Interspeech 2018
2-6 September 2018, Hyderabad

Investigating Utterance Level Representations for Detecting Intent from Acoustics

SaiKrishna Rallabandi[1], Bhavya Karki[1], Carla Viegas[1,2], Eric Nyberg[1] and Alan W Black[1]

Language Technologies Institute,Carnegie Mellon University, PA, USA [1]
NOVA Laboratory for Computer Science and Informatics, FCT NOVA,
Campus Caparica, Almada, Portugal [2]

lti.cs.cmu.edu

Abstract

Recognizing paralinguistic cues from speech has applications in varied domains of speech processing. In this paper we present approaches to identify the expressed intent from acoustics in the context of INTERSPEECH 2018 ComParE challenge. We have made submissions in three sub-challenges: prediction of 1) self-assessed affect and 2) atypical affect 3) Crying Sub challenge. Since emotion and intent are perceived at suprasegmental levels, we explore a variety of utterance level embeddings. The work includes experiments with both automatically derived as well as knowledge-inspired features that capture spoken intent at various acoustic levels. Incorporation of utterance level embeddings at the text level using an off the shelf phone decoder has also been investigated. The experiments impose constraints and manipulate the training procedure using heuristics from the data distribution. We conclude by presenting the preliminary results on the development and blind test sets.

Index Terms: speech processing, convolutional neural networks, strength of excitation, classification, emotion

1. Introduction

Applications of Computational Paralinguistics have grown rapidly over the last decade and span both human-human as well as human-machine interactions. The ComPare Paralinguistics challenges have been playing a significant role in driving progress in the diverse use of paralinguistics. Besides the traditional task of affect recognition using suprasegmental nonverbal aspects of speech, novel tasks were introduced, such as, the detection of speaker traits, deception, conflict, eating and autism [1, 2, 3, 4]. These challenges have shown that paralinguistic information can be used not only to identify affect but also clues that are helpful to detect abnormalities indicating disorders. Paralinguistic information also has applications in other domains of speech processing such as dialog systems, speech synthesis, voice conversion, assistance systems, and eHealth systems.

In this paper, we present our approach to three of the INTERSPEECH 2018 ComPare sub-challenges [5]: prediction of 1) self-assessed affect, 2) atypical affect and 3) types of crying. The *Self-Assessed Affect **(S)** Sub-Challenge* and the *Atypical Affect **(A)** Sub-Challenge* aim to classify affect from speech. In **(S)** ground-truth labels are provided by the speaker itself. The prediction of affect from speech oriented by the own assessment, could be used as a support in eHealth systems for individuals with affective disorders, such that a therapist can monitor the emotional state of their clients. In **(A)** the goal is to determine the affect of mentally, neurologically, and/or physically disabled individuals. The challenge is that some disorders also affect way people express their emotions. However, having a system able to detect distress in workplaces of disabled individuals can be helpful to make supervisors aware to suggest breaks or divide tasks in smaller ones, improving the emotional state of workers and therefore their concentration. The *Crying **(C)** Sub-Challenge* focuses on using paralinguistic information to identify affect in vocalisations of infants. Experts in the field of early speech-language development labeled audio-video clips into three classes of vocalisations: neutral/positive , fussing, and crying.

Typical approaches for classification and prediction of paralinguistic features include extraction of low level descriptive features followed by a machine learning model. Examples of low level descriptors are Mel-Frequency Cepstral Coefficients (MFCCs), log Mel-scale filter banks energies (FBANK) and several suprasegmental acoustic features that can be extracted using the openSMILE tool [6]. These features act as general purpose feature set and are expected to achieve competitive results in a wide range of paralinguistic problems. However, derived neural representations using unsupervised learning have shown impressive results on many speech and image based tasks recently [7]. These features usually embed the task relevant information from the entire utterance in a compact form. Also end-to-end learning models have been employed in affect classification using Long Short-Term Memories (LSTMs) or Gated Recurrent Units (GRUs) [8, 5].

Motivated by this, we explore different utterance level representations and end-to-end approaches in the context of sub-challenges. Specifically, we investigate the significance of using both utterance level acoustic and derived linguistic features. We further employ data augmentation using utterance emphasis (see section 2.3.4) and random utterance segmentation (section 2.3.2), as a strategy to cope with class imbalance. For obtaining linguistic features we first obtain the text for each of the utterances using a pretrained English ASPire model. We then train a Recurrent Neural Network language model on the obtained text at the phone level and use the representation at the hidden state as the embedding of the utterance. Apart from this, we explore the applications of various Convolutional Neural Network models and chart their performance. It has to be noted that even though acoustic and phonetic embeddings use identical inputs, they differ in the higher level features learned internally. Therefore we believe that they complement each other producing a superior fusion result.

2. Framework

In this section, we present different features and classifiers used for all three sub-challenges. We used two different classification models: 1) Bidirectional LSTM using low-level features which

10.21437/Interspeech.2018-2149

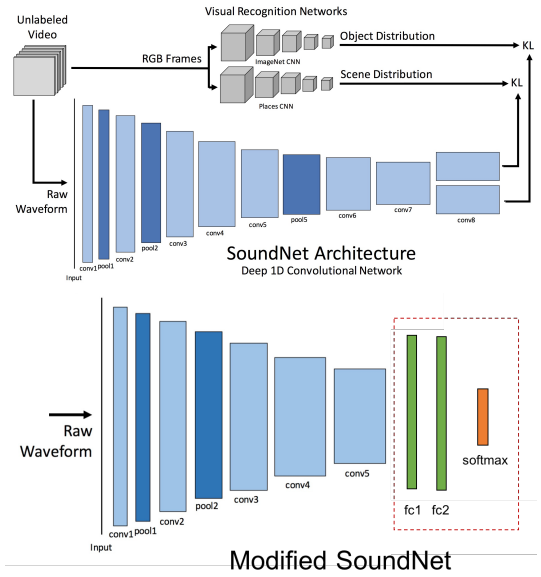

Figure 1: *Original SoundNet architecture [7] on top and modified SoundNet architecture at the bottom. The modified version uses 2 layers of 512 fully connected (fc) units and a softmax layer of 3 units.*

uses temporal information , and 2) Random Forest classifier or SVM Classifier using high-level features, which are utterance based, combined with utterance level embeddings.

2.1. Temporal classification

2.1.1. Low level features

For acoustic feature extraction we divided each utterance (length is 8 s) into 25 ms segments with a 10 ms frame shift. For each frame we extract 13 mel-frequency cepstral coefficients and their deltas and double-deltas obtaining a feature vector of 39 dimensions. We further extract the log pitch ($f0$) and strengths of excitation (5 dim) [9]. In addition, we also obtain 40 dimensional filter banks and 23 dimensional PLP based features. Filter banks have been obtained using the open source toolkit Kaldi [10] with 'dithering' enabled as it was shown to be robust in other experiments. We have also extracted several features using Opensmile toolkit [6] and performed singular value decomposition with the intention of obtaining an acoustic representation. This procedure also results in a dense low dimensional representation. This representation was later used in combination with the high level features we obtained in the spirit of early fusion.

2.1.2. Classifier

Using all previously mentioned features, we train a 2 layer bidirectional LSTM network with 512 units in each cell. This is followed by 2 fully connected layers each with 512 units. The final softmax layer dimensions were dependent on the sub challenge. The network is trained by minimizing the expected divergence between the classes using cylindrical SGD [11].

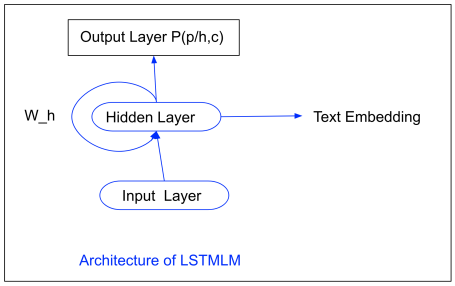

Figure 2: *Architecture for extracting textual embedding. The hidden layer obtained after passing all phonemes of an utterance, is used as embedding of the same utterance. The phonetic decoding is then obtained using a pre-trained acoustic model.*

2.2. Utterance-based classification

Recently, end-to-end approaches have shown impressive results on many speech based tasks [8]. Specifically combinations of CNN and fully connected layers with a global pooling layer have obtained human level recognition rates on speaker and language recognition tasks. The global pooling layer functions as averaging sequential inputs therefore aggregating frame level representations to utterance level. This is advantageous for end-to-end learning.

2.2.1. Extracting high level acoustic representations using Modified SoundNet

SoundNet [7] is a convolutional network operates on raw waveforms and is trained to predict the objects and scenes in video streams at certain points. After the network is trained, the activations of its intermediate layers can be considered a representation of the audio suitable for classification. It has to be noted that SoundNet is a fully convolutional network, in which the frame rate decreases with each layer. Since we need to predict the emotions with reasonable recall, we cannot extract features from the higher layers of SoundNet directly.

The original SoundNet network has seven hidden convolutional layers interspersed with maxpooling layers. Each convolutional layer essentially doubles the number of feature maps and halves the frame rate. The network is trained to minimize the KL divergence from the ground truth distributions to the predicted distributions. In the original SoundNet architecture, the higher layers have been subsampled too much to be used directly for feature extraction. In order to fully exploit the information in the higher layers, we train a fully connected variant of SoundNet (see Fig. 1). Instead of using convolutional layers all the way up, we switch to fully connected layers after the 5th layer. We have also changed the input sampling rate to 16 kHz to match the provided data.

2.2.2. Linguistic features

An informal analysis of the recordings indicated that the content being spoken plays a non trivial role in the valence of the utterance. A simple manifestation of this is the distribution of filled pauses and hesitations in the provided data across the classes. In the Self Assessed Affect dataset, examples belonging to the class 'low' have higher number of such irregularities compared to the other two classes. Note that these features are not extracted for the Crying dataset. Therefore we hypothesize that using an off the shelf phoneme decoder to recognize such

events might be beneficial. For this we first obtain the text at the phoneme level for each of the utterances using a pretrained English ASPire model from the toolbox Kaldi [10]. We then train a Recurrent Neural Network language model on the obtained text at the phoneme level and use the representation at the hidden state as the embedding of the utterance. The architecture is depicted in Fig. 2.

2.2.3. Classifier

We obtain the prediction scores from our models using either a Random Forest Classifier or a one-vs-rest classifier implemented using a binary SVM classifier depending on the performance. It is a known fact that SVM models perform better on sparse data than does trees in general. Therefore depending on the data augmentation techniques, we choose the classifier.

2.3. Data Manipulation & Enhancement

In this section we present various data engineering approaches that make the data more suitable for our models. Specifically, we explore approaches that aim to (a) obliterate the imbalance in class, (b) extract derived features which might help in distinguishing between the classes, (c) downsizing and normalizing on the duration of clips, etc.

2.3.1. Class balancing by data restriction

In order to address the class imbalance present in the original data, we reduce the number of samples used for the classes that are dominant in the dataset. We hypothesize that the skewness of the original data causes low recall for classes that are in minority. Therefore, we study the effects of attempting to artificially balancing the classes by using less samples of dominant classes.

2.3.2. Class balancing by data augmentation

The objective function we minimize in this approach is the expected divergence between the classes. An analysis of the original data points to the imbalance between the classes: For example, in Self Assessed Affect subchallenge, there are almost 3 times less number of examples for the 'low' class compared to the other classes in the training set. To alleviate this, we look at approaches to augment the existing data. Since our model operates on the sequence of frames, we hypothesize that segmenting the audio data into chunks [12] exposes the model to different distributional properties. We obtain 4 times the original data for the class with less number of examples in Self Assessed Affect challenge by chopping the original signal between (0-2), (0-4), (0-6) and (0-8) seconds.

2.3.3. Deriving Speaker Identity

Speaker normalization and adaptation have been widely documented as significant for a speech recognition system. As the original data did not have speakers tagged per utterance, we have tried to do speaker recognition using length normalized i Vector. i-Vectors are low-dimensional representation of GMM supervectors in a single subspace which have been formulated to include all characteristics of speaker and intersession variability. Mathematically, given an observation set X_s, the adapted mean super-vector m_s is modeled as,

$$m_s = m_0 + \mathbf{T}w_s + \theta \quad (1)$$

where m_0 is the Universal Background Model (UBM) supervector, and θ is the residual term which accounts for the variability not captured by T. Following Garcia-Romero and Epsy-Wilson [13], we perform a within class covariance normalization followed by length normalization of i vectors. These have been shown to 'gaussianize' the distribution and improve the performance of PLDA. iVectors have been extracted after log energy based voice activity detection on the utterances. This system was built within framework of Kaldi toolkit[10].

2.3.4. Improving contrastiveness of features

We have tried to improve the contrastive nature among the classes artificially. An informal analysis of the recordings from Self assessed affect subchallenge led to the observation that the utterances with high valence were also relatively at a higher speed compared to the utterances with lower rate. Therefore we increased the rate of speech for the high valence utterances by 10 percent while simultaneously decreasing the rate of speech for low valence utterances by 10 percent. We performed similar perturbations with respect to pitch: boosting the pitch of the samples from 'high' class and lowering the pitch for the samples from 'low'. The samples for 'medium' class have not been subjected to any modification.

2.4. Early Fusion - Combining different representations

We have experimented with a feature level fusion of Soundnet layer 5 and ResNet50 [14] features extracted from the audio files. Resnet has been trained on around 1.28 images from the Imagenet dataset and has a top 5 error of 3.57% beating all other CNN image classifiers. We aim to systematically study the strategies of combining representations from multiple feature extractors.

3. Datasets

3.1. Self-Assessed Affect Recognition

The dataset used in this sub-challenge is the Ulm State-of-Mind in Speech (USoMS). It contains recordings of 100 students. The labels were obtained from the subjects themselves obtaining 3 classes: low, medium, and high.The class distributionfor combined train and dev sets are: 716 high, 698 medium, and 174 low. This highlights skewness in the data distribution.

3.2. Atypical Affect Recognition

The dataset comprised of a total of 10677 audio files out of which there are 3342 training, 4186 validation files and the remaining test files. There are four target classes that pertain to the four emotions - neutral, happy, sad and angry. The distribution of classes is again skewed with 5209, 1708, 516 and 175 being the total numbers of neutral, happy, sad and angry labels on the train and validation sets.

3.3. CRYING

This dataset is obtained from the Cry Recognition In Early Development (CRIED) database. It consists of 5588 vocalizations of 20 infants sampled at 44.1kHz in mpeg format. The objective is to identify three mood-related types of infant vocalization - neutral/positive, fussing and crying. The class distribution is as follows: 2292 cases of neutral/positive mood, 368 files of class fussing and the remaining 178 belonging to the class crying. The dataset is clean of vegetative sounds such as breathing

sounds, smacking sounds, hiccups and so on. Further details about the datasets can be obtained from [15].

4. Experiments

In the following we present the preliminary results obtained using the systems we investigated on the Self Assessed Affect sub challenge. We further present the results of UAR for blind tests for all the three sub-challenges.

4.1. Class balancing by data restriction(System CBR)

We systematically try to reduce the data points from the classes with higher number of examples. The results from this experiment are depicted in Table 1.

Table 1: *UAR for class balancing by data restriction*

Data split			UAR[%]
100% Low	100% High	90% Medium	56.8
		70% Medium	55.0
		40% Medium	52.1
	100% Medium	90% High	**59.1**
		70% High	56.8
		40% High	51.5
All Data			57.2

4.2. Speaker identity based experiments(System SI)

Table 2: *UAR for Speaker identity based experiments*

		Normalization	
	UAR[%]	used	not used
Speaker ID	used	62.2	54.0
	not used	61.1	**64.7**

Since the classifiers we use are discriminative in nature, we experiment with two ways of incorporating speakers or subject specific information:

- (1) We add the identity of the speaker as an extra dimension thus forcing the model to build speaker specific models. For example, in case of decision trees, this forces the model to split at the identity of speaker.
- (2) Normalizing with respect to the speaker, following the procedure typically used in speech recognition.

The results from these experiments have been depicted in table 2.

4.3. Improving contrastiveness of features(System CTR)

We have explored two ways of artificially increasing the contrastiveness of the features, based on observations on the original data. Since the different classes appear to have a different distribution of artifacts such as hesitation, we have tried to use signal processing techniques to further separate the classes. Specifically, we have used festival toolkit [16] to decompose the signal into its spectrum, pitch and then apply class specific modifications to the utterances in the train set. The waveform was reconstructed using the vocoding framework within festvox voice building tools. We have used WSOLA [17] to accomplish duration based manipulations. The results from these experiments are shown in the table 3.

Table 3: *UAR for Emphasis and Data Augmentation Experiments*

Augmentation [%]	UAR [%]
100	54.4
200	58.3
300	57.2
400	58.8

4.4. Blind Test Results and Discussion

The evaluation results on blind test set for the three sub-challenges is mentioned in the table 4. Based on the preliminary experiments, system **SI** appears to achieve a significant boost over the baseline before fusion. This seems plausible due to the nature of task at hand: emotions and intent have been known to be speaker specific. System **CTR** surprisingly does not have the expected gain in performance. We hypothesize that even though the premise of improving the class statistics by enhancing contrastiveness is valid, the manner in which we have performed the manipulation might be flaky. For example, given manipulating pitch might not be the best way to improve contrastiveness when the classes are separated by valence. However, we do see improvements with the Atypical affect subchallenge. Specifically, the recall for the class angry seems to improve with very little augmentation. Another observation with respect to system **CBR** is that the 'neutral' class seems to be very sensitive to any subsampling.

Table 4: *UAR Blind test summary*

Sub-challenge	UAR
Self Assessed Affect	48.3
Atypical Affect	34.2
CRYING	71.406

5. Conclusion

In this work, we present the submission from CMU for COMPARE challenge 2018. We have explored the usage of both low level and high level features aimed at deciphering the intent from acoustics. In the preliminary experiments, higher level features seem to effectively embed the holistic information required for intent recognition. Since the datasets were highly skewed, we have explored various data augmentation and class balancing techniques. It might be beneficial to design architectures that exploit the nature of data and the constraints of the task.

6. Acknowledgments

This work was supported in part by a fellowship from the Portuguese Foundation for Science and Technology through the CMU Portugal Program and the BioVisualSpeech project (grant CMUP-ERI/TIC/0033/2014) to Carla Viegas.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2794–2797.

[3] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2017, pp. 3442–3446.

[5] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *INTERSPEECH 2018 – 18th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings*, 2018.

[6] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[7] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

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

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for hmm-based speech synthesis," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[11] L. N. Smith, "Cyclical learning rates for training neural networks," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 464–472.

[12] A. Agrima, L. Elmazouzi, I. Mounir, and A. Farchi, "Detection of negative emotion using acoustic cues and machine learning algorithms in moroccan dialect," in *International Conference on Soft Computing and Pattern Recognition*. Springer, 2017, pp. 100–110.

[13] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[15] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom *et al.*, "The interspeech 2018 computational paralinguistics challenge atypical and self-assessed affect, crying and heart beats," in *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, 2018, pp. 3442–3446.

[16] A. Black, P. Taylor, R. Caley, and R. Clark, "The festival speech synthesis system," 1998.

[17] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE, 1993, pp. 554–557.