# A survey on the Semi Supervised Learning paradigm in the context of Speech Emotion Recognition

Guilherme Andrade, Manuel Rodrigues, Paulo Novais
Informatics Department/ Computer Science and Tech. Centre
University of Minho
Braga, Portugal
a80426@alunos.uminho.pt, mfsr@di.uminho.pt, pjon@di.uminho.pt

**The area of Automatic Speech Emotion Recognition has been a hot topic for researchers for quite some time now. The recent breakthroughs on technology in the field of Machine Learning opens up doors for multiple approaches of many kinds. However, some concerns have been persistent throughout the years where we highlight the design and collection of data. Proper annotation of data can be quite expensive and sometimes not even viable, as specialists are often needed for such a complex task as emotion recognition, even for humans themselves. The evolution of the semi supervised learning paradigm tries to drag down the high dependency on labelled data, potentially facilitating the design of a proper pipeline of tasks, single or multi modal, towards the final objective of the recognition of the human emotional mental state. In this paper, a review of the current single modal (audio) Semi Supervised Learning state of art is explored, as a away to help future researches refer to when getting to the planning phase of such task.**

***Index Terms*—Machine learning, Semi Supervised Learning, Speech Emotion Recognition, Deep Learning, Classification**

*This work has been supported by FCT - Fundação para a Ciencia e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020*

## I. INTRODUCTION

EMOTIONS are a big part of the human essence. They have the power to completely drive our actions and portrait behaviors that model the human society. Its complex nature is somewhat uncertain as multiple men- tal states can overlap, originating different perspectives regarding each individual [1]. For a relatively long time, a lot of researchers built theories attempting to discretize these mental states. Paul Ekman initially focused on six basic emotions: anger, disgust, fear, happiness, sadness and surprise. He based his conclusions on empirically universally recognized emotions, independent of culture [2]. Robert Plutchik proposed a psychoevolutionary clas- sification approach for emotional responses [3]. He began from the point where he took into account a few basic, primary emotions, anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. From these, different combi- nations would arise, giving origin to more complexsets of emotions, much like the basic colors and their derivatives. Note that it is not possible to combine just any of the basic emotions as some were proposed to be mutually exclusive. The line of work, where emotions become discrete values allowed for discriminative models to appear, and the recog- nition of emotions by a machine found its base to grow. The traditional Automatic Speech Emotion Recognition (ASER) framework divides itself in three components, data collection, feature extraction and classification [4] [5]. Data collection on speech has had multiple approaches as well as categories where they fit, regarding the context and restrictions it has [6]. But the issue focused here revolves around the expensive task of labelling necessary large amounts of data to be fed into state of art algorithms. It does not scale to the required proportions of instances to make a reliable ASER system [7]. So something like Semi Supervised Learning (SSL) could have a big impact on the design of such collection, where costs are severely reduced, making big scale projects viable, and so, more likely to perform ASER in a more naturalistic way.

With today's advances in deep learning, the feature extraction on ASER tasks require less and less human intervention as the neural networks designed can build complex functions that identify important features on data [8]. So the need for hand crafted features has been decreasing in favour of the automatically extracted ones. It opens up a bigger margin for training as the time wasted on feature crafting and selection might grow quite large. With automatic feature extraction, it might be possible to try out the most diverse scenarios [9].

The classification schemes have already been multiple. The more traditional ones consisted on Hidden Markov Model (HMM) [10], Gaussian Mixture Model (GMM) [11], Bayesian Networks [12] and Support Vector Machine (SVM) [13]. However, neural networks have been the best choice for the past few years due to their discriminant ability and efficiency. In this paper, mainly Deep Neural Network theory will be exploited for the purpose of SSL [14] [15].

In Section 2, important concepts will be described as a baseline to SSL. In Section 3, we will review the work done on ASER, under the SSL paradigm, and the respective state of art, to our best knowledge. In section 4, a brief discussion on the future of the research will be exposed as a personal opinion by the authors. In section 5, a conclusion and future research perspectives will be shared.

## II. IMPORTANT NOTIONS

On the SSL paradigm, multiple approaches were ex- ploited so far. [14] distinguishes and defines four categories on SSL algorithms: Consistency training, Proxy label methods, Generative models and Graph based methods.

Consistency training relays on the assumption that perturbations of a certain degree to a data instance would not change its output class, providing a characteristic robustness on unlabelled data and its perturbed versions. From this statement, a model would be trained in such a way that the decision boundary would lie in a low-density region of the data space, making the probability of one example to switch classes after a small perturbation much thinner.

Proxy label methods look to take more of an explicit advantage on labelled data instances combined with some heuristic to assign classes to some instances of unlabelled data, providing information on the modelling of the final function, even if there is some error to the automatically labelled instances. Self training and multiview learning can be seen as two major approaches in this category, where the first focus on the very own model producing the new labels, and the second one uses models trained on different views of the data the come up with the new labels.

Generative models are models with the ability of gen- erating new data instances that follow the training data distribution. This implies that the model should learn important features on the data presented, opening doors for use on other downstream tasks. More importantly, with generative models, a direct approach with concepts such as deep learning [15] have risen, meaning neural networks with multiple layers with a remarkable ability of feature extraction and selection. By using deep learning models as generative models, impressive results have been achieved on many areas of study, making it worth to highlight and focus around.

Graph Based methods look for direct comparison be- tween data instances through specific measures like simi- larity metrics or prior knowledge derived values, on which it attempts to correctly propagate labels through unla- belled data.

It is important to note that algorithms from different categories can complement each other to take advantages from each one on multiple stages.

## III. RESEARCh ON ASER

On the context of ASER, a lot of research has been coming by for the past years, but the technological rev- olution of machine learning started relatively recently, meaning hardware and algorithms are being refined for many different tasks with a considerable rate of success. SSL itself is also very new, on early stages of development, but due to its nature of handling unlabelled data, it might just guarantee a spot on future approaches. Labelled data can present issues like costs, time and difficulty of obtaining. In this case, SSL looks to capitalize on the abun- dance of unlabelled data to improve learning performance, sometimes even outperforming supervised learning itself [7].

The field of ASER has always suffered from the dataset design perspective [16], as both collection and labelling present themselves as expensive tasks and on many cases even not viable, thus strongly justifying approaches based on SLL [17] [19] [20] .

[17] went on to taking advantage of the concept of Generative Adversarial Network (GAN) [21] regarding generative models combined with distribution smoothness taken out of Adversarial Training [22] and Virtual Ad- versarial Training [23] [14]. A performance comparison is made between Semi Supervised Generative Adversarial Network (SSGAN), Smooth Semi Supervised Generative Adversarial Network (SSSGAN) and Virtual Smooth Semi Supervised Generative Adversarial Network (VSSSGAN).

The first model displayed at Fig.1 represents the stan- dard procedure of a SSGAN incorporating the audio signal processed on the OpenSmile toolkit [52]. This model is used as control to compare with the proposed method- ologies. It’s always important to set good comparison frameworks so the value of a certain project won’t be ever undermined under such volatile criteria.

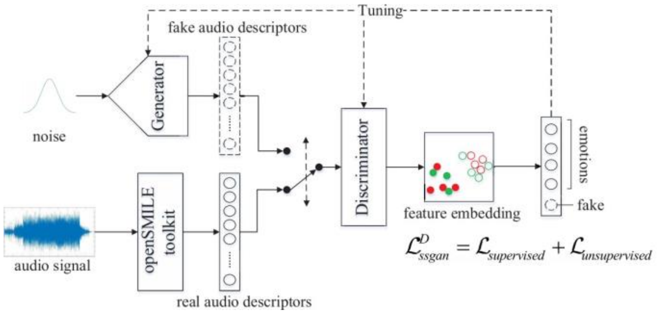


Fig. 1. Semi Supervised Generative Adversarial Network standard framework

The next model, displayed at Fig.2, adds the Adversarial Training component to the equation. The loss function has an extra portion comprehending the adversarial loss. This is one of the two proposed variations on the GAN training.

Note that Adversarial Training utilizes labelled data to smooth the decision boundary area, discarding completely the information contained in the unlabelled portion of data in terms of Adversarial Training itself. Taking the current context into account, it may be perceived as irrelevant, but it’s important to face it up against the use of unlabelled instances on the algorithm.

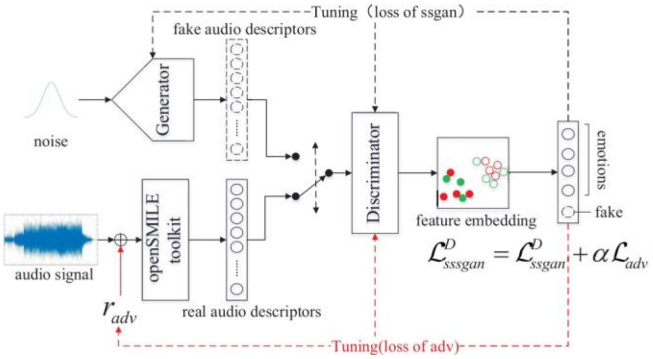


Fig. 2. Proposed Smooth Semi Supervised Generative Adversarial Network framework

The model displayed at Fig.3 represents the applied concept of Virtual Smoothing. The changes on Virtual Adversarial Training, contrarily to standard Adversarial Training, allow it to make use of potentially important features on unlabelled data, complementing the Semi Su- pervised Paradigm and providing value to the approach. This specific model makes use of virtually created labels for unlabelled data to incorporate these instances on the calculation on estimates of adversarial directions.

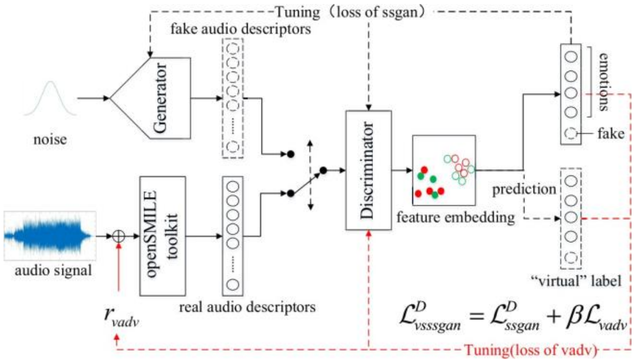


Fig. 3. Proposed Virtual Smooth Semi Supervised Generative Ad- versarial Network framework

The chosen dataset is the IEMOCAP [24], collected by the Speech Analysis and Interpretation Laboratory at the University of Southern California, due its common use in the ASER community. Three other datasets were chosen as unlabelled training sets: the EmoDB [25],AEC [26] and MSP-IMPROV [27].

The chosen features are the baseline from [28], consisted of 16 Low Level Descriptors (LLD) extracted from the raw signals: zero-crossing rate, root mean square, pitch frequency normalized to 500 Hz, harmonics-to-noise rate by autocorrelation function, and Mel-frequency cepstrum coefficients 1-12 in full accordance to HTK-based computation. Then, the first order of the 16 LLD is calculated to append to the feature set. Finally, a set of functionals are applied to the 16 LLD and their first orders, i. e., mean, standard deviation, kurtosis, skewness, minimum and maximum values, relative position and range, as well as the offset and slope of the linear regression line and their mean square error. Finally, a total of (16+16)× 12 = 384 acoustic features are taken into consideration. Fig.4 summarizes the features just described for a more brief overview.

| **LLDs**$(16 \times 2)$ | **Functions(l2)** |
|---|---|
| $(\Delta)$ ZCR | mean |
| $(\Delta)$ RMS Energy | standard deviation |
| $(\Delta)$ F0 | kurtosis, skewness |
| $(\Delta)$ HNR | extremes: value, relative position, range |
| $(\Delta)$ MFCC 1-12 | linear regression: offset, slope, MSE |

Fig. 4. Specified features used in the baseline of the INTERSPEECH 2009 emotion challenge

With their methods, the obtained results were 59.3% and 58.7% Unweighted Average Recall (UAR), at 2400 labelled data, on the proposed methods of SSSGAN and VSSSGAN, respectively, claiming to outperform the till date state of art work on the task of ASER, under the specified context. Fig.5 shows the aggregated results on the experiments done over 300, 600, 1200 and 2400 labels. Both the networks on Adversarial and Virtual Adversarial Training present better results on average over all exper- iments done. This shows the potential fine tuning aspect that it can bring into models in future research, as valuable performance gains might come in crucial later on.

| | | | | |
|---|---|---|---|---|
| SSGAN | $51.6_{\pm 2.0}$ | $54.2_{\pm 1.5}$ | $56.7_{\pm 0.7}$ | $57.8_{\pm 1.6}$ |
| SSSGAN | $51.9_{\pm 1.3}$ | $55.3_{\pm 0.9}$ | $\mathbf{57.8}_{\pm 2.0}$ | $\mathbf{59.3}_{\pm 1.3}$ |
| VSSSGAN | $\mathbf{52.3}_{\pm 2.1}$ | $\mathbf{55.4}_{\pm 1.7}$ | $57.1_{\pm 1.5}$ | $58.7_{\pm 0.9}$ |

Fig. 5. Values obtained on the approaches proposed

The graph on Fig.6 can be used on the purpose of scaling of UAR regarding the number of labelled samples utilized during training.

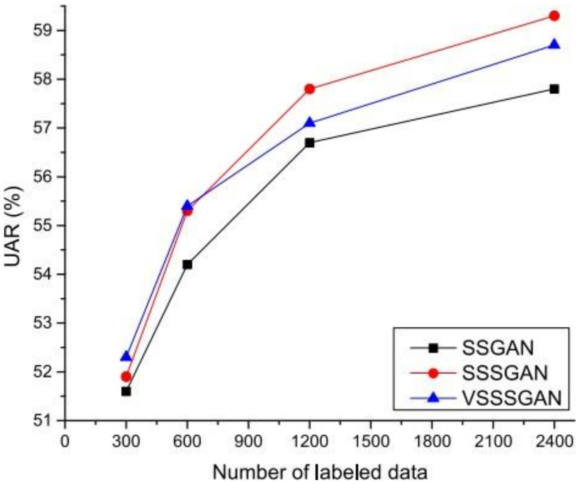


Fig. 6. Graph compiled from results, showing short range scaling on number of labelled instances

An interesting take on the work developed is the use of multiple classes on the discriminator, where a label is assigned to which emotion plus one to determine the authenticity of a sample. The purpose of it was to train the discriminator to also define features that differentiate the considered real classes, instead of only if it comes from a theoretically real distribution or not. This can be seen as a take off based on the previous work on Categorical Generative Adversarial Network (CatGAN) [29] and Semisupervised Generative Adversarial Network (SGAN) [30].

[18] utilized a GAN variant in a different type of pipeline task, the Boundary Equilibrium Generative Adversarial Network (BEGAN) [31], where an AutoEncoder (AE) is used in place of the discriminant and trained on unlabelled data, in an total unsupervised way, by a GAN mechanism. The encoder obtained is then used in the building of a classifier, consisted of convolutions, that is then trained on labelled data. Fig.7 represents an overview on the model designed for the task.

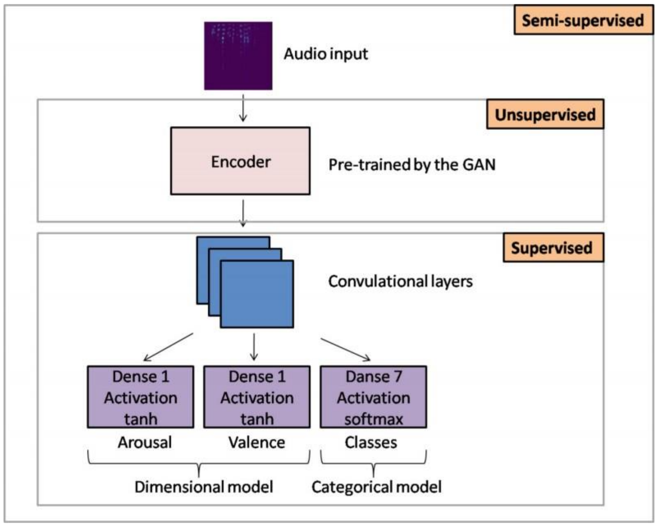


Fig. 7. Abstract overview of the approach proposed

The AE is introduced as a mean of taking advantage of the pieces of unlabelled information. An important note is on the use of the Convolutional Neural Network (CNN) as a classifier, as it represents a supervised component of the process. This type of pipeline may present traces of the main issue on labelled data, but the use of generative model to learn features on unlabelled instances are the highlight, as the challenge of this sort of task is to extract valuable information without the labelling cost associated to it.

The system's performance is measured on its application to three diferent datasets: SAVEE [32], OMG Emotion [33] and EmoDB [25].The reason behind this choice lies around the diversity of design, modality, language and context on which these datasets were conceived on. For the unsupervised module, the LibriSpeech [34] dataset was used due to its size and variability of multiple components like speakers and condition scenarios. Regarding features and preprocessing, it is mentioned that there was a change in the audio frequency to 16 KHz followed by a decomposition into 1 second chunks without overlapping. The raw audio was then converted to a spectrogram with Short Time Fourier Transform [35] of size 1024 and stride 512.

The results obtained on SAVEE, OMG Emotion and EmoDB were shown using diverse metrics so that those could be matched to previous work. For SAVEE, accuracy was used on four different speakers, DC, JE, JK, KL each one achieving, respectively, 80.69%, 80.96%, 80.15% and 82.46%, acomplishing superior results to those of the baseline mentioned in the article [36], displayed on Table I.

| Accuracy averages (%) | | | | |
|---|---|---|---|---|
| | DC | JE | JK | KL |
| Ashwin et al. | 79 | 78 | 76 | 80 |
| I. Pereira et al. | 80.69 | 80.96 | 80.15 | 82.46 |

TABLE I
Results on SAVEE dataset

For the OMG Emotion dataset, the F-score,and Arousal and Valence concordance correlation coefficient (CCC). The values reported for this work were respectively 0.73, 0.17 and 0.16, surpassing the baseline work for the dataset, with results at Table II

| Performance metrics | | | |
|---|---|---|---|
| | F-score | Arousal CCC | Valence CCC |
| Barros et al. | 79 | 78 | 76 |
| OMG emotion | - | 0.29 | 0.36 |
| I. Pereira et al. | 0.73 | 0.17 | 0.16 |

TABLE II
Results on OMG dataset

On the EmoDB dataset, accuracy values are provided as results and goes around 72% accuracy, showcased at Table III.

| Accuracy (%) | |
|---|---|
| Deb and Dandapat | 79 |
| I. Pereira et al. | 80.69 |

TABLE III
Results on EmoDB dataset

This approach builds on models that should try and generalize for most cases and conditions on which speech is collected. It can be seen there are competitive results with previous comparable work done, as it was shown by the authors.

GANs show promising results and the margin of improvement is stretching far away into the horizon [30] , but its complexity and computational cost might be a breaking factor when it comes to choice of algorithm, as there are many factors about GAN that aren't quite standardized yet as well as other intrinsic problems regarding its performance, training and design [38].

[19] recurred to a model architecture built around two main paths, a supervised and an unsupervised one, mod- eling a Denoising AutoEncoder (DAE) [40] [37], with the same root layers, so that the shared parameters would be able to retain information from both labelled and unlabelled data. It is added a pseudo class for recognition of unlabelled data so it can complement the joint loss function. The use of AE, compared to that of GANs, show a lower degree of complexity as the amount of hyper parameters to test from is smaller and training itself can be more stable [39], this might be important on research where the focus is not on the algorithm itself but on the pipeline designed for the task, although it would be essential to keep this kind of trade off in mind. As a claimed novelty, the authors introduce the concept of identity skip connection, where the output of one layer would skip a few of the next to a certain point in the chain of layers, justified by the smooth flowing of information across multiple layers during training, displayed at Fig.8. The idea of skipping connections is not a new thing, as it has already been used in other different contexts [48], but the use of it in the SSL paradigm based models is not such a banality itself.

This works allows for the use of AEs [40] in the semi supervised learning paradigm, claiming to achieve state of the art results on the performance test executed, under the specified circumstances.

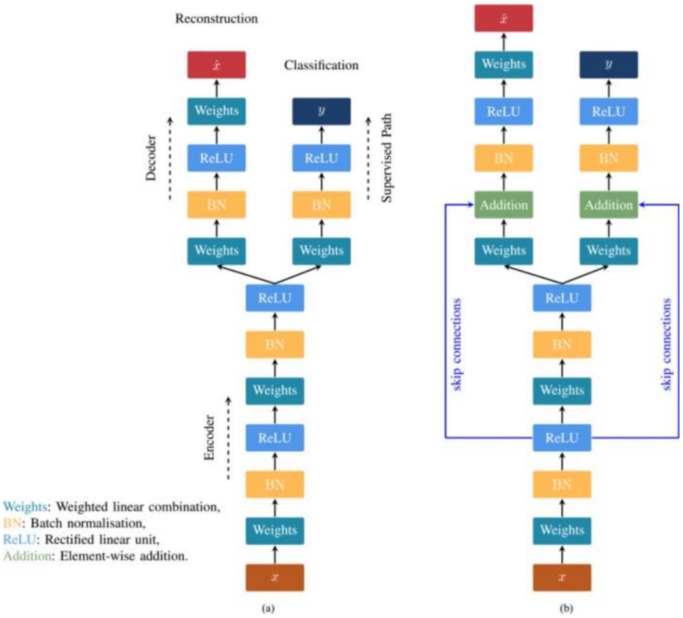
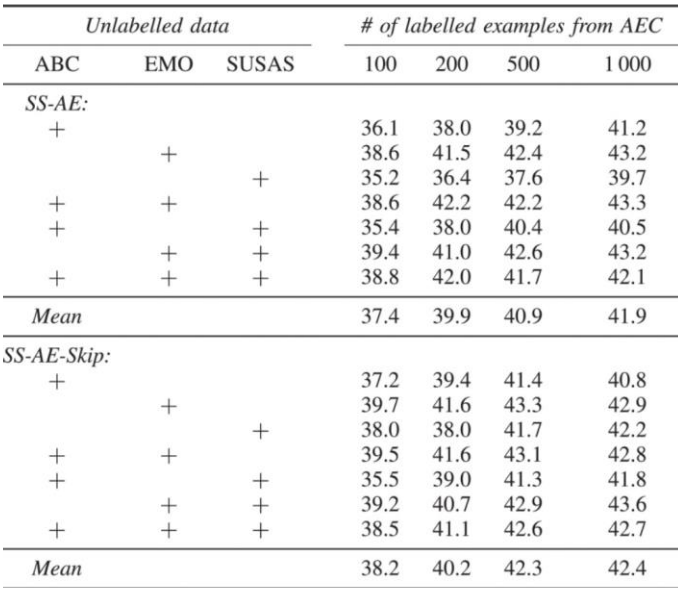


Fig. 8. Proposed model by [19], with two paths sharing the same root layers (left) and the proposed model with the concept of identity skip connections integrated (right)

For the datasets used on the performance evaluation, five different public datasets were chosen: the INTER- SPEECH 2009 baseline dataset FAU AEC [28] [26], the GeWEC [41], ABC [42], the EmoDB [25] and the SUSAS dataset [43].

The chosen set of features was the one used on the [28] displayed previously at Fig.4. There were 16 LLD extracted from the raw signals: zero-crossing rate, root mean square, pitch frequency normalized to 500 Hz, harmonics-to-noise rate by autocorrelation function, and Mel-frequency cepstrum coefficients 1-12 in full accordance to HTK-based computation. Then, the first order of the 16 LLD is calculated to append to the feature set. Finally, a set of functionals are applied to the 16 LLD and their first orders, i. e., mean, standard deviation, kurtosis, skew- ness, minimum and maximum values, relative position and range, as well as the offset and slope of the linear regression line and their mean square error. Finally, a total of (16+16)× 12 = 384 acoustic features are taken into consideration.

As an evaluation metric, the standard Unweighted Av- erage Recall was used. On the AEC dataset, for 100, 200, 500 and 1000 labelled instances of data, respectively,
this work achieved 36.6% , 38.4%, 40.1%, 41.5% with the Semi Supervised Autoencoder solution and 36.5%, 38.5%, 41.1% and 41.8% with the Semi Supervised Autoencoder skip connections concept. These results surpass multiple approaches till the date of this work and stays competitive with previous fully supervised work that, according to this paper, it reaches a top of 45.6%. Further experiments are conducted on unlabelled out of domain data, where the labelled data from training dataset belongs to the AEC and the unlabelled data is from ABC or EMO or SUSAS or a mix of them. The test set still belongs to the AEC dataset. With these trials, the authors are testing the use of unlabelled data on the model's capacity of generalization.



| *Unlabelled data* | | | *# of labelled examples from AEC* | | | |
|---|---|---|---|---|---|---|
| ABC | EMO | SUSAS | 100 | 200 | 500 | 1 000 |
| *SS-AE:* | | | | | | |
| + | | | 36.1 | 38.0 | 39.2 | 41.2 |
| | + | | 38.6 | 41.5 | 42.4 | 43.2 |
| | | + | 35.2 | 36.4 | 37.6 | 39.7 |
| + | + | | 38.6 | 42.2 | 42.2 | 43.3 |
| + | | + | 35.4 | 38.0 | 40.4 | 40.5 |
| | + | + | 39.4 | 41.0 | 42.6 | 43.2 |
| + | + | + | 38.8 | 42.0 | 41.7 | 42.1 |
| *Mean* | | | 37.4 | 39.9 | 40.9 | 41.9 |
| *SS-AE-Skip:* | | | | | | |
| + | | | 37.2 | 39.4 | 41.4 | 40.8 |
| | + | | 39.7 | 41.6 | 43.3 | 42.9 |
| | | + | 38.0 | 38.0 | 41.7 | 42.2 |
| + | + | | 39.5 | 41.6 | 43.1 | 42.8 |
| + | | + | 35.5 | 39.0 | 41.3 | 41.8 |
| | + | + | 39.2 | 40.7 | 42.9 | 43.6 |
| + | + | + | 38.5 | 41.1 | 42.6 | 42.7 |
| *Mean* | | | 38.2 | 40.2 | 42.3 | 42.4 |

Fig. 9. Results on unlabelled multi domain data, on combinations of ABC,EMO and SUSAS, with AEC as labelled data, tested on the AEC test set, from Semi Supervised Autoencoders and Semi Supervised Autoencoders with skip connections

Comparing to the within corpus experiments, it can be drawn that the pick of unlabelled data, in relation to the labelled data, comes up with a big importance. The results observed are still quite impressive, taking into account the mixed domains on multiple datasets. The next step would be to introduce data of the same domain of the labelled data and test set. To the unlabelled data, partitions of the AEC are added on every experiment joint with the other datasets that keep the previous experiment's configurations.

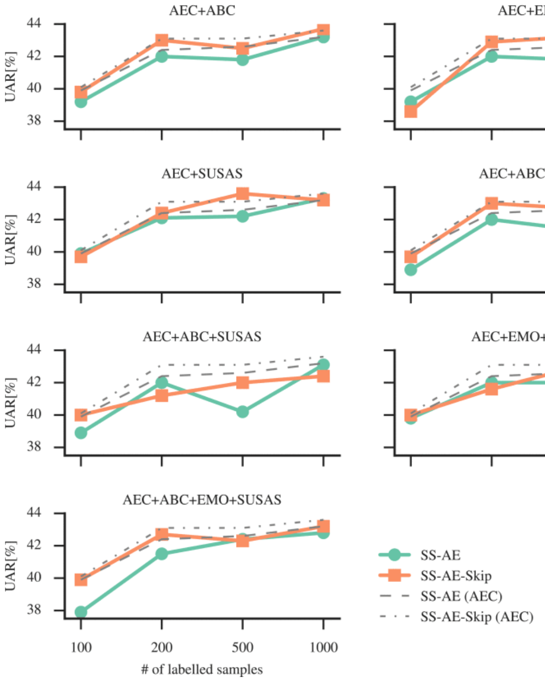


Fig. 10. Results on unlabelled multi domain data, on combinations of ABC,EMO,SUSAS and AEC, with AEC as labelled data, tested on the AEC test set, from Semi Supervised Autoencoders and Semi Supervised Autoencoders with skip connections, and the previous within corpus results

The main purpose of the cross domain unlabelled dataset revolved around checking the effects of using other different domain datasets as a mean of data augmentation to a particular context, in this case, the context of AEC. The results from Fig.10 show that no significant improvement was made on this particular premise but the model still manages to capture important features from the the data as whole, leading to acceptable results on the reality of the research done. Another positive aspect for this work is that the skip connections managed to consistently remain on top of the normal implementation, making it promising for fine-tuning models and push for the extra performance boost in the future.

To take it further, experiments are done on the GeWEC dataset, replacing the role of AEC. The GeWEC introduces a different problem, where it consists of both whispered and normally enunciated speech. A portion of the whispered instances is used as labelled training data where the normally enunciated speech instances are used as test data. As unlabelled data, the ABC, EMO and SUSAS are once again referred to. Three pieces of previous work on GeWEC are used as comparison, two transfer learning methods, uLSIF [49] and DAE [50], and asuper- vised learning method, with Modified Group Delay and SVM [41], with values displayed on Fig.11 as a matter of reference. It is important to try and understand the impact of SSL algorithms have over other more solidified ones like Supervised Learning and such, because the viability of such concept is only valid as long as it holds a considerable amount of success on the task in hand, under the the circumstances that current state of art allows.

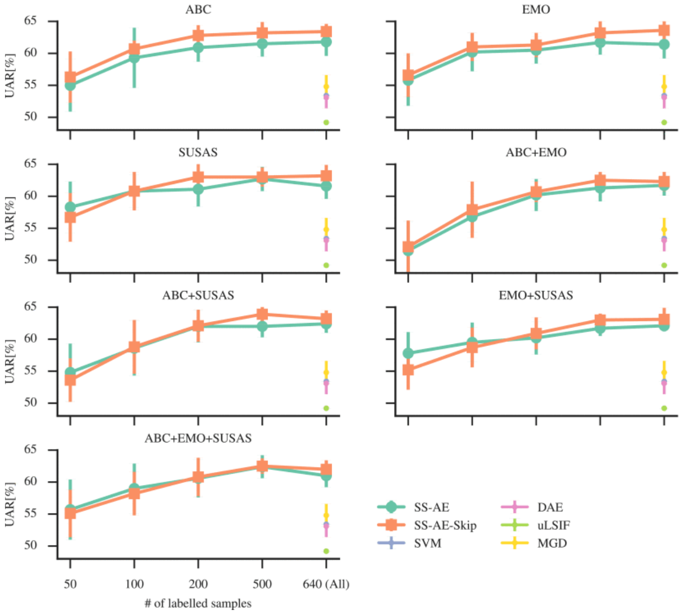


Fig. 11. Results on unlabelled multi domain data, on combinations of ABC,EMO and SUSAS, with GeWEC as labelled data, tested on the GeWEC test set, from Semi Supervised Autoencoders and Semi Supervised Autoencoders with skip connections, and the comparison results on uLSIF, DAE and MGD+SVM

As seen on Fig.11, only a very limited amount oflabels is needed for the proposed model to reach competitive results on the previous work chosen as baseline on the dataset. With the maximum amount of labels, significant improvements are made over the baselines. This can come up as very important detail as it can open up doors for large scale projects on data collection, where unlabelled data still has a great value to the model, but it might be able to function properly with low resources when it comes to labelled data.

A piece of interesting work comes with the use of Ladder Networks [14] [44] on an ASER task by [20]. They claim to use as motivation the fact there are problems on ASER models regarding generalization degree of models, where one model trained on a determined dataset will most likely perform considerably worse at another dataset. Theyalso focus around the issue of multi and single task learning, where they set as target the arousal, valence and domi- nance of utterances and use those as metrics of comparison to previous work. This research is highlighted on multiple but somewhat simple experimental setups that allow for a fair and highly detailed comparison to other methods done on previous research. Training is done on different feature sets (not shown here) as well as architectures, both fully connected and convolutional layers [45]. Fig.12 evidences the building of a standard ladder network, where two encoders and a decoder complement each other to extract information from data.

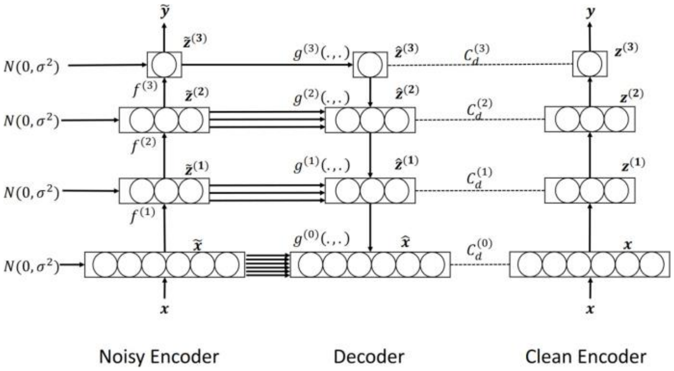


Fig. 12. Schematics on a ladder network model.

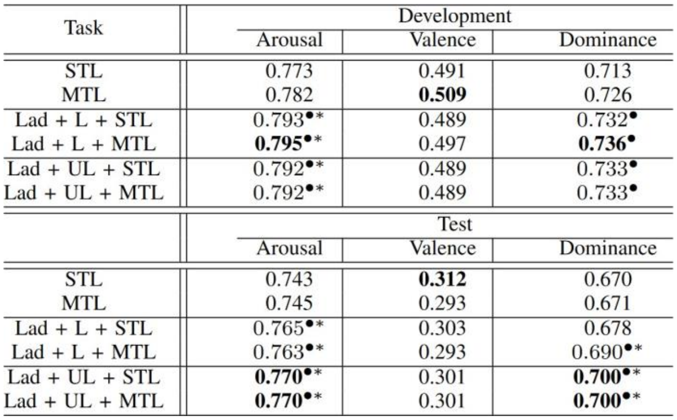
As to the datasets chosen, the MSP-Podcast [46] is used to train, fine-tune and initially evaluate the model. Then, a characteristic cross corpus evaluation is performed on the MSP-Improv [27] and the IEMOCAP [24]. Feature wise, according to the paper, the paralinguistic challenge at Interspeech 2013 [47] was used as reference. The fea- ture set is called the ComParE feature set. LLD are extracted throughout 20 millisecond worth of frames. The LLD include loudness, mel-frequency cepstral coefficients, fundamental frequency, spectral flux, spectral slope, jitter and shimmer. Then, segment-level features are calculated over the LLD mentioned, leading to a fixed dimensional feature vector. This is important due to the model's nature on the input length. These statistics are referred to as high-level descriptors and include various functionals such as the arithmetic and geometric means, standard deviations, peak to peak distances and rise and fall times. The ComParE feature set contains 130 LLD (65 LLD + 65 delta) and 6,373 High Level Descriptors (HLD). This is a specific description used from [47].

The experiments made goes around two baseline results on methods previously implemented on singletask and multitask learning showcased on the article.

For this work specifically, a setup on ladder networks with Single Task Learning (STL) or Multi Task Learning (MTL) as well as with labelled or labelled+unlabelled data, denoted respectively by L and UL, is formed. Later, the model's architecture is altered to CNN on the place of the Fully Connected layers. The evaluation metric is the concordance correlation coefficient, known as CCC, already mentioned previously on this survey. The results within the MSP-Podcast dataset, with fully connected layers, showing in Fig.13, are made on every configuration available.

The following conclusions are drawn directly from the original paper [20].

“On the development set, [...] the best performing sys- tems for ladder network architectures are significantly better than the STL baseline for arousal and dominance.



| Task | Development | | |
|---|---|---|---|
| | Arousal | Valence | Dominance |
| STL | 0.773 | 0.491 | 0.713 |
| MTL | 0.782 | **0.509** | 0.726 |
| Lad + L + STL | 0.793$^{\bullet *}$ | 0.489 | 0.732$^{\bullet}$ |
| Lad + L + MTL | **0.795**$^{\bullet *}$ | 0.497 | **0.736**$^{\bullet}$ |
| Lad + UL + STL | 0.792$^{\bullet *}$ | 0.489 | 0.733$^{\bullet}$ |
| Lad + UL + MTL | 0.792$^{\bullet *}$ | 0.489 | 0.733$^{\bullet}$ |
| | Test | | |
| | Arousal | Valence | Dominance |
| STL | 0.743 | **0.312** | 0.670 |
| MTL | 0.745 | 0.293 | 0.671 |
| Lad + L + STL | 0.765$^{\bullet *}$ | 0.303 | 0.678 |
| Lad + L + MTL | 0.763$^{\bullet *}$ | 0.293 | 0.690$^{\bullet *}$ |
| Lad + UL + STL | **0.770**$^{\bullet *}$ | 0.301 | **0.700**$^{\bullet *}$ |
| Lad + UL + MTL | **0.770**$^{\bullet *}$ | 0.301 | **0.700**$^{\bullet *}$ |

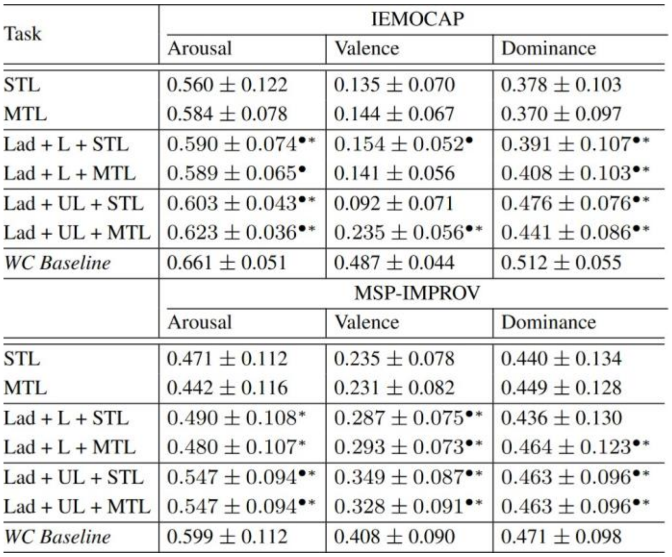
Fig. 13. Within the MSP-Podcast corpus results.

For these emotional attributes, the best performance is achieved by the ladder network implemented with MTL with only labeled data. The results on the test set are very consistent with the trends observed in the development set, demonstrating the generalization of the models. For arousal, the results of the ladder network frameworks are statistically significantly better than the results achieved by both baseline methods. For dominance, the ladder network architectures trained with labeled and unlabeled data lead to statistically significant improvements over both baseline frameworks. The frameworks trained with unlabeled data give the best performance for both arousal and dominance. Under this setting, the ladder network truly utilizes the abundant unlabeled data and generalizes to unseen data.”

Based on these within corpus results, it is quite interest- ing to note that the results on the proposed model trained with labelled+unlabelled data achieve similar or better results than the same model trained fully on labelled data. As verified previously, it is the model that uses unlabelled data that achieves best statistically speaking performance on the test set of the corpus, where the generalization capability of the model is put under stress to some degree and manages to beat both baseline models and its equal trained on fully supervised learning. For a more challenging problem, the cross corpus evaluations, more specifically, the IEMOCAP and the MSP-IMPROV, where the labels on the previously trained model's dataset are adjusted to fit this task, further results are shown in Fig.14. Note that the development set is still from the MSP-Podcast dataset and will not beshown results for it again.

In the cross corpus results, with the settings of the ladder network and unlabelled+labelled data, it can be seen a significant improvement not only over baseline work but over fully labelled data training as well. It is fascinating to realize that the non existence of labels might actually come as beneficial to certain types of models, promoting the power of abstraction of such.

An alternative experiment is done, this time using Con- volutional layers joint with the ladder network configura-



| Task | IEMOCAP | | |
|---|---|---|---|
| | Arousal | Valence | Dominance |
| STL | 0.560 ± 0.122 | 0.135 ± 0.070 | 0.378 ± 0.103 |
| MTL | 0.584 ± 0.078 | 0.144 ± 0.067 | 0.370 ± 0.097 |
| Lad + L + STL | 0.590 ± 0.074$^{\bullet *}$ | 0.154 ± 0.052$^{\bullet}$ | 0.391 ± 0.107$^{\bullet *}$ |
| Lad + L + MTL | 0.589 ± 0.065$^{\bullet}$ | 0.141 ± 0.056 | 0.408 ± 0.103$^{\bullet *}$ |
| Lad + UL + STL | 0.603 ± 0.043$^{\bullet *}$ | 0.092 ± 0.071 | 0.476 ± 0.076$^{\bullet *}$ |
| Lad + UL + MTL | 0.623 ± 0.036$^{\bullet *}$ | 0.235 ± 0.056$^{\bullet *}$ | 0.441 ± 0.086$^{\bullet *}$ |
| *WC Baseline* | 0.661 ± 0.051 | 0.487 ± 0.044 | 0.512 ± 0.055 |
| | MSP-IMPROV | | |
| | Arousal | Valence | Dominance |
| STL | 0.471 ± 0.112 | 0.235 ± 0.078 | 0.440 ± 0.134 |
| MTL | 0.442 ± 0.116 | 0.231 ± 0.082 | 0.449 ± 0.128 |
| Lad + L + STL | 0.490 ± 0.108$^{*}$ | 0.287 ± 0.075$^{\bullet *}$ | 0.436 ± 0.130 |
| Lad + L + MTL | 0.480 ± 0.107$^{*}$ | 0.293 ± 0.073$^{\bullet *}$ | 0.464 ± 0.123$^{\bullet *}$ |
| Lad + UL + STL | 0.547 ± 0.094$^{\bullet *}$ | 0.349 ± 0.087$^{\bullet *}$ | 0.463 ± 0.096$^{\bullet *}$ |
| Lad + UL + MTL | 0.547 ± 0.094$^{\bullet *}$ | 0.328 ± 0.091$^{\bullet *}$ | 0.463 ± 0.096$^{\bullet *}$ |
| *WC Baseline* | 0.599 ± 0.112 | 0.408 ± 0.090 | 0.471 ± 0.098 |

Fig. 14. Cross corpus evaluation with the IEMOCAP and MSP Improv dataset.

tion on a set up of frame-level features (instead of sentence- level features), where spectograms are used and inputted through a CNN, treated like an image. The architecture is composed by four convolutional layers that perform 1D convolutions, a flattening layer and two fully connected layers. Max pooling is done after each convolutional layer. The schematics of the convolutional component of the model is showcased on Fig.15. The model is composed by four convolutional layers, followed by the flatenning layer, two fully connected layers and one linear output layers.

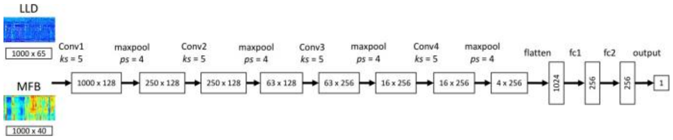
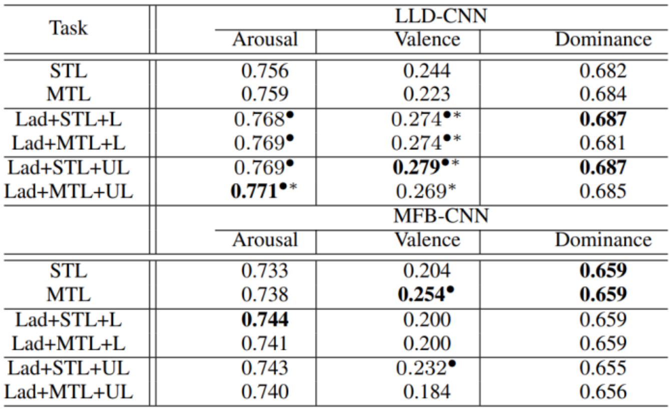


Fig. 15. Proposed CNN composed model to integrate in the ladder network.

The results will present itself as within corpus, on the MSP Podcast. Two different kinds of features are men- tioned and trialed separately , LLD from the ComParE feature set, used originally in this work, and MFB energies with n=40 bands [51]. The time limit on each sample to 10 seconds where every instance crossing the limited is truncated where as instances with less than 10 seconds are padded with zeros. Due to the costs of training the ladder network on frame-level features, the authors chose to impose limitations on the process: the reconstruction loss is only applied only to the fully connected layers afters after the flattening layer; a fixed equal number of labelled and unlabelled data is specified. Both multitasking and singletasking is still taken into account. These models are compared with systems trained with HLD, where such models are not specified.

On the features composed by LLD, the CNN shows progress over the STL baseline on arousal and over both STL and MTL baselines on valence. The performance on dominance is said to be statistically non significant. On the other hand, with Mel-frequency Band (MFB), there is a certain consistency associated to the performance of the model with the LLD in relation to the baselines. It is said the statistically significant improvement is observed over the STL baseline for valence alone with the Lad+STL+UL set up. The authors of this just described work indicate that implementing the reconstruction loss on the CNN component would likely increase performance even further. The CNN approach on frame-level features presents itself as a different computational perspective. The ladder network set up already shows an impressive capacity of abstraction towards speech and the recognition of the components presented on this work. The fact that the authors can identify possible improvements towards the next step is very important, as it shows as a baseline where future research on ladder networks, on multiple architectures, can stand and rise to the expectation behind

Artificial Intelligence (AI).



| Task | LLD-CNN | | |
|---|---|---|---|
| | Arousal | Valence | Dominance |
| STL | 0.756 | 0.244 | 0.682 |
| MTL | 0.759 | 0.223 | 0.684 |
| Lad+STL+L | 0.768$^{\bullet}$ | 0.274$^{\bullet *}$ | **0.687** |
| Lad+MTL+L | 0.769$^{\bullet}$ | 0.274$^{\bullet *}$ | 0.681 |
| Lad+STL+UL | 0.769$^{\bullet}$ | **0.279**$^{\bullet *}$ | **0.687** |
| Lad+MTL+UL | **0.771**$^{\bullet *}$ | 0.269$^{*}$ | 0.685 |
| | MFB-CNN | | |
| | Arousal | Valence | Dominance |
| STL | 0.733 | 0.204 | **0.659** |
| MTL | 0.738 | **0.254**$^{\bullet}$ | **0.659** |
| Lad+STL+L | **0.744** | 0.200 | 0.659 |
| Lad+MTL+L | 0.741 | 0.200 | 0.659 |
| Lad+STL+UL | 0.743 | 0.232$^{\bullet}$ | 0.655 |
| Lad+MTL+UL | 0.740 | 0.184 | 0.656 |

Fig. 16. Results on frame-level features, through the CNN compo- nent model

## IV. MODALITY TRADE off

All the pieces of work mentioned here are composed by a single modal task. The use of various modalities have been exploited for some time and recently, interesting results were achieved on such paradigm [53].

However, some issues might come up on multi modal work. Many types of modalities other than voice can be identified on the problem of Emotion recognition, such as words used, facial expression, body language.

Firstly, depending on the modalities chosen, the over- head of collecting, processing and predicting can increase dramatically due to each modality demand, and even cause considerable performance and logistic issues on the devices used. For example, video recording brings many problems as a clean shot of facial expressions or body language would probably be needed for effective use; if we're talking about an application used in an office, it just might work as the environment reunites all the conditions needed for clean data, but on everyday use, it might raise more issues than it solves, due to the criteria demanded on recordings such as angle, quality and duration. If certain goals are made, such as portability, ease of use and low hardware requirements, then a lot of the viability of the project itself can get compromised.

Another particularly characteristic issue is the design of such tractable models on multiple modalities, more specifi- cally, the capacity of achieving satisfactory performance on means available. For example speech recognition is a rather explored area and Machine Learning algorithms already achieved very satisfactory results [54]. Such systems as well as the data used are often not open source. In theory, they could be integrated with Natural Language Processing systems designed for the task of ASER but it could raise copyright issues. It would be also possible to build one from scratch, but most likely incapable of achieving the desired results due to the demanding on computation power and data storage/usage. Even if such capacity is available, the high complexity needed prompts to the issues reported just previously on overhead of execution and adds another one where, with the complexity increase, the unpredictability of the model also increases [55], which contributes to the instability of the model itself. It takes experience to design such systems, even more so if adopted a multitask learning approach (explored on this survey), so it becomes harder to control the outcome of theexper- iment.

Is the performance achieved on ASER good enough to justify the use of an above average set of hardware, most likely unavailable for most of the common users? The same question can be turned around where we question onhow low can the performance be dropped to be available on a larger scale. It goes around efficiency, juggling resources, a trade off involving complexity, performance and accessi- bility;

Even if the hardware used is the state of the art and there is no worrying about usage scaling, can the model achievethe high expectations on accuracy for ASER? Even with the hardware at hand, it is still difficult to design a proper system as many specialist spend months or even years developing state of art projects.

Keep in mind that all this revolves around the final purpose of the system designed. Medical or legal purposes? Then performance will most likely be favoured over usage scaling. Emotional self awareness or self control purposes? Then usage scaling can be the priority as the system looks on to fit the common citizen devices.

Many more variations and custom purposes can be extracted from the rationalization above but it is every important to have it clear when designing the final system.

With the evolution of hardware, we might make it to a point where such trade off on complexity of modalities won’t be taken into account as much as it should at the present date, and no doubt that multi modal tasks will dominate single modal ones in the future, but as long as there are big limitations, production and consumption wise, careful planning is needed on the deployment of models.

## V. Conclusion

Many different approaches on the ASER task have been coming up on the last passing years. The development on Deep learning opened up many doors to different complex tasks, and more and more diverse pipelines have been showing up in the area. As algorithms get more complex, more computational power is required which may in many cases, together with time, be a limiting factor for many experiments. Never the less, progress is being made on many ends, such as computer science, psychology, neurology, all contributing with improvements toward the final goal on the field of ASER.

Semi Supervised Learning is relatively recentparadigm that looks to build on the huge of amounts of data getting collected and alleviate the costs on labelling and hand processing that same data. As it was shared during this article, it has been showing amazingly promising results, many times staying even or better than fully supervised learning. This might allow for large scale projects to run smoother as the dependence on labelled instances goes down.

An important aspect is the managing of comparison frameworks. It is noticeable the differences on performance measures in between multiple lines of work. Of course it might be considered that different purposes and objectives are set for each piece of research done, but overall, for a steady and healthy grow, solid benchmarking foundations need be settled. Such thing might only be possible with the time, as many different valid approaches present them- selves every other day.

There is still a long path to go on reaching a somewhat reliable performance on ASER models, but newproposals with improved results keep showing up every day, meaning researchers all around the world are committed to find way where speech emotion recognition by machine learning is crystal clear viable.

## References

[1] Frijda, N. H. (2004). Emotion and action. In A. S. R. Manstead, N. Frijda, & A. Fischer (Eds.), Feelings and emotions: The Amsterdam symposium(pp. 158–173). Cambridge, England: Cambridge University Press.

[2] Shiota, Michelle N. (2016). ”Ekman’s theory of basic emo- tions”. In Miller, Harold L. (ed.). The Sage encyclopedia of theory in psychology. Thousand Oaks, CA: Sage Publications. pp. 248–250. doi:10.4135/9781483346274.n85.

[3] Plutchik, Robert (2000). Emotions in the practice of psychotherapy: clinical implications of affect theories. Washington, DC: American Psychological Association. doi:10.1037/10366-000. ISBN 1557986940. OCLC 44110498.

[4] Aeluri, Pramod Vijayarajan, V. (2017). Extraction of Emo- tions from Speech-A Survey. International Journal of Applied Engineering Research. 12. 5760-5767.

[5] Pathak, S., Kolhe, V.L. (2016). Emotion Recognition from Speech Signals Using Deep Learning Methods. Imperial journal of interdisciplinary research, 2.

[6] Drakopoulos, Georgios Pikramenos, George Spyrou, Evagge- los Perantonis, Stavros. (2019). Emotion Recognition From Speech: A Survey. 10.5220/0008495004320439.

[7] Singh, Aarti Nowak, Robert Zhu, Xiaojin. (2008). Unlabeled data: Now it helps, now it doesn’t. NIPS. 1513-1520.

[8] Khalil, Ruhul Amin Jones, Edward Babar, Mohammad Jan, Tariqullah Zafar, Mohammad Alhussain, Thamer. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2936124.

[9] Shaheen, Fatma Verma, Brijesh Asafuddoula, Md. (2016). Impact of Automatic Feature Extraction in Deep Learning Architecture. 1-8. 10.1109/DICTA.2016.7797053.

[10]L. Rabiner and B. Juang, ”An introduction to hidden Markov models,” in IEEE ASSP Magazine, vol. 3, no. 1, pp. 4-16, Jan 1986, doi: 10.1109/MASSP.1986.1165342.

[11]Reynolds, Douglas. (2008). Gaussian Mixture Models. Ency- clopedia of Biometrics. 10.1007/978-0-387-73003-5 196. [12]Heckerman, David. (2008). A Tutorial on Learning With Bayesian Networks. 10.1007/978-3-540-85066-3 3.

[13] Ben-Hur, Asa Weston, Jason. (2010). A User’s Guide to Sup- port Vector Machines. Methods in molecular biology (Clifton, N.J.). 609. 223-39. 10.1007/978-1-60327-241-4 13.

[14] Ouali, Yassine Hudelot, Céline Tami, Myriam. (2020). An Overview of Deep Semi-Supervised Learning.

[15] Goodfellow, I., Bengio, Y.„ Courville, A. (2016). Deep Learn- ing. MIT Press.

[16] Ayadi, Moataz Kamel, Mohamed S. Karray, Fakhri. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 44. 572-587. 10.1016/j.patcog.2010.09.020.

[17] Zhao, Huan Yufeng, Xiao Zhang, Zixing. (2020). Robust Semisupervised Generative Adversarial Networks for Speech Emotion Recognition via Distribution Smoothness. IEEE Ac- cess. PP. 1-1. 10.1109/ACCESS.2020.3000751.

[18] Pereira, Ingryd Santos, Diego Maciel, Alexandre Barros, Pablo. (2018). Semi-supervised Model for Emotion Recognition in Speech: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I. 10.1007/978-3-030-01418-6 77.

[19] Deng, Jun Xu, Xinzhou Zhang, Zixing Frühholz, Sascha Schuller, Björn. (2017). Semi-Supervised Autoen- coders for Speech Emotion Recognition. IEEE/ACM Trans- actions on Audio, Speech, and Language Processing. PP. 1-1. 10.1109/TASLP.2017.2759338.

[20] Parthasarathy, Srinivas Busso, Carlos. (2019). Semi-Supervised Speech Emotion Recognition with Ladder Networks.

[21] Goodfellow, Ian Pouget-Abadie, Jean Mirza, Mehdi Xu, Bing Warde-Farley, David Ozair, Sherjil Courville, Aaron Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neu- ral Information Processing Systems. 3. 10.1145/3422622.

[22] Goodfellow, Ian Shlens, Jonathon Szegedy, Christian. (2014). Explaining and Harnessing Adversarial Examples. arXiv 1412.6572.

[23] Miyato, Takeru Maeda, Shin-ichi Koyama, Masanori Ishii, Shin. (2017). Virtual Adversarial Training: A Regularization Method for Supervised and Semi-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 10.1109/TPAMI.2018.2858821.

[24] Busso, Carlos Bulut, Murtaza Lee, Chi-Chun Kazemzadeh, Abe Mower Provost, Emily Kim, Samuel Chang, Jeannette Lee, Sungbok Narayanan, Shrikanth. (2008). IEMOCAP: In- teractive emotional dyadic motion capture database. Language Resources and Evaluation. 42. 335-359. 10.1007/s10579-008-9076-6.

[25] Burkhardt, Felix Paeschke, Astrid Rolfes, M. Sendlmeier, Walter Weiss, Benjamin. (2005). A database of German emo- tional speech. 9th European Conference on Speech Communi- cation and Technology. 5. 1517-1520.

[26] Batliner, Anton Steidl, Stefan Noeth, Elmar. (2008). Re- leasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus.

[27] Busso, Carlos Parthasarathy, Srinivas Burmania, Alec Abdel- Wahab, Mohammed Sadoughi, Najmeh Mower Provost, Emily. (2016). MSP-IMPROV: An Acted Corpus of Dyadic In- teractions to Study Emotion Perception. IEEE Transactions on Affective Computing. 8. 1-1.10.1109/TAFFC.2016.2515617.

[28] Schuller, Björn Steidl, Stefan Batliner, Anton. (2009). The Interspeech 2009 Emotion Challenge. Proc. Interspeech. 312- 315.

[29] Springenberg, Jost. (2015). Unsupervised and Semi-supervised Learning with Categorical Generative Adversarial Networks. [30]Odena, Augustus. (2016). Semi-Supervised Learning with Gen- erative Adversarial Networks.

[31]Berthelot, David Schumm, Tom Metz, Luke. (2017). BEGAN: Boundary Equilibrium Generative Adversarial Networks. [32]Jackson, Philip ul haq, Sana. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database.

[33] Barros, Pablo Churamani, Nikhil Lakomkin, Egor Siqueira, Henrique Sutherland, Alexander Wermter, Stefan. (2018). The OMG-Emotion Behavior Dataset. 1-7. 10.1109/IJCNN.2018.8489099.

[34] Panayotov, Vassil Chen, Guoguo Povey, Daniel Khudanpur, Sanjeev. (2015). Librispeech: An ASR corpus based on public domain audio books. 5206-5210. 10.1109/ICASSP.2015.7178964.

[35] Sejdic, Ervin Djurovic, Igor Jiang, Jin. (2009). Time–frequency feature representation using energy concentration: An overview of recent advances. Digital Signal Processing. 19. 153-183. 10.1016/j.dsp.2007.12.004.

[36]T S, Ashwin Saran, Sai Reddy, G. (2016). Video Affec- tive Content Analysis based on multimodal features using a novel hybrid SVM-RBM classifier. 416-421. 10.1109/UPCON.2016.7894690.

[37] Vincent, Pascal Larochelle, Hugo Bengio, Y. Manzagol, Pierre-Antoine. (2008). Extracting and composing robust fea- tures with denoising autoencoders. Proceedings of the 25th International Conference on Machine Learning. 1096-1103. 10.1145/1390156.1390294.

[38] Saxena, Divya Cao, Jiannong. (2020). Generative Adversarial Networks (GANs): Challenges, Solutions, and Future Direc- tions.

[39] Munjal, Prateek Paul, Akanksha Krishnan, Narayanan. (2019). Implicit Discriminator in Variational Autoencoder.

[40] Hinton, G.E. Salakhutdinov, R.R.. (2006). Reducing the Di- mensionality of Data with Neural Networks. Science (New York, N.Y.). 313. 504-7. 10.1126/science.1127647.

[41] Deng, Jun Xu, Xinzhou Zhang, Zixing Frühholz, Sascha Grandjean, Didier Schuller, Björn. (2017). Fisher Kernels on Phase-Based Features for Speech Emotion Recognition. 10.1007/978-981-10-2585-3 15.

[42] Schuller, Björn Arsic, Dejan Rigoll, Gerhard Wim- mer, Matthias Radig, Bernd. (2007). Audiovisual Behav- ior Modeling by Combined Feature Spaces. 2. II-733 . 10.1109/ICASSP.2007.366340.

[43] Hansen, J., Bou-Ghazale, S.E. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. EUROSPEECH.

[44] Valpola, Harri. (2014). From neural PCA to deep unsupervised learning. From Neural PCA to Deep Unsupervised Learning. 10.1016/B978-0-12-802806-3.00008-7.

[45] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. The MIT Press.

[46] Mariooryad, Soroosh Lotfian, R. Busso, Carlos. (2014). Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora. Proceedings of the Annual Conference of the International Speech Commu- nication Association, INTERSPEECH. 238-242.

[47] Schuller, Björn Steidl, Stefan Batliner, Anton Vinciarelli, Alessandro Scherer, Klaus Ringeval, Fabien Chetouani, Mohamed Weninger, Felix Eyben, Florian Marchi, Erik Mortillaro, Marcello Salamin, Hugues Polychroniou, Anna Valente, Fabio Kim, Samuel. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. Proceedings of the Annual Confer- ence of the International Speech Communication Association, INTERSPEECH. 148-152.

[48]K. He, X. Zhang, S. Ren and J. Sun, ”Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[49]Hassan, Ali Damper, Robert Niranjan, Mahesan. (2013). On Acoustic Emotion Recognition: Compensating for Covariate Shift. Audio, Speech, and Language Processing, IEEE Trans- actions on. 21. 1458-1468. 10.1109/TASL.2013.2255278.

[50]J. Deng, Z. Zhang, F. Eyben and B. Schuller, ”Autoencoder- based Unsupervised Domain Adaptation for Speech Emotion

Recognition," in IEEE Signal Processing Letters, vol. 21, no. 9, pp. 1068-1072, Sept. 2014, doi: 10.1109/LSP.2014.2324759.

[51]Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2741-2745, doi: 10.1109/ICASSP.2017.7952655.

[52] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy, ISBN 978-1-60558-933-6, pp. 1459-1462, 25.-29.10.2010.

[53] Latif, Siddique Rana, Rajib Khalifa, Sara Jurdak, Raja Epps, Julien Schuller, Björn. (2019). Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion.

[54] Tulshan, Amrita Dhage, Sudhir. (2019). Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa: 4th Inter- national Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers. 10.1007/978-981-13-5758-9_17.

[55] Yampolskiy, Roman. (2019). Unpredictability of AI.