ROLE OF LEXICAL BOUNDARY INFORMATION IN CHUNK-LEVEL SEGMENTATION FOR SPEECH EMOTION RECOGNITION

Wei-Cheng Lin and Carlos Busso

Multimodal Signal Processing (MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

wei-cheng.lin@utdallas.edu, busso@utdallas.edu

ABSTRACT

Chunk-level *speech emotion recognition* (SER) is a common modeling scheme to obtain better recognition performance than sentence-level formulations. A key open question is the role of lexical boundary information in the process of splitting a sentence into small chunks. Is there any benefit in providing precise lexical boundary information to segment the speech into chunks (e.g., word-level alignments)? This study analyzes the role of lexical boundary information by exploring alternative segmentation strategies for chunk-level SER. We compare six chunk-level segmentation strategies that either consider word-level alignments or traditional time-based segmentation methods by varying the number of chunks and the duration of the chunks. We conduct extensive experiments to evaluate these chunk-level segmentation approaches using multiples corpora, and multiple acoustic feature sets. The results show a minor contribution of the word-level timing boundaries, where centering the chunks around words does not lead to significant performance gains. Instead, the critical factor to effectively segment a sentence into data chunks is to define the number of chunks according to the number of spoken words in the sentence.

Index Terms— lexical information, speech emotion recognition, chunk-level modeling, data segmentation approach

1. INTRODUCTION

Speech emotion recognition (SER) system aims to automatically identify the emotional states of the input voice signals, which has important applications for advanced *human-computer interaction* (HCI). In general, SER recognizes either discrete emotional categories (e.g., happiness and sadness), or continuous emotional attributes (e.g., arousal, dominance, and valence) from acoustic features, forming a sequence-to-one recognition task [1, 2]. One of the core modeling problems in SER is how to extract reliable sentence-level representations from frame-level acoustic features of sentences with varied lengths (i.e., temporal modeling). Traditional methods rely on statistical descriptors over frame-level acoustic features (e.g., the mean of the fundamental frequencies), resulting in a high-dimensional vector to represent the audio signal [3]. Other approaches utilize deep learning models such as *convolutional and recurrent neural networks* (CRNN) to capture the temporal dynamics within the signal for better and robust sentence-level representation [4, 5]. However, the use of frame-level features results in long input sequences, which increases the complexity of the model [6], or leads to severe gradient vanishing issues [7]. To solve these problems, chunk-level learning schemes are widely used in SER [8–11], where the speaking turns are split into small chunks that are individually processed. The chunk-level representations are later aggregated to obtain a sentence-level representation.

Conventionally, chunk-level SER splits each sentence into smaller data chunks with a predefined segmentation approach to train the emotional classifier. The duration of the chunks varies, but even one second is able to carry enough information for determining emotions [12]. It is still an open question to find the optimal choice for segmenting the sentence into chunks. In particular, it is not clear if segmenting the sentence into chunks according to the correct lexical boundary information (e.g., word boundaries) leads to a better modeling strategy. The existing evidence is not conclusive. Jeon *et al.* [13] found superior performances by using pure time-based segmentation, ignoring any lexical knowledge over an approach that relied on a segmentation into word-level units. The results in Schuller and Rigoll [14] showed the highest SER accuracy by using the fusion of diverse time-level segments (i.e., different lengths of the data chunks) without relying on the actual timing provided by the lexical boundaries. Nonetheless, other studies also obtained improved recognition accuracy by introducing phonetic segmentation via pitch detection [15] or lexical content alignments [16, 17]. How critical is the role of the lexical boundary content in the chunk segmentation process? What are the key factors in the chunk segmentation that impact the model performance?

This study investigates the role of precise lexical boundaries in the segmentation of sentences into chunks. In particular, we study if centering the chunks around words leads to performance improvements. The analysis compares different segmentation approaches based on either word-level alignments or pure time-based methods. We consider two important variables in the chunk segmentation process: (a) the number of chunks in a sentence, C, and (b) the duration of the chunk, W. These parameters are either fixed or varied depending on the number and duration of the words in the sentence, creating six alternative chunk segmentation strategies. These six chunk segmentation methods have different levels of dependencies on the lexical boundary information. We evaluate the chunk-based segmentation strategies with the IEMOCAP [18] and MSP-Podcast [19] corpora. We implement our SER system with traditional *low-level descriptors* (LLDs) [20], and with the self-supervised speech representations Wav2Vec2 model [21]. We also compare these models with the temporal advanced deep chunk-level modeling framework proposed by Lin and Busso [9]. These experimental settings allow us to analyze trends across approaches and conditions, obtaining reliable insights to understand the role of lexical boundary information in the segmentation of chunks. In summary, the main findings in the study are:

- The use of the lexical boundary information for chunk-based segmentation has a minor role in creating an effective chunk-based segmentation strategy.
- The key benefit provided by the lexical information in the segmentation process is the number of words, which we used to deter-

This study was funded by the NSF under grant CNS-2016719.

mine the number of chunks used to segment the sentence.

2. BACKGROUND

2.1. Chunk-Level SER in Deep Learning Framework

Studies have used chunk-based segmentation for SER tasks [8–11, 22, 23]. This section describes some of these methods which have relied on deep learning formulations. The majority of the studies using chunk-based segmentation for SER have used pure time-based segmentation criteria [8, 10, 22, 23]. More specifically, the studies have defined a predefined fixed chunk window size with a shifting step (e.g., 1-sec chunks with 0.5 secs hop size) to split a sentence into multiple small data chunks. These data chunks are assigned the same sentence-level emotional label to train a chunk-level recognition model. The common approach to combine the results across chunks is to use a simple majority vote rule [10, 24]. Alternatively, the chunks can be combined by using a statistical (e.g., mean) pooling operation to create a sentence-level representation [8, 22]. However, these sentence-level aggregation approaches are not necessarily optimal given that a varied number of data chunks are produced depending on the sentence duration (i.e., a longer sentence is segmented into more chunks). Lin and Busso [9] proposed a dynamic chunk segmentation process to split a sentence with varied length into a fixed number of data chunks with fixed duration by dynamically adjusting the chunk step size according to the sentence duration.

2.2. Lexical Information for Data Segmentation

While lexical information has been commonly used to recognize emotions [25], our interest is to explore if lexical boundary information should be considered in the chunk-based segmentation process. One straightforward approach to incorporate lexical-based segmentations is to rely on a fundamental frequency detection approach to identify voiced regions in a sentence [26]. The detected pitch contour directly indicates the phonetic boundaries of voiced segments, which can be used as the speech unit to train an SER model [15]. Similarly, a syllable can be used as a chunk-based unit [27] by estimating the *vowel onset points* (VOPs) on spectrum and energy features. The speech segment between two successive VOPs can be considered as the syllable boundaries. Another conventional method is to perform forced alignment to generate word-level temporal boundaries based on the given transcription and trained acoustic model. The segmenting unit can also be extended to phrase-level boundaries (e.g., noun phrases) by detecting the syntax structure of the sentence [13]. In this study, we use word-level alignments as the cue to segment data chunks in the lexical-dependent approaches included in our analyses.

3. CHUNK-BASED SEGMENTATION

We explore alternative chunk-based segmentation strategies to study the role of lexical boundary information in finding an optimal segmentation. These methods have different levels of dependencies on the word boundaries, and the number of words in the speaking turn.

We define C as the number of chunks and W as the size of the chunk. These two parameters are important for splitting a sentence of varied duration into chunks. We evaluate six alternative segmentation strategies to investigate the role of lexical boundary information in the chunk-based segmentation process. The core concept is to determine how many and where to put the data chunks in a sentence. Some methods incorporate additional word-level alignment information to determine the placement of the chunks. Other methods rely exclusively on time-based strategies that place the chunks regardless of the lexical boundary information. The segmentation can have fixed or varied values for C and W. *FixedW* (FW) uses a predefined fixed chunk size W during the segmentation process. In contrast, *VariedW* (VW) uses a varied length for W, which depends on the target word durations. Similarly, *FixedC* (FC) assigns a fixed number of data chunks C to split a sentence. In contrast, *VariedC* (VC) assigns a varied number of data chunks C based on how many words are spoken in the sentence. These combinations result in the following different segmentation approaches, which are visualized in Figure 1.

• *Time-based FixedW-FixedC (tFW-FC)*: This strategy splits the sentences using FW and FC, regardless of the duration of the sentence. This approach corresponds to the dynamic chunk segmentation strategy presented by Lin and Busso [9, 11]. It is a time-based segmentation approach, which aims to split the sentence into a fixed number of chunks C with a fixed duration W. Equation 1 shows the key formula to achieve this goal, where it dynamically adjusts the step size Δc_i between the data chunks according to the duration of the ith sentence T_i. In this approach, the overlap between chunks decreases for longer sentences. This data segmentation approach does not rely on any lexical information (Fig. 1(a)).

$$\Delta c_i = \frac{T_i - W}{C - 1} \tag{1}$$

• *Lexical FixedW-FixedC (FW-FC)*: This approach splits the sentence using FW and FC with the word boundaries from the forced alignment results. This approach is achieved by centering the data chunks obtained in tFW-FC (i.e., the red arrows in Fig. 1(a)) around the center of their nearest word. For instance in Figure 1(b), the first data chunk from Figure 1(a) (see the first red arrow) is closer to the word "swear" than "I." Therefore, we place the first chunk centered around the word "swear" (Fig. 1(b)).

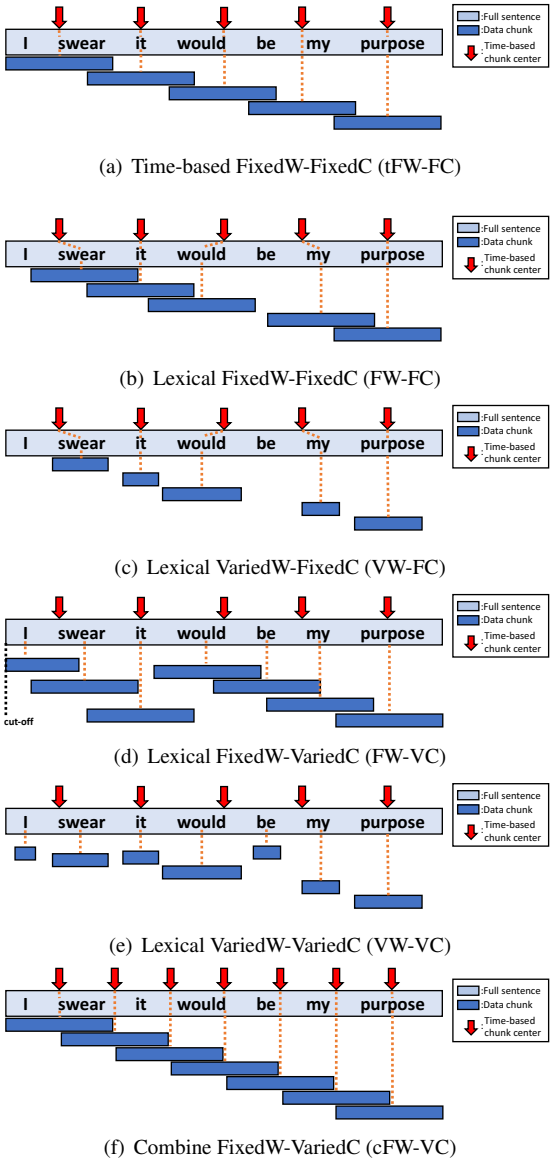
• *Lexical VariedW-FixedC (VW-FC)*: This segmentation strategy is similar to the FW-FC approach, but it varies the chunk window size to strictly match the exact word spoken duration (VW) rather than a predefined fixed window length (Fig. 1(c)).

• *Lexical FixedW-VariedC (FW-VC)*: This segmentation strategy places a data chunk centered at every uttered word in the sentence. The duration of the chunk is fixed with a predefined window length (FW). The number of data chunks depends on the number of spoken words (Fig. 1(d)).

• *Lexical VariedW-VariedC (VW-VC)*: This strategy corresponds to the most intuitive way to apply word-level chunk segmentation based on the alignment information. We add one chunk per word in the sentence, with the duration of the chunks matching the duration of the corresponding words (Fig. 1(e)).

• *Combine FixedW-VariedC (cFW-VC)*: This strategy is similar to the FW-VC approach. The key difference is that the data chunks are equally distributed instead of being placed around the words' centers. The number of chunks is equal to the number of words. Therefore, this approach uses a time-based segmentation, where the number of words is the only cue used by the strategy (Fig. 1(f)).

Once the data is segmented, we process each chunk with *long short-term memory* (LSTM) layers, creating a chunk-based representation. We aggregate the representations across chunks using the *multi-head self-attention* (MH). The combination of LSTM for chunk encoder and MH for temporal aggregation is one of the best combinations reported in Lin and Busso [9]. For approaches using FC, the number of chunks is fixed so it is easy to implement the MH approach as an end-to-end framework. For approaches using VC, the number of chunks depends on the number of words in the sen-



(a) Time-based FixedW-FixedC (tFW-FC)

(b) Lexical FixedW-FixedC (FW-FC)

(c) Lexical VariedW-FixedC (VW-FC)

(d) Lexical FixedW-VariedC (FW-VC)

(e) Lexical VariedW-VariedC (VW-VC)

(f) Combine FixedW-VariedC (cFW-VC)

Fig. 1. Visualization of different segmentation approaches explored in this study for one sentence. The red arrows show the center of the chunks if a time-based segmentation strategy would have been used.

tence creating a varied number of data chunks. Therefore, the entire model is implemented with a two-stage training strategy, where the LSTM encoder and MH attention model are separately trained. We apply zero-padding to match the same size across samples, using the maximum number of chunks in the sentences as the target size (i.e., the number of spoken words). Notice that the different training schemes do not impact the model architecture. It still has exactly the same model complexity for a fair comparison between segmentation approaches.

4. RESOURCES AND EXPERIMENTAL SETTINGS

4.1. Datasets and Acoustic Features

We perform the experimental analysis on two emotional databases. The first corpus is the MSP-Podcast database [19], which is the largest spontaneous speech emotion corpus. We use the release version 1.10 to conduct our experiments, which provides predefined train/development/test partitions with a total of 104,267 audio clips (∼166 hrs). These audios are processed from real-world audio conversations that are segmented into speaking turns with a duration ranging from 2.75 to 11 secs. The IEMOCAP database [18] is currently the most popular benchmark dataset in SER. The corpus consists of dyadic interactions collected from ten actors to elicit natural emotional expressions. The corpus has a total of 10,039 audio clips (∼12 hrs). We exclude files that are shorter than 1 sec and longer than 17 secs, retaining 98.7% of the data. Both datasets provide human transcriptions for each speaking turn. As a reference, the MSP-Podcast corpus has a higher average number of words per sentence (16.1) than the IEMOCAP corpus (11.5). For the recognition task, we train the SER model to predict the values for valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong). We formulate the SER task as a regression problem, since these emotion attributes are represented by continuous scores.

We implement our evaluation with traditional features and a self-supervised speech representation model. For the traditional acoustic features, we use the OpenSmile [28] toolkit to extract frame-level LLDs. We use the configuration for the *computational paralinguistics challenge* (ComParE) set presented at Interspeech 2013 [20]. This set consists of common acoustic features such as fundamental frequency and *Mel-frequency cepstral coefficients* (MFCCs). The total feature dimension is 130. For the self-supervised speech representation, we use the pretrained *wav2vec2-large-robust* model [21] from the Huggingface library [29]. This model is based on the transformer, and extracts frame-level representations. We use the last hidden state output of the model, which has a 1,024 dimension. Notice that we only treat the pretrained Wav2Vec2 model as a feature extractor, which is not finetuned or incorporated in the model architecture mentioned in Section 3.

4.2. Experimental Setups

The model is trained with either LLDs (130D) or Wav2Vec2 (1,024D) features. There is a small model architecture difference when we train with the Wav2Vec2 features. To reduce the model complexity, we add an additional fully connected layer to map the Wav2Vec2 feature representation into a 128D space. This vector corresponds to the input of the LSTM block. The implementation of the LSTM and MH models follows the architecture described by Lin and Busso [9]. We use two LSTM layers implemented with 128 nodes to process the chunk-level representation. We implement the MH model with 1,024 hidden dimensions, using two heads to aggregate the sentence-level representation. We train this model with batches of size 128, using the Adam optimizer (lr=0.0001). The loss function is based on the *concordance correlation coefficient* (CCC) since we are predicting emotional attributes. We use an early stopping criterion to check the performance of the development set. We also use dropout with a rate set to $p = 0.5$ for the LSTM and $p = 0.1$ for the MH model. The models are coded in PyTorch.

We utilize the *Montreal Forced Aligner* (MFA) [30] to obtain our word-level alignment results from the transcriptions. For methods implemented with FW, we set the chunk duration to $W = 1$ sec. Notice that over 98.5% of the words for both corpora have durations under 1 sec. Therefore, 1-sec chunks are long enough to cover most words. For the FC methods, we set the number of chunks according to the maximum sentence duration in the corpora, using $C = 11$ for the MSP-Podcast corpus and $C = 17$ for the IEMOCAP corpus.

We evaluate the SER models using CCC as the performance metric. All the CCC values are reported by considering multiple running trials with random network initializations. For the MSP-

Table 1. Summary of CCC performance for different segmentation approaches on the MSP-Podcast and IEMOCAP corpora, using LLDs and Wav2Vec2 features. The best result per column is highlighted in bold. Results tagged with $*$ show statistically significant better performance over other approaches without a marker. Results tagged with † indicate that the results are statistically significantly better than all other approaches.

	MSP-Podcast v1.10						
Method	*CovrR/OvrR*	*LLDs* (CCC)			*Wav2Vec2* (CCC)		
	[%]	Aro.	Val.	Dom.	Aro.	Val.	Dom.
tFW-FC	99 / 50	0.528	**0.216***	0.430	0.604	**0.352**	0.478
FW-FC	90 / 56	0.529	0.191	0.423	0.598	0.344	0.475
VW-FC	57 / 35	0.534	0.170	0.427	0.595	0.352	0.460
FW-VC	94 / 68	0.544*	0.141	0.455*	**0.620***	0.349	0.497*
VW-VC	82 / 0	0.546*	0.118	0.459*	0.613*	0.336	0.492*
cFW-VC	99 / 65	**0.562**†	0.207*	**0.468**†	0.616*	0.343	**0.499***
	IEMOCAP						
Method	*CovrR/OvrR*	*LLDs* (CCC)			*Wav2Vec2* (CCC)		
	[%]	Aro.	Val.	Dom.	Aro.	Val.	Dom.
tFW-FC	99 / 79	0.614	0.353	0.406	0.709*	0.554	0.531
FW-FC	82 / 83	0.593	0.257	0.411	0.700	0.537	0.539
VW-FC	47 / 71	0.595	0.279	0.409	0.688	0.532	0.526
FW-VC	81 / 67	0.633*	0.378	0.451*	0.713*	0.582*	0.538
VW-VC	61 / 0	0.626*	0.395*	0.433	**0.719***	0.577*	**0.558***
cFW-VC	99 / 60	**0.636***	**0.404**†	**0.463**†	0.718*	**0.584***	0.549*

Podcast, we run three model trials. We randomly split the original test set into ten subsets. This approach results in $3 \times 10 = 30$ evaluation data points to conduct statistical analysis between methods. Since the IEMOCAP corpus does not have predefined partitions, we implement a *leave-one-session-out cross-validation* (LOSO CV) strategy to perform speaker-independent evaluations. There are five dyadic sessions in the IEMOCAP corpus. We use three sessions for the train set, one session for the development set, and one session for the test set. The test set is randomly split into six subsets, producing $5 \times 6 = 30$ evaluation data points. We use a two-tailed t-test for the statistical analysis, defining significance with p-value ≤ 0.05.

5. RESULTS AND ANALYSES

Table 1 reports the results for our evaluation, including the CCC values for arousal, dominance, and valence obtained with the six-chunk segmentation strategies. The table also provides the average *coverage ratio* (CovrR) of the sentences obtained with the chunks, and the chunk *overlap ratio* (OvrR) for all the six segmentation methods. These metrics are used to quantify the difference between the segmentation approaches. The CovrR metric calculates the ratio (in percentage) of the actual feature frames that are covered within the data chunks of the full sentence. For example, the FW-FC strategy using the MSP-Podcast corpus has a CovrR equal to 90%, which indicates that 90% of the feature frames of the sentences are included in the chunks. The OvrR metric computes the percentage of overlapped frames between chunks. For instance, the VW-VC strategy has 0% overlaps between data chunks for both corpora, since it strictly follows the word boundaries obtained from the forced alignment results, without any overlap. Furthermore, the lower CovrR value in the IEMOCAP corpus (61%) indicates a more sparse distribution of the spoken content than in the MSP-Podcast corpus (82%).

Table 1 reveals some major points to understand the role of lexical boundary information. First, the VW-FC approach usually leads to lower performance compared to other segmentation strategies. One reason that explains this result is the low CovrR value for this method, indicating a lack of sentence coverage with the chunks. It is important to have enough coverage of the sentence with the chunks when considering the segmentation approach.

Second, we consistently find significantly better predictions when using the VC segmentation scheme (i.e., FW-VC, VW-VC and cFW-VC). Knowing the exact number of data chunks (i.e., the number of spoken words) to split a sentence brings benefits for chunk-level SER modeling. However, the knowledge of the word boundaries used to set the window size W does not lead to significant improvements. This result can be observed by directly comparing the FW-VC and VW-VC methods, or the FW-FC and VW-FC methods. In both comparisons, splitting the sentence into chunks using the strict word duration does not provide significant advantages over using a predefined fixed window size. For methods using the FW strategy, we generally observe high values for CovrR and OvrR. Therefore, using a window size of 1 sec for the chunks may be sufficient not only to cover the entire words, but also to produce high overlap between data chunks to better preserve the temporal continuity in the sentence.

Third, in general, we obtain the best performance with the cFW-VC segmentation approach, which only requires the number of words in the sentence. This method combines a time-based segmentation with partial lexical boundary information. The comparison between the FW-VC and cFW-VC approaches indicates that the placement of the chunks does not have to be aligned with the words. The cFW-VC strategy obtains better results without requiring any prior knowledge about the word locations. From a practical perspective, this is good news, since a simple time-based chunk segmentation method is enough to achieve the best performance, as long as we have an equivalent number of chunks as the number of words. In summary, the key contributing factor from the lexical information is to know how many data chunks the sentence needs to be split. In contrast, having a precise word alignment for segmenting the sentence into chunks plays a minor role in achieving the best SER performance. As an aside, Wav2Vec2 features consistently outperforms traditional acoustic LLDs, suggesting that self-supervised speech representation models are attractive options for SER tasks.

6. CONCLUSIONS

This study investigated different chunk-level SER segmentation approaches leveraging word-level alignments to explore the role of lexical boundary information in the chunk segmentation process. Our experimental results demonstrated that setting the number of chunks to match the number of spoken words in the sentence (VC scheme) is consistently better than having a fixed number of chunks per sentence (FC scheme). The key cue from lexical boundary information that benefits the most the chunk-based segmentation process was knowing how many spoken words are included in the sentence, which can be used to determine the number of chunks. In contrast, centering the chunks around the words did not lead to significant improvements, showing a minor role in the performance. We observe similar results by placing the chunks based on time-based strategies that do not rely on lexical boundary information.

Our analysis has important implications for SER tasks. It is enough to segment the sentence into chunks by setting the number of chunks to be similar to the number of words in the sentence. This strategy simplifies SER formulations. While these results are true for SER, we are interested in exploring if this result also holds in multimodal emotion recognition. We hypothesize the benefits of having chunk-based segmentations using lexical boundary information by improving the alignment between lexical and acoustic modalities.

7. REFERENCES

[1] S. Ntalampiras, "Toward language-agnostic speech emotion recognition," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 7–13, January 2020.

[2] S. Ntalampiras, "Speech emotion recognition via learning analogies," *Pattern Recognition Letters*, vol. 144, pp. 21–26, April 2021.

[3] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088, IEEE.

[4] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[5] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, December 2014.

[6] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *International Conference on Learning Representations (ICLR 2020*, Addis Ababa Ethiopia, April-May 2020, pp. 1–12.

[7] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[8] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Singapore, September 2014, pp. 223–227.

[9] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. Early Access, 2022.

[10] S. Sahoo, P. Kumar, B. Raman, and P. Pratim Roy, "A segment level approach to speech emotion recognition using transfer learning," in *Asian Conference on Pattern Recognition (ACPR 2019)*, S. Palaiahnakote, G. Sanniti di Baja, L. Wang, and W. Yan, Eds., vol. 12047 of *Lecture Notes in Computer Science*, pp. 435–448. Springer Berlin Heidelberg, Auckland, New Zealand, November 2019.

[11] W.-C. Lin and C. Busso, "An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 2322–2326.

[12] B. Schuller and L. Devillers, "Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 801–804.

[13] J.H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4940–4943.

[14] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *International Conference on Spoken Language (ICSLP 2006)*, Pittsburgh, PA, USA, September 2006, pp. 1818–1821.

[15] M. T. Shami and M. S. Kamel, "Segment-based approach to the recognition of emotions in speech," in *IEEE International Conference on Multimedia and Expo (ICME 2005)*, Amsterdam, Netherlands, July 2005, pp. 1–4.

[16] J.P. Arias, C. Busso, and N.B. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 278–294, January 2014.

[17] B. Vlasenko, B. Schuller, K. T. Mengistu, G. Rigoll, and A. Wendemuth, "Balancing spoken content adaptation and unit length in the recognition of emotion and interest," in *Interspeech 2008*, Brisbane, Australia, September 2008, pp. 805–808.

[18] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[19] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, Virtual, December 2020, vol. 33, pp. 12449–12460.

[22] L. Tarantino, P.N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2578–2582.

[23] E. Tzinis and A. Potamianos, "Segment-based speech emotion recognition using recurrent neural networks," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 190–195.

[24] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613–625, July-August 2010.

[25] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 604–624, February 2021.

[26] C. Busso, S. Lee, and S.S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[27] S. G. Koolagudi, N. Kumar, and K. S. Rao, "Speech emotion recognition using segmental level prosodic analysis," in *International Conference on Devices and Communications (ICDeCom 2011)*, Mesra, India, February 2011, pp. 1–5.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[29] T. Wolf et al., "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[30] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable text-speech alignment using Kaldi," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 498–502.