ConflictNET: End-to-End Learning for Speech-based Conflict Intensity Estimation

Vandana Rajan, Alessio Brutti, and Andrea Cavallaro

***Abstract*—Computational paralinguistics aims to infer human emotions, personality traits and behavioural patterns from speech signals. In particular, verbal conflict is an important example of human-interaction behaviour, whose detection would enable monitoring and feedback in a variety of applications. The majority of methods for detection and intensity estimation of verbal conflict apply off-the-shelf classifiers/regressors to generic handcrafted acoustic features. Generating conflict-specific features requires refinement steps and the availability of metadata, such as the number of speakers and their speech overlap duration. Moreover, most techniques treat feature extraction and regression as independent modules, which require separate training and parameter tuning. To address these limitations, we propose the first end-to-end convolutional-recurrent neural network architecture that learns conflict-specific features directly from raw speech waveforms, without using explicit domain knowledge or metadata. Additionally, to selectively focus the model on portions of speech containing verbal conflict instances, we include a global attention interface that learns the alignment between layers of the recurrent network. Experimental results on the SSPNet Conflict Corpus show that our end-to-end architecture achieves state-of-the-art performance in terms of Pearson Correlation Coefficient.**

***Index Terms*—Computational Paralinguistics, Conflict Intensity Estimation, Convolutional-Recurrent Network.**

I. Introduction

THE recognition of emotions from speech [1], the classification of infant vocalization [19], the detection of depression and other health conditions [6][27][28], as well as the estimation of conflict intensity [18] are important computational paralinguistics problems. In particular, the automatic estimation of conflict from speech signals has several important applications, such as monitoring conflicts during meetings and in call centers to help employees handle difficult interactions and thereby reduce stress and anxiety.

Conflict is an interaction process between parties who pursue incompatible goals [26]: each party perceives that their interests are being opposed or negatively affected by another party [14]. While goals and interests are not directly observable, they influence human behaviour through gestures, facial expressions and speech [5].

Verbal conflict analysis can be formulated as detection or estimation problem. *Conflict detection* aims to identify if a given temporal interval of speech contains verbal conflict [3][10][15]. *Conflict intensity estimation* is a regression task that aims to determine a continuous level of conflict intensity [9][12], which is more informative than the binary class label generated by conflict detection methods [9]. Traditional conflict detection and conflict intensity estimation methods use state-of-the-art classifiers/regressors on generic hand-crafted acoustic features, which require further manual refinement and time-consuming feature pruning [8][9][15][17]. Task-specific hypotheses and metadata, like the number of speakers and the ratio of their speech overlaps, may also be needed to extract conflict-specific features from standard acoustic features [2][3][10]. Another drawback of these methods is the need for separate training and parameter tuning of the feature extractor and the classifier/regressor. An alternative approach is end-to-end learning, which trains models directly from raw input data: since the parameters are trained jointly, the end-to-end model learns task-specific features from the input, without requiring any guidance other than the objective function and the training dataset.

In this work, we propose *ConflictNET*[1], an end-to-end Convolutional-Recurrent-Neural-Network (CRNN) architecture, that learns to estimate conflict intensity directly from raw speech. Regression with an end-to-end architecture eliminates the need for metadata or task-specific knowledge. Feature extraction and regression are combined in a single deep neural network that can be trained in an end-to-end fashion. We use a CRNN architecture similar to those in [21], [22], [25] and [29], where successive convolutional layers learn features at different levels and a Long-Short-Term Memory (LSTM) network learns the temporal relationships between features. Additionally, to enable the network to focus on temporal intervals that are more relevant for conflict intensity estimation, we introduce a global attention mechanism between LSTM layers by using weighted combinations of hidden states from several time-steps. Finally, we add a temporal average pooling layer to reduce the number of input time-steps to the LSTM layers and show that it improves the performance of our model.

To the best of our knowledge, there is no prior work on end-to-end deep learning for verbal conflict intensity estimation from raw speech signals.

II. Related Work

In this section, we discuss end-to-end learning using Deep-Neural-Networks (DNNs) applied to related computational paralinguistics tasks. We also discuss speech-based conflict detection and conflict intensity estimation methods. The key methods are summarised in Table I.

CRNN-based end-to-end learning has been recently applied to emotion recognition [21][22], where the input is raw speech

V. Rajan and A. Cavallaro are with the Centre for Intelligent Sensing, Queen Mary University of London, UK (e-mail: v.rajan@qmul.ac.uk, a.cavallaro@qmul.ac.uk). A. Brutti is with the Fondazione Bruno Kessler, Italy (e-mail: brutti@fbk.eu).

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

and the output is a pair of continuous values indicating the level of emotional valence and arousal. Similar CRNN architectures have also been used for emotion classification [29] and infant vocalization classification [25]. These end-to-end learning methods use single stream networks consisting of multiple convolution-maxpooling layers followed by various choices of recurrent layers like LSTM [22][29], Bidirectional LSTM (BLSTM) [21] or Gated Recurrent Units (GRUs) [25]. A Convolutional Neural Network (CNN) based end-to-end architecture for customer satisfaction prediction from contact center phone calls uses conflict detection as an auxiliary task to initialize the network weights [20].

Verbal conflict detection and intensity estimation were popularized by the *conflict sub-challenge* of the INTERSPEECH 2013 Computational Paralinguistics Challenge [18], whose baseline relied on 6,373 acoustic features extracted using OpenSMILE [7]. Most of the conflict detection and conflict intensity estimation methods either identify a subset of these features that are relevant for this task [8][12][15][17] or, to generate conflict-specific features, use metadata like the number of simultaneous speakers, interruptions and turn-taking characteristics [2][3][9][10].

The relevance of features can be determined by repeated classification using random feature subset selection [17], canonical correlation analysis based discriminative projection [15], greedy forward-backward feature selection [8] or ensemble Nyström method on manually partitioned feature subsets [12]. A major drawback of these methods is that they require checking all possible feature subsets to reduce feature redundancy and identify conflict-specific features. For example, [17] performs 300,000 iterations to identify 349 conflict specific features out of the 6,373 baseline features.

A Support Vector Machine (SVM) classifier can be used for conflict detection using predicted speech overlap ratio [10] or speech overlap based features [3]. Speech overlap predictions generated by a BLSTM can also be used for conflict detection using a DNN classifier [2]. Utterance-level features, obtained by combining frame-level DNN predicted speech overlap posteriors along with a subset of the baseline features, can be used for conflict intensity estimation using Support Vector Regressors (SVR) [9]. These methods require the availability of metadata, like the number of speakers and speech overlap duration.

The number of papers on speech-based conflict intensity estimation is scarce since [9]. Recently, a multi-modal conflict estimation method used a concatenation of audio and visual features as input to an LSTM-based encoder-decoder architecture with attention. This method focuses on visual features (facial gestures) and uses 65 audio Low-Level Descriptors (LLD) features, sampled at 25 Hz [23]. While hand-crafted features may facilitate interpretation of specific characteristics of the speech signal that are used as predictors for the task at hand, we aim to explore if an end-to-end learning framework can be used for a complex paralinguistic task such as verbal conflict intensity estimation by automatically learning relevant acoustic features for this task.

TABLE I
Summary of features, refinement methods and classifiers/regressors. KEY - IS13: Interspeech 2013 Conflict sub-challenge baseline features; IS10: Interspeech 2010 paralinguistics challenge baseline features; rep. class.: repeated classification; conv.: conversational features; pros.: prosodic features; over.: overlap features; Class/Reg: Classifier/Regressor; KNN: K Nearest Neighbour; SVM: Support Vector Machine; SVR: Support Vector Regressor; LSTM: Long Short Term Memory; BLSTM: Bi-directional LSTM; SPLSR: Sparse Partial Least Squares Regression; FPF: Facial Point Features; LLD: Low Level Descriptors; CRNN: Convolutional Recurrent Neural Network

Ref.	Input	Feature Refinement Method	Class/Reg
[17]	IS13	relevance adjustment by rep. class.	KNN
[15]	IS13	canonical correlation analysis	SVM
[8]	IS13	forward-backward pass	SVR
[12]	IS13	manual feature partitioning + ensemble Nyström	ensemble SPLSR
[10]	IS13	speech overlap ratio using SVR	SVM
[2]	conv. & pros.	speech overlap ratio using BLSTM	DNN
[9]	IS13 & over.	forward-backward pass	SVR
[3]	IS10 & IS13	overlap detection using SVR + backward selection	SVM
[23]	FPF & LLD	LSTM based encoder-decoder network	
[20]	raw speech	End-to-End Convolutional Neural Network	
Ours	raw speech	End-to-End CRNN with attention	

III. ConflictNET: End-to-End Model Design

In this section, we present *ConflictNET*, an end-to-end model that, given raw speech waveforms, predicts a continuous value representing the level of conflict. Our model combines feature extraction and regression in a unified framework, which contains six types of layers (convolutional, max-pooling, average pooling, LSTM, attention and fully connected layers) arranged in a single stream (see Figure 1).

Features from the speech signal are extracted by 1D *convolutional* layers with learnable filters. We have 3 1D strided convolutional layers, with 64, 128 and 256 filters respectively. Each convolutional layer uses ReLu activation. 1D filters of successive convolutional layers, each with stride 1, are of size 6, 4 and 4 respectively. A progressive increase in the number of filters as well as decrease in filter size after the first convolutional layer is due to the fact that, with increased depth, the network learns more detailed features.

Changes in the parameters of network layers during training modify the distribution of the input to their subsequent layers, a phenomenon known as internal covariate shift [13]. To reduce the effect of this phenomenon and thereby accelerate the training, we perform *batch normalization* after each convolutional layer. Successive *max-pooling* layers downsample the convolution outputs and reduce the number of network parameters. The pooling size is determined by considering the rate of overlap, R, between convolution filter size, F, and pooling size, P [22]:

$$R = \frac{F-1}{F+P-1}. \quad (1)$$

We keep $R < 0.4$ and use a stride size equal to the pool size in all the max pooling layers.

A given input speech signal can contain multiple instances of verbal conflict spread across time. A common choice to

TABLE II
TRAIN-VAL-TEST SPLIT [18] FOR THE SSPNET CONFLICT CORPUS

	Train	Val	Test	Total
Low (conflict$<$0)	471	127	226	824
High (conflict$\geq$0)	322	113	171	606
Total	793	240	397	1430

model such temporal sequential data is to use a Recurrent Neural Network (RNN). However, as the length of the input signal increases over time, it becomes harder to train a vanilla RNN due to the vanishing gradient problem, which can be attenuated using an LSTM [11]. Thus, we use two tanh-activated LSTM layers, with 128 and 64 units respectively, to capture the inter-dependencies between features across time.

Although, theoretically, there is no limitation on the number of time-steps an LSTM can process, our experiments showed that restricting the number of time steps to fewer than 250 improves performance. Thus, we use a *temporal average pooling* layer of pool size 4 to reduce the number of input time-steps to the first LSTM layer.

Intuitively, not all portions of an input speech signal will contribute equally towards the conflict intensity estimate of the entire signal. Thus, to enable the network to focus on portions of the signal that are more relevant for conflict intensity estimation, we add an *attention mechanism* between the LSTM layers. The LSTM layer with 128 units provides a sequence output rather than a single value to the attention layer, which assigns different weights to hidden states across different time-steps. We use a global additive self-attention mechanism [30], which considers the whole context to calculate relevance:

$$\begin{cases} g(t,t^{'}) = \tanh(W_g h_t + W_{g'} h_{t'} + b_g), \\ e(t,t^{'}) = \tanh(W_a g(t,t^{'}) + b_a), \\ a(t) = \text{softmax}(e(t)), \\ l_t = \sum_{t'} a(t,t^{'}) h^{'}_t, \end{cases} \tag{2}$$

where W_g and $W_{g'}$ are weight matrices corresponding to hidden states h_t and $h_{t'}$ respectively; W_a is the weight matrix corresponding to their non-linear combination; b_g and b_a are the bias vectors; $a(t,t^{'})$ captures the similarity between h_t and $h_{t'}$; l_t represents the attention focused hidden state representation, which is then given as input to the second LSTM at time-step t.

The conflict intensity value is predicted by a *fully connected layer* with a linearly activated single output neuron, which is connected to the final time-step of the last LSTM layer.

Finally, we design our loss function, L, to maximise the Pearson Correlation Coefficient (PCC):

$$L = 1 - PCC = 1 - \frac{1}{N\sigma\hat{\sigma}} \sum_{i=1}^{N} (y_i - \mu)(\hat{y}_i - \hat{\mu}), \tag{3}$$

where N is the number of labels; y_i and $\hat{y}_i$ are true and predicted labels, respectively; and (μ,σ) and ($\hat{\mu}$,$\hat{\sigma}$) are their corresponding mean and standard deviation.

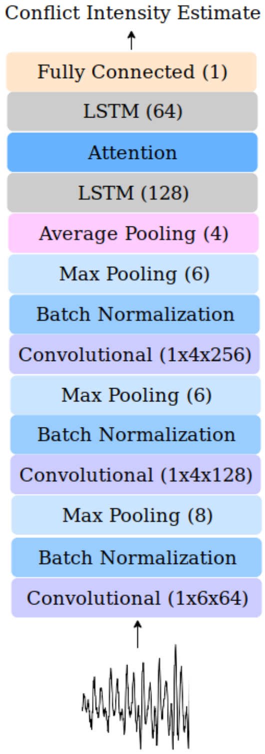

Fig. 1. The proposed *ConflictNET* architecture for conflict intensity estimation.

IV. VALIDATION

In this section, we compare the performance of *ConflictNET* with other conflict intensity estimation and classification methods. We also present an ablation study that quantifies the contributions of different parts of our network, starting from a baseline convolutional-recurrent model composed of three sets of convolution, max pooling and batch normalization layers as well as two LSTM layers and a fully connected layer. We refer to this model as *ParaNET*.

A. Dataset

We adopt the SSPNet Conflict Corpus [16], an audio-visual corpus that consists of 1,430 clips, each of duration 30 seconds, totalling ~12 hours of recordings extracted from a collection of political debates in French. Each clip is rated by 10 different non-French speaking assessors and the conflict intensity value assigned to each clip is the average of individual scores [24]. These values are in the range [-10,10], from no conflict (-10) to high level of conflict (+10), thus making the dataset suitable for regression tasks. Audio signals are sampled at 48KHz, resulting in 1,440,000 samples per clip. This dataset was adopted in the *conflict sub-challenge* of the INTERSPEECH 2013 Computational Paralinguistics Challenge [18] using only the audio signal. In our experiments, we follow the same training-validation-testing data split as defined in the challenge (see Table II). Note that the challenge considered a binary classification task, obtained by classifying the conflict level into high ($\geq$ 0) or low ($<$ 0). Also, we convert the target labels from the range [-10,10] to [-1,1] for compatibility with the activations of the neural network.

Due to memory considerations, we downsample each clip to 8 KHz, thus reducing the number of samples per clip to 240,000. Because of the uneven distribution of verbal conflict instances, we process the input signal as a whole. To even the loudness over the entire input signal S, we perform root-mean-square normalization as follows:

$$s = \frac{S}{\sqrt{\frac{\sum_{i=1}^{M}|S_i|^2}{M}}}, \tag{4}$$

where S_i is the i^{th} sample, *M* is the total number of samples of the input signal and s is the normalized signal.

B. Evaluation measures

We consider three evaluation measures: Pearson Correlation Coefficient (PCC) for regression, Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) for classification. Note that we obtain the classification outputs after binarizing the predicted continuous labels into high and low conflict levels. Also, we map the predicted output values to the same range as the training labels before calculating UAR and WAR, which helps to improve these evaluation measures without changes in the PCC value. The results we report are average values obtained after training and testing the model for 10 times.

C. Training

The model was developed, trained and tested using Keras with Tensorflow backend [4]. The model was trained using the training set and the validation set was used to identify the epoch for early stopping and model saving callbacks. We used the Adam optimizer with a learning rate of 0.01 and decay of 0.6 for training the network with mini-batches of size 32. The model was selected based on the highest PCC value on the validation set.

D. Ablation Study and Results

The results obtained by adding subsequent layers to *ParaNET* are summarized in the bottom of Table III. The performance of our baseline model *ParaNET* is much better than the expected measure by chance values (PCC = -0.008 ± 0.023, UAR = 50%) given in [18]. An average pooling operation at the input of the first LSTM layer improves the performance on all the 3 evaluation measures, which can be attributed to the better performance of the LSTM obtained by reducing the number of input time-steps. An attention layer added to *ParaNET* improves its performance by a noticeable margin of 0.162, 9.8% and 8.1% in PCC, UAR and WAR, respectively. This supports our intuition that weighted combinations of hidden states across multiple time-steps can result in performance improvement of the LSTM layers. Further, adding both average pooling and attention layers to *ParaNET* improves the PCC value to 0.853 ± 0.003. We also experimented by using a Global Average Pooling layer that took an average of the entire output sequence of the second LSTM layer before feeding it to the fully connected layer. However, adding this layer resulted in a slight decrease of 0.002 in PCC and a slight improvement of 0.2% and 0.5% in UAR and WAR values, respectively. It is worthwhile to note that the standard deviation of UAR and WAR values (0.43% and 0.51%, respectively) are higher than that of PCC. This is not surprising since we optimized our network in terms of PCC alone.

TABLE III
PERFORMANCE COMPARISONS ON THE SSPNET CONFLICT CORPUS TEST SET. NOTE THAT THE RANGE OF PCC IS [-1,1], AND THAT OF UAR AND WAR ARE IN PERCENTAGE. KEY - '*' RESULTS REPORTED BY TRAINING ON BOTH TRAINING AND VALIDATION SETS; '-' VALUES NOT REPORTED; REF: REFERENCE; PCC: PEARSON CORRELATION COEFFICIENT; WAR: WEIGHTED AVERAGE RECALL; UAR: UNWEIGHTED AVERAGE RECALL; NN: NEURAL NETWORK; DNN: DEEP NEURAL NETWORK; AP: AVERAGE POOLING; ATTN: ATTENTION; GAP: GLOBAL AVERAGE POOLING

Ref	Method	PCC	UAR	WAR
[18]	INTERSPEECH'13 baseline	.826*	80.8*	-
[17]	Random subset feature selection	.826	81.6	82.1
[15]	Random discriminative projection	-	84.6*	-
[2]	Deep hierarchical neural networks	.838*	84.3*	-
[8]	Greedy forward-backward	.842*	85.6*	-
[12]	Ensemble Nyström method	.849*	-	-
[10]	Detection using speaker overlap	-	83.1	-
[3]	Speech interruption detection	-	85.3	-
[9]	DNN-based feature extraction	.856	84.7	-
[20]	End-to-End Convolutional NN	.779	79.8	-
	ParaNET	.675	72.4	75.3
	ParaNET + AP	.781	79.9	81.3
	ParaNET + Attn	.837	82.2	83.4
	ParaNET + AP + Attn + GAP	.850	84.5	84.8
	ConflictNET: *ParaNET* + AP + Attn	.853	84.3	84.3

The comparison[2] in Table III shows that the performance of *ParaNET*+AP is similar to that of the end-to-end solution in [20]. Our best performing model *ConflictNET* outperforms in terms of PCC all but one method ([9]). *ConflictNET* achieves almost the same performance as [9], a model with DNN based speech overlap feature set and feature pruning based conflict specific subset of standard acoustic features. This suggests that our end-to-end architecture has automatically learned task-specific information from the raw speech input.

V. CONCLUSION

We proposed a new convolutional-recurrent neural network architecture for end-to-end conflict intensity estimation from raw speech data, the first of its kind. We quantified the effectiveness of adding an attention mechanism and an average pooling layer to a baseline convolutional-recurrent architecture. Unlike previous works on this topic, our end-to-end model implicitly learns conflict-specific features directly from the input speech waveform. The performance of the proposed model is on par with the state-of-the-art method in terms of Pearson Correlation Coefficient on the SSPNet Conflict Corpus dataset. A future research objective is to understand the evolution of negative emotions, like anger and aggression, arising from verbal conflicts.

Acknowledgement. This research used QMUL's Apocrita HPC facility, supported by QMUL Research-IT.

[2] As the results of [23] on the SSPNet Conflict Corpus are not available, this method in not included in the comparison.

References

[1] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, 2015.

[2] Raymond Brueckner and Björn Schuller. Be at odds? Deep and hierarchical neural networks for classification and regression of conflict in speech. In *Conflict and Multimodal Communication*, pages 403–429. Springer, 2015.

[3] Marie-José Caraty and Claude Montacié. Detecting speech interruptions for automatic conflict detection. In *Conflict and Multimodal Communication*, pages 377–401. Springer, 2015.

[4] François Chollet et al. Keras. https://keras.io, 2015. (last accessed on 26 August 2019).

[5] Virginia W Cooper. Participant and observer attribution of affect in interpersonal conflict: an examination of noncontent verbal behavior. *Journal of Nonverbal Behavior*, 10(2):134–144, 1986.

[6] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71(C):10–49, 2015.

[7] Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 1459–1462, Firenze, Italy, October 2010.

[8] Gábor Gosztolya. Conflict intensity estimation from speech using greedy forward-backward feature selection. In *Proceedings of INTERSPEECH*, pages 1339–1343, Dresden, Germany, September 2015.

[9] Gábor Gosztolya and László Tóth. DNN-based feature extraction for conflict intensity estimation from speech. *IEEE Signal Processing Letters*, 24(12):1837–1841, 2017.

[10] Félix Grezes, Justin Richards, and Andrew Rosenberg. Let me finish: automatic conflict detection using speaker overlap. In *Proceedings of INTERSPEECH*, pages 200–204, Lyon, France, August 2013.

[11] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[12] Dong-Yan Huang, Haizhou Li, and Minghui Dong. Ensemble Nyström method for predicting conflict level from speech. In *Proceedings of Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–5, Siem Reap, Cambodia, December 2014.

[13] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of International Conference on Machine Learning*, pages 448–456, Lille, France, July 2015.

[14] Charles M Judd. Cognitive effects of attitude conflict resolution. *Journal of Conflict Resolution*, 22(3):483–498, 1978.

[15] Heysem Kaya, Tuğçe Özkaptan, Albert Ali Salah, and Fikret Gürgen. Random discriminative projection based feature selection with application to conflict recognition. *IEEE Signal Processing Letters*, 22(6):671–675, 2014.

[16] Samuel Kim, Fabio Valente, Maurizio Filippone, and Alessandro Vinciarelli. Predicting continuous conflict perception with Bayesian Gaussian processes. *IEEE Transactions on Affective Computing*, 5(2):187–200, 2014.

[17] Okko Räsänen and Jouni Pohjalainen. Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech. In *Proceedings of INTERSPEECH*, pages 210–214, Lyon, France, August 2013.

[18] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings of INTERSPEECH*, pages 148–152, Lyon, France, August 2013.

[19] Björn W Schuller, Stefan Steidl, Anton Batliner, Peter B Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian Pokorny, Eva-Maria Rathner, Katrin D Bartl-Pokorny, et al. The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In *Proceedings of INTERSPEECH*, Hyderabad, India, September 2018.

[20] Carlos Segura, Daniel Balcells, Martí Umbert, Javier Arias, and Jordi Luque. Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls. In *Proceedings of International Conference on Advances in Speech and Language Technologies for Iberian Languages*, pages 255–265, Lisbon, Portugal, November 2016.

[21] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, Shanghai, China, March 2016.

[22] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093, Calgary, Canada, April 2018.

[23] Ruben Vereecken, Stavros Petridis, Yiannis Panagakis, and Maja Pantic. Online attention for interpretable conflict estimation in political debates. In *Proceedings of 2018 IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 389–393, Xian, China, May 2018.

[24] Alessandro Vinciarelli, Samuel Kim, Fabio Valente, and Hugues Salamin. Collecting data for socially intelligent surveillance and monitoring approaches: The case of conflict in competitive conversations. In *Proceedings of 2012 International Symposium on Communications, Control and Signal Processing*, pages 1–4, Roma, Italy, May 2012.

[25] Johannes Wagner, Dominik Schiller, Andreas Seiderer, and Elisabeth André. Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant? In *Proceedings of INTERSPEECH*, pages 147–151, Hyderabad, India, September 2018.

[26] James A Wall Jr and Ronda Roberts Callister. Conflict and its management. *Journal of management*, 21(3):515–558, 1995.

[27] Jun Wang, Prasanna V Kothalkar, Beiming Cao, and Daragh Heitzman. Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples. In *Proceedings of INTERSPEECH*, pages 1195–1199, San Francisco, USA, September 2016.

[28] Jochen Weiner, Christian Herff, and Tanja Schultz. Speech-based detection of Alzheimer's disease in conversational German. In *Proceedings of INTERSPEECH*, pages 1938–1942, San Francisco, USA, September 2016.

[29] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[30] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. OpenTag: Open attribute value extraction from product profiles. In *Proceedings of the International Conference on Knowledge Discovery & Data Mining*, pages 1049–1058, London, UK, August 2018.