helpful information for hotspot detection. The combination of all scores provides the lowest EERs on both test corpora. Adapting the conflict intensity estimation to the ICSI development set helps for the ICSI testset but does not seem to reduce the EER on the AMI testset – this indicates that the ConflictNET model may learn corpus-specific information during adaptation, e.g. speakers. On the other hand, the adapted model does not perform much worse on the AMI corpus indicating that overfitting could be avoided.

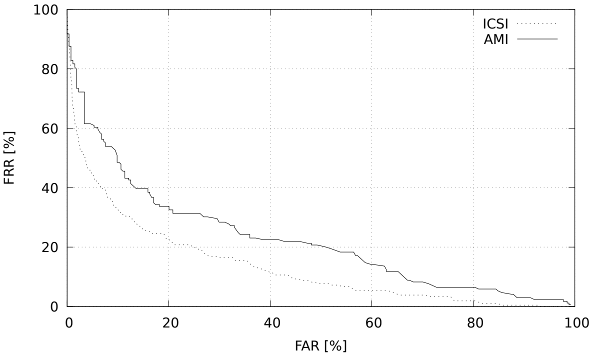

Figure 1: *DET for the best model ensemble on ICSI and AMI.*

When considering the proposed approach for an actual application, one is interested mainly in the FAR vs. FRR at a specific detection threshold. The Detection Error Tradeoff (DET) curve in Fig. 1 is the best combined model from the last row of Tab. 4. For a FAR of around 10%, the system misses about 30% of the ICSI hotspots and about 45% of the AMI hotspots, which means that the error rate is still too high for most real-world applications.

For a more detailed analysis the confusion table Tab. 5 is presented, which evaluates the recognition rates for the original labels in the ICSI test set at the detection threshold that minimizes EER. As expected from listening to the files, the 'lukewarm' hotspots, equally classified as 'no hotspot' or 'hotspot', are a significant source of errors.

Table 5: *Example confusion table for the best ICSI system.*

| Reference | Original label | Hotspot Detected [%] | |
|---|---|---|---|
| | | no | yes |
| no hotspot | none | 81.6 | 18.4 |
| | 'lukewarm' | 49.3 | 50.7 |
| hotspot | 'warm' | 21.1 | 78.9 |
| | 'hot' | 12.5 | 87.5 |

## 5. Conclusions and Next Steps

We have shown that a cross-corpus approach is effective in detecting emotional hotspots in natural meetings. The results are surprisingly good, given the high training and test data mismatch. However, error rates are too high for an actual application. For future research, we will finalize the labeling of the AMI corpus to confirm our findings and publish the labels. We will start applying methods that increase generalization, e.g., ladder networks [15], and plan to include features derived from an automatic speech recognizer.

## 6. References

[1] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, *The Automatic Recognition of Emotions in Speech.* Berlin, Heidelberg:

Springer Berlin Heidelberg, 2011, pp. 71–99. [Online]. Available: https://doi.org/10.1007/978-3-642-15184-2_6

[2] B. Wrede and E. Shriberg, “Spotting ”hot spots” in meetings: Human judgments and prosodic cues,” in *INTERSPEECH*, 2003.

[3] 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 *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03).*, vol. 1, 2003, pp. I–I.

[4] I. Mccowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska Masson, W. Post, D. Reidsma, and P. Wellner, “The ami meeting corpus,” *Int’l. Conf. on Methods and Techniques in Behavioral Research*, 01 2005.

[5] V. Rajan, A. Brutti, and A. Cavallaro, “Conflictnet: End-to-end learning for speech-based conflict intensity estimation,” *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1668–1672, 2019.

[6] J. Gillick, W. Deng, K. Ryokai, and D. Bamman, “Robust laughter detection in noisy environments,” in *Interspeech 2021*, 08 2021, pp. 2481–2485.

[7] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[8] M. Filippone, S. Kim, F. Valente, and A. Vinciarelli, “Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes,” in *MM 2012 - Proceedings of the 20th ACM International Conference on Multimedia*, 10 2012.

[9] J. Godfrey, E. Holliman, and J. McDaniel, “Switchboard: telephone speech corpus for research and development,” in *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520 vol.1.

[10] C. Busso, M. Bulut, C.-C. Lee, E. A. Kazemzadeh, E. M. Provost, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[11] K. Laskowski, “Modeling vocal interaction for text-independent detection of involvement hotspots in multi-party meetings,” in *2008 IEEE Spoken Language Technology Workshop*, 2008, pp. 81–84.

[12] D. Makhervaks, W. Hinthorn, D. Dimitriadis, and A. Stolcke, “Combining acoustics, content and interaction features to find hot spots in meetings,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8054–8058.

[13] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, “Cross-corpus acoustic emotion recognition: Variances and strategies,” *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[14] T. X. Z. X. Zhang S, Liu R, “Deep cross-corpus speech emotion recognition: Recent advances and perspectives.” *Frontiers in Neurorobotics*, November 2021.

[15] S. Parthasarathy and C. Busso, “Semi-supervised speech emotion recognition with ladder networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, 2020.

[16] R. Milner, M. A. Jalal, R. W. M. Ng, and T. Hain, “A cross-corpus study on speech emotion recognition,” in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 304–311.

[17] B. Wrede, S. Bhagat, R. Dhillon, and E. Shriberg, “Meeting recorder project: Hot spot labeling guide, ICSI technical report TR-05-004,” Tech. Rep., 2005.

[18] J. Eggink, “Krippendorff’s alpha,” 2023, MATLAB Central File Exchange. Retrieved March 4, 2023. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/36016-krippendorff-s-alpha

[19] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673, https://github.com/facebookresearch/libri-light.

[20] M.-J. Caraty and C. Montacié, *Detecting Speech Interruptions for Automatic Conflict Detection*. Springer International Publishing, 2015, pp. 377–401. [Online]. Available: https://doi.org/10.1007/978-3-319-14081-0_18

[21] W. Kirch, Ed., *Pearson’s Correlation Coefficient*. Dordrecht: Springer Netherlands, 2008, pp. 1090–1091.

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “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.

[23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6