A New Perspective on Smiling and Laughter Detection: Intensity Levels Matter

Hugo Bohy[§]
ISIA Lab
University of Mons
Mons, Belgium
hugo.bohy@umons.ac.be

Kevin El Haddad[§]
ISIA Lab
University of Mons
Mons, Belgium
kevin.elhaddad@umons.ac.be

Thierry Dutoit
ISIA Lab
University of Mons
Mons, Belgium
thierry.dutoit@umons.ac.be

***Abstract*—Smiles and laughs detection systems have attracted a lot of attention in the past decade contributing to the improvement of human-agent interaction systems. But very few considered these expressions as distinct, although no prior work clearly proves them to belong to the same category or not. In this work, we present a deep learning-based multimodal smile and laugh classification system, considering them as two different entities. We compare the use of audio and vision-based models as well as a fusion approach. We show that, as expected, the fusion leads to a better generalization on unseen data. We also present an in-depth analysis of the behavior of these models on the smiles and laughs intensity levels. The analyses on the intensity levels show that the relationship between smiles and laughs might not be as simple as a binary one or even grouping them in a single category, and so, a more complex approach should be taken when dealing with them. We also tackle the problem of limited resources by showing that transfer learning allows the models to improve the detection of confusing intensity levels.**

***Index Terms*—laugh, smile, multimodal, transfer learning, laughter detection, smiles detection, intensity levels, arousal levels**

I. Introduction

With the growth of virtual agents and other human-centric applications, detection systems for nonverbal expressions have been attracted the attention of the research community in the past decade, especially concerning smiles and laughter (S&L). This is due to the importance of these expressions in human communications. Indeed they not only have emotional but also social functionalities: according to [1], S&L tend to happen during interactions with others rather than alone. They can control the flow of a conversations: change the current topic [2], [3] or encourage a person to carry on speaking [4]. Laughter can be contagious to listeners and can lighten the mood of the conversation [5]. S&L are also expressions used frequently in human-human interactions. Indeed the ICSI corpus [6] counts about 10% of its total verbalizing time as being laughter [7], [8] and Chovil in [9] reports not even considering smiles in the study due to its high frequency of occurrence in the data compared to other expressions.

It is therefore not surprising that the detection of laughs or smiles became an attractive field rising alongside the deep learning technologies and AI-backed human-agent interaction systems.

[§]Equal contribution

A plethora of work can be found on smile detection. We estimate that the vast majority of them are based on the visual cue as we could find very few work based on other modalities [10]–[12], notably the audio cue was rather absent from the state-of-the-art although smiles were proven to be recognizable audible [13]–[15]

Fewer work can be found on laughter detection. They focus on the audio and the visual modalities individually but also in multimodal approaches. Kantharaju et. al. in [16] present an automatic detection of different categories of laughter using audio-visual data. The authors in [17] use full-body motion capture data to detect laughter while [18] investigates the laughter detection based on audio and facial motion capture.

Surprisingly, very few work can be found where S&L are considered as two distinct expressions, and none of them attempts to classify/detect them as different entities. Indeed, even though the authors in [19] annotate them as two expressions in their work, they build classifiers considering them as the same class. The authors in [20] propose a system classifying smiles vs non-smiles based on the visual cue and laugh/non-laugh based on a single modality and on multimodal data, but no smile/laugh discrimination is presented. One reason for this might be the difficulty for the models to learn the differences between smiles and laughs, especially given the limited amount of resources available. Another reason might be the common representation for some, of smiling being a less intense expression of laughter or both even both being the same expression, which is to the best of our knowledge, unproven yet.

Although S&L are commonly defined as the former being a purely facial expression while the latter being an audio-visual one, no clear answer can be found in the literature as of the relationship between these two: are they the same expression at different intensity levels ? Are they distinct expressions although smiles can be perceived in laughs ? Or does it depend on the context/situation ? Ruch and Ekman observe in [21] that enjoyment smiles were involved in laughter while Trouvain's perception study in [22] revealed that some participants preferred to categorize speech-laughs into smiles and laughs (speech-laughs being laughter-speech co-articulation phenomenon involving laughter intermingling with speech). In [23] the authors present existing relationships between S&L

978-1-6654-5908-2/22/$31.00

on several levels, suggesting a common ancestry of these expressions and therefore that at least some relationship exist between them.

Since no study showing proof of smiles and laughs being separated expressions exists, nor a smiling-laughter continuum established, we consider it important, in this work, to approach S&L as two distinct expressions. Doing this makes it easier to analyse the common and differentiating points, and allows us to leverage the feature extraction power of deep learning to further examine the intrinsic problems of smiles and laughter detection systems.

In this paper, we present several contributions. First, we propose a first step towards an efficient S&L detection system discriminating between laughs and smiles, as opposed to systems detecting a single category of smiles and laughs, by building and analysing classifiers based on the audio and the visual cues. Given the observations mentioned above regarding the relationship between S&L , we push the analyses further by examining the behavior of models with regard to S&L intensity levels, while being trained without any supervised knowledge of these intensities. These analyses reveal that not all S&L levels are equal in the eyes of deep learning systems. This understanding might change the simplistic approaches taken for building laughter or smiling detection systems.

S&L are difficult to collect in naturalistic setups and to annotate. This difficulty to access accurately annotated data represents a challenge to develop efficient systems and a significant barrier of entry for new contributions in the field. Indeed this lack of well annotated data makes it more difficult to leverage the efficiency of deep learning methods, and thus stalls the improvement of S&L detection systems. So, as a second main contribution, we apply transfer-learning by leveraging the knowledge learned with speech data by the models to improve the efficiency and generalisation capabilities of the models for S&L detection.

The following is a more detailed summary of our main contributions in this work:

1) we propose the first deep learning-based S&L classification system that we know of that considers S&L as two different entities
2) a deeper analysis than what can be found in the literature of the model's behavior showing that deep learning-based systems implicitly take into account the S&L intensity levels in their learning process without being trained with any explicit knowledge of them
3) we show that transferring knowledge from visual lipreading task and from audio word classification improves the performance of the models and help tackle the problem of limited resources

The paper is organised as follows: in Section II we present the datasets used for our experiments. Section III contains the description of the model architectures for the audio and the visual modalities, as well as for the fusion of both. We describe the experimental protocol followed to train the models in Section IV and we discuss the results of the aforementioned experiments in Section V.

II. Dataset

The data used here are subsets of the Nonverbal Dyadic Conversation on Moral Emotions (NDC-ME) [24], and of the IFA Corpus (IFADV) [25] for which the S&L were annotated. The S&L were segmented and the intensity level was added to each segment. We followed the annotation protocol described in [26]. The annotations were made using the ELAN software [27] by two annotators on average and are available to the community [28].

NDC-ME is an audiovisual collection of dyadic interactions focusing on the emotions expressed during speaker-listener interactions. The subset we use is distributed in 17 dyadic interactions split between 10 male and 4 female individuals, with 7 male-male, 6 male-female and 4 female-female pairs. During these interactions, each duo discusses emotional topics introduced by an open question. Since some of those interactions are not fully annotated, the total duration of annotated data is about 90 minutes with an unbalanced distribution between individuals.

IFADV is also a collection of audio-visual recordings of dyadic conversations. The subset we used contains 23 dyadic interactions of 15 male and 28 female individuals with 4 male-male, 8 male-female and 11 female-female pairs of interactions. The annotations cover only the first two minutes of each file, leading two around 46 minutes of annotated data.

The laughs intensities are divided in three levels (low, medium and high) and the smiles intensities in four (subtle, low, medium and high). According to the authors of the previously mentioned papers, the subtle level was added to capture all the levels of smiles even the ones that are normally left out because of the difficulty to annotate them: subtle smiles co-occurring with other expressions for instance. A third class, referred to as the **None** class, includes all segments of the recordings that contain neither laughter nor smiles, such as neutral expressions and speech. Therefore we ended up with three main classes **Laughs**, **Smiles** and **None**, which will be used for training without taking into account the intensity levels.

III. Classifiers/Detection systems

In this section, we describe the deep learning architectures used for classification. Since S&L are distinguishable through both audio and visual, we first separated our system in two models, one per modality. We used the audio and visual models proposed in [29] and [30], which present audio word recognition/classification and visual lip reading applications. We then perform the fusion of both modalities. Fig. 1 displays a schematic representation of the system architecture.

There were several intuitions behind these choices. The first one is the fact that modalities, as mentioned before, could be a major factor helping a model discriminate between smiles, laughs and everything else. This type of architecture already takes this aspect into account and successfully applies it on speech, making it a good candidate for our experiments. The second intuition is the fact that our classification learning could potentially benefit from knowledge learned from speech.

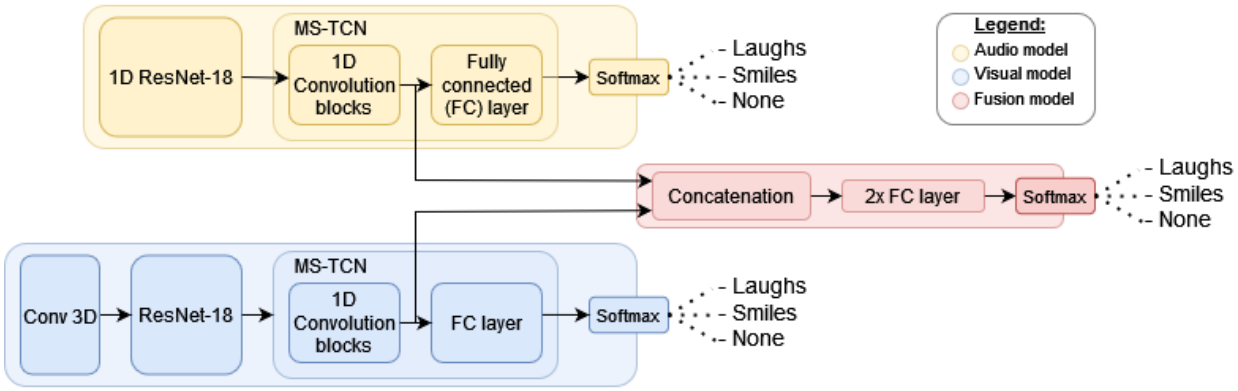


Fig. 1: Architectures of the audio model, the visual model and the fusion of both.

Indeed by learning to recognise speech pattern, these models learn audio and visual features as well as pattern that could also be useful for classifying smiling, laughter and others. In fact, given the limited amount of S&L data available, a transfer learning technique is a good option to optimise learning with a small dataset. Transfer learning based on facial recognition has already been used successfully in the context of smiles detection [31].

A. Audio Cue

Our audio model is a mix between the backbone architecture proposed in [29] for speech audio data and the frontend layers proposed in [30]. The model consists of a modified 18-layer ResNet backbone using 1D kernels fed to a multi-scale temporal convolutional network (MS-TCN). We used 1D operations since audio waveforms are unidimensional signals. The MS-TCN is a multi-layered combination of 1D convolutions with batch normalisation and PReLU activation layers, designed in such a way that it models both short and long term temporal information simultaneously. The final layer of the MS-TCN block is a fully connected (FC) layer with a softmax activation function to perform the classification.

B. Visual Cue

We used the architecture proposed in [30] for the visual recordings. The standard ResNet18 backbone was modified to have its first layer to be a 3D convolution of kernel size $5 \times 7 \times 7$. This layer extracts spatiotemporal features from the sequences of image given as input to the model. The backbone output is fed to the MS-TCN block of layers described above. In the same fashion as for the audio, we used a softmax activation function after the MS-TCN block to classify data as **Laughs**, **Smiles** or **None**.

C. Fusion

We performed the fusion by feeding the output of each modality to a network of two fully connected layers of size 1024 and 3 respectively. We froze the weights of each modality and we used the concatenation of their output as input of the fusion models.

IV. Experiments

For our experiments, we extracted from the NDC-ME data 8,352 videos with 1.22s windows overlapping by 0.4s and split them with respect to their classes giving us 446 **Laughs**, 4,858 **Smiles** and 3,048 **None**. We re-sampled the audio data at 16 kHz and we converted the visuals from RGB to grey scale and then extracted a 96 x 96 pixel region of interest (ROI) around the mouth. We then distributed 70% of each class as training data, 15% as validation data and 15% as test data. Classes were balanced during partitioning using a random weighted batch sampler with each class given a weight proportional to the inverse of the number of elements in the class.

We conducted three different training sessions for each modality using exclusively the NDC-ME subset: a training with weights randomly initialised (referred to as *from scratch* models), an other one where we trained all layers from the pre-trained lipreading model (*fully fine-tuned* models) and the last one where we trained only the MS-TCN related layers of the pre-trained model (*last layers fine-tuned* models). Each training session consisted of 80 epochs with an evaluation of the model at the end of each epoch using the validation partition. Since we had limited amount of data, we used an initial learning rate of 3×10^{-6} and a batch size of 16. The learning rate decreased using a cosine annealing schedule. The models were then evaluated with the test partition of our NDC-ME subset on the one hand and on the other hand with all annotated IFADV data to assess the generalisation of the models.

For the fusion, we trained our network of fully connected layers initialised with random weights using several configurations of audio and visual models. In this paper, we present only two configurations: the first one combines both modalities trained from scratch while the second configuration is based on the results obtained for each modality individually. This second fusion model combines the output from the audio model fine-tuned on its MS-TCN layers only and the visual model fine-tuned on all layers. We used an initial learning rate of 3×10^{-6} and a batch size of 16, and the training lasted 30 epochs. The other possible combinations were indeed tested during the experimentation phase, but did not bear any

interesting results (either performed generally less well than the ones presented here or from which we could not draw any relevant conclusion).

V. Results and Discussions

A. Results

Table I contains the Precision, Recall, F1-score and UAR metrics for the S&L classifications. The configurations are separated by modality and evaluation dataset. The highest value per metric and per separation is highlighted in green. We can observe better generalisation to other datasets for the fine-tuned models compared to the models trained from scratch. Fine-tuning is also better for classification except on the NDC-ME evaluation of the fusion model.

Fig. 2a, Fig. 2b and Fig. 2c each contains four heatmaps representing, just like confusion matrices, the way the **Laughs**, **Smiles** and **None** were classified but breaking down the classification results of the laughs and smiles into their different intensity levels. Each column shows the class predicted by the model, while each row is the ground truth intensity levels of each expression. The values are presented in percentages of the sum of the corresponding row/ground truth: for example, X% in (row 1, column 1) represents X% of the sum of row 1.

In this work, we present and discuss only the results of the configurations for which we were able to draw interesting conclusions. But the readers should note that all the configurations not reported for a specific modality but that were reported for another one were indeed carried on during the experimentation phase but left out of this paper on purpose. This is for similar reasons as the fusions configuration in Section IV: the results not reported here, did not bear any interesting conclusion or the performances were not good enough to even be relevant for this work (for example fine-tuning all the layers the audio modality gave very poor classification results, no interpretation of the results could be made, and so was not reported here).

Fig. 2a shows heatmaps related to the audio models trained either from scratch or by fine-tuning the last layers only. The first and second heatmaps from the left are the results of the tests on the NDC-ME data. We can observe that the medium and high levels of laughs have more True Positives than the other cases for both models, and that the fine-tuned model achieves higher results for low level laughs than the model trained from scratch. The third and fourth heatmaps are the results for the same models in the first and second heatmaps (trained on NDC-ME data) but tested on IFADV data. The model trained from scratch tends to classify the samples mostly into the **None** category, while the fine-tuned model shows better performance for all laughter intensity levels.

Heatmaps in Fig. 2b shows the results for the visual models trained either from scratch or by fine-tuning all layers. The first and second heatmaps from the left show that both models have similar results for lower levels (subtle and low) smiles and for **None**. The model from scratch seems to confuse laughter with smiles no matter the intensity, while the fine-tuned model performs better for high level laughs. The third and fourth heatmaps contain the results for the same models mentioned before but applied on the IFADV data. The model trained from scratch shows, as expected since applied on a different dataset, decreased performances but the fine-tuned one seems to perform well on laughs. It seems to improve the performance on smiles compared to the model trained from scratch.

Fig. 2c displays the performance of the fusion of the models trained from scratch for audio and visual modalities (labelled as "From scratch") and the fusion of the fine-tuned audio and visual models mentioned above (labelled as "Fine tuning"). The first and second heatmaps show that the fusion appears to have similar results on NDC-ME data, with better a performance on high level laughter with the models trained from scratch. The third heatmap shows that the low generalisation rate of both modalities trained from scratch induces the same behaviour on their fusion model, while the fourth one presents better True Positives on both laughter and smiles detection.

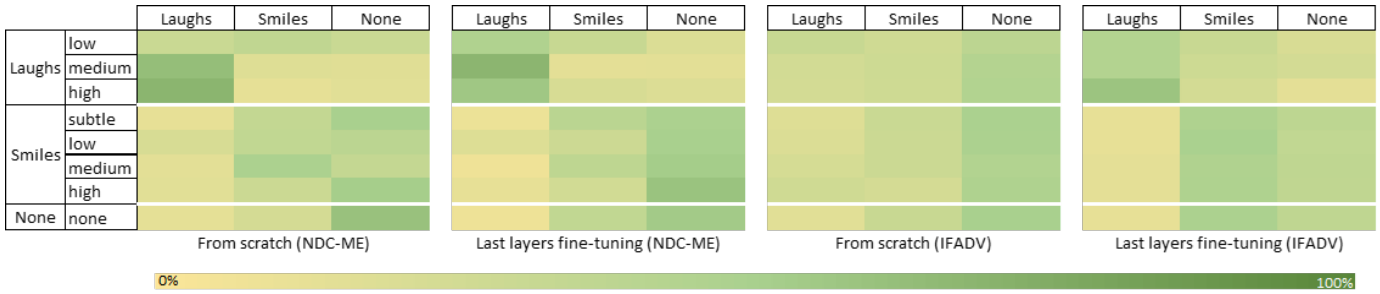
B. Discussion

Firstly it is clear that not one model performed better than all the others in all categories. But by considering overall results, we can argue that, when training and testing on the same dataset, models fusion trained from scratch performs relatively well on all classes, even better than the fine-tuned visual model which, interestingly, seems to confuse low level laughs with smiles. The fusion model seems able to keep the overall good performances of the visual modality while improving the bad ones (low level laughs notably). It is worthy to note though, that in this work, a simple fusion mechanism and training were applied. Improving these should allow to take better advantage of both modalities. We can also note that audio laughs, when misclassified, are most often confused with smiles, especially low-level laughs. Which is an interesting point suggesting that a relationship might exist even in the audio modality. However, this modality does not perform as well on smiles, for either evaluation datasets. It is true that the smiles true positives are quite high but so are the false positives represented by the **None** being misclassified as smiles. On an intuitive level, this makes sense. Indeed, although smiles have been shown to be audibly recognisable, smiled speech is more a change of voice than a burst of affect as is laughter, which makes it more complicated to discriminate from non-smiling speech, especially with the limited amount of data at our disposal. The audio modality seems to perform rather well on laughter, but the smile misclassification leads to poor metrics value. The visual models seem to perform overall better for the smiles than audio modality. This also intuitively makes sense since an obvious discriminating feature between smiles and laughs is the audio cue. Nevertheless the models also seem to perform rather well on laughs especially when fine-tuned, this is probably due to the physical movements accompanying the laughs that are less present when smiling.

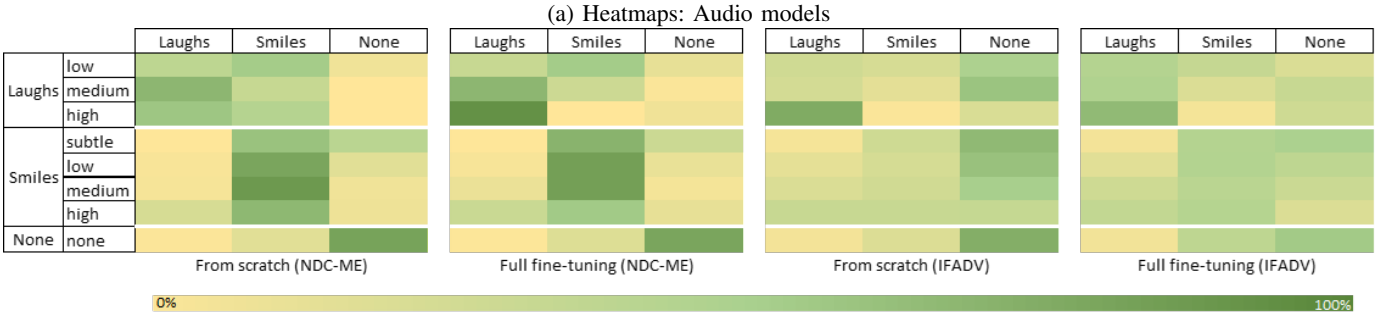
Some interesting notes can also be taken concerning the fusion. First, the fusion surprisingly seems to work better when fused models were trained from scratch, than from fine-tuned ones. This fusion of models trained from scratch seems to

TABLE I: Precision, Recall, F1-score and UAR metrics per configuration for S&L classification. The configuration name follows a XYZZZ pattern with X being the modality: A (Audio), V (Video) or F (Fusion); Y the training method: S (from Scratch) or F (corresponding Finetuning method); ZZZ the evaluation dataset: NDC (NDC-ME) or IFA (IFADV). The highest value for each configuration and dataset is coloured in green.

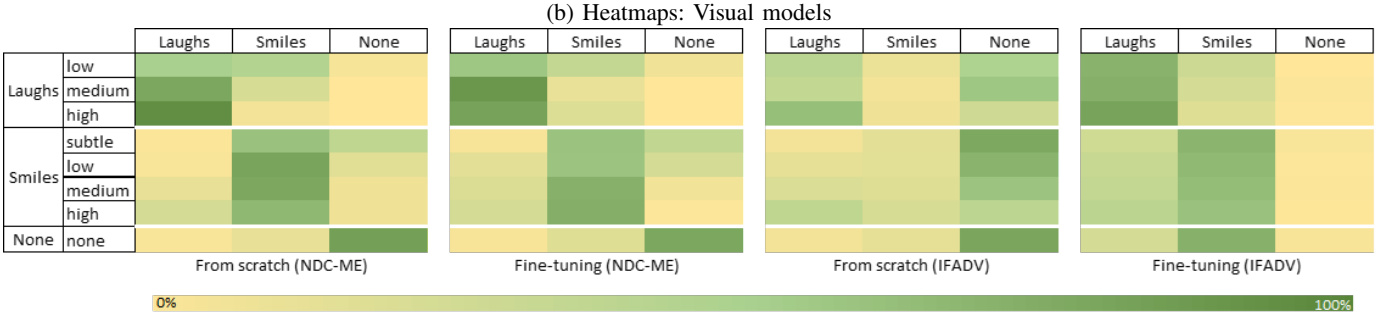
Metrics	**AUDIO**				**VIDEO**				**FUSION**			
	ASNDC	*AFNDC*	*ASIFA*	*AFIFA*	*VSNDC*	*VFNDC*	*VSIFA*	*VFIFA*	*FSNDC*	*FFNDC*	*FSIFA*	*FFIFA*
Precsion	0.4828	0.4937	0.3431	0.3719	0.7116	0.7074	0.3993	0.4137	0.6154	0.6018	0.3837	0.4986
Recall	0.4769	0.4757	0.3689	0.4379	0.6743	0.6793	0.4316	0.4770	0.7829	0.7138	0.3847	0.4639
F1-score	0.4798	0.4845	0.3555	0.4022	0.6924	0.6930	0.4148	0.4431	0.6892	0.6530	0.3842	0.4807
UAR	0.6081	0.6058	0.5206	0.5578	0.7571	0.7611	0.5731	0.5929	0.8077	0.7364	0.5422	0.6046



(a) Heatmaps: Audio models



(b) Heatmaps: Visual models



(c) Heatmaps: Fusion models

Fig. 2: Class distribution heatmaps. Each column corresponds to the predicted class while each row shows the ground-truth label and its intensity. The colour gradient expresses the distribution in percentage per row (the sum of each row should be 100%).

allow the model to use the best prediction of both modalities in one system by improving the recall at the cost of a decrease in precision. Another interesting point to note regarding fine-tuning in general is that it improves laughter classification and generalisation (when applied to IFADV data) in all cases. It also seems to improve smiles true positives score but at the expense in some cases (audio and fusion - fourth heatmap from the left) of the smiles false positives, represented by the confusion of **None** with smiles. For the visual modality, fine-tuning seems to improve the performance of the models both for smiles and laughs detection and its generalizability most of the time which is observed on the results of the models on the IFADV data. The only slight deterioration that we can observe is that more **None** samples are confused for smiles than in the model trained from scratch. We can deduce that fine-tuning allows a model to use the knowledge gained from prior training on speech or lip-reading data to increase its robustness to other datasets.

With the goal to get a better understanding of the models' data representation especially on the impact of fine-tuning, we present a visualisation of the ending layers of the models. For this, we extract embeddings from the output of the MS-TCN block's last layer. We then apply a t-SNE [32] method to reduce the embeddings dimensions to a two-dimensional space while retaining the most relevant features. The general process is depicted in Fig. 3 while the results on the audio modality are shown in Fig. 4 and the ones on the visual modality in Fig. 5. For both modalities, we can see that fine-tuning allows to discriminate better the three classes. Indeed the audio laughs (shades of orange on the figure) are pushed at the extremities of the pattern, while the smiles are still rather mixed with the **None** class, which is coherent with the results presented above. An even more interesting observation can be made on the visual data: we can clearly see the laughs being pushed at the left of the pattern, the low level laughs (yellow dots) tend to also be present in the centre of the pattern, the higher level smiles (darker blue) tend to be more mixed with the laughs and lower level smiles (lighter blue) with the **None** - all coherent with our observations made above.

An analysis of the results with respect to intensity levels shows that the system tends to learn implicit knowledge of those levels from the data. Apart from the high level laughs, the levels on the extremes seem to be more often confused by the models than the medium ones. Low level laughter are in general mostly confused as smiles and the high levels of smiles (medium and high) are mostly confused as laughs while the low levels (subtle and low) are mostly confused as being **None** (which, as we could observe in our dataset, contains a majority of neutral expressions or speech). These observations can be seen in almost all the presented results from the visual modality. This confusion by models are intuitive to us. Indeed, although they were not given any information about the intensity levels of the expressions during training, the models seem to have more difficulty with some intensities on the extreme levels than with others. In the visual modality, if we revised the current results by considering the samples

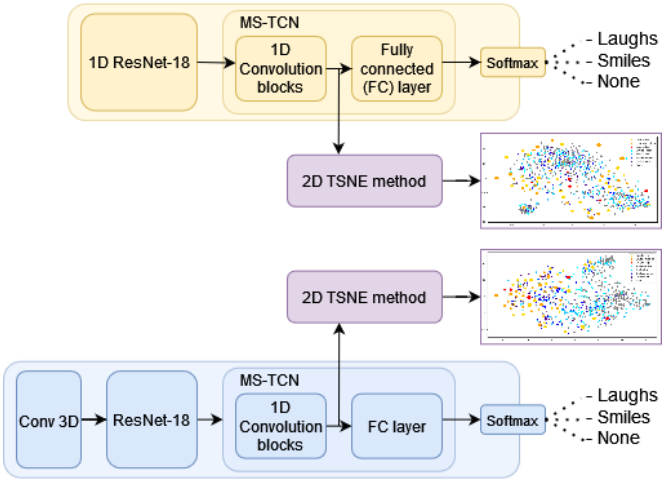


Fig. 3: t-SNE dimensional reduction applied to each modality. Graphic results are display in Fig. 4 and 5.

classified as higher levels of smiles (medium and high) as laughs and lower level smiles (subtle and low) as **None** (thus having only 2 classes at the end instead of 3), the data would be correctly classified as laughs at an average 69.15% with an standard deviation of 5.58% compared to the current average rate 66.46% with an standard deviation of 10.06%. We assume that this is due to the nature of the expressions themselves, since the features representative of some intensities in one expression can be shared with features in another expression (high level smiles and laughs can both show pulled lip corners and raised cheeks for instance).

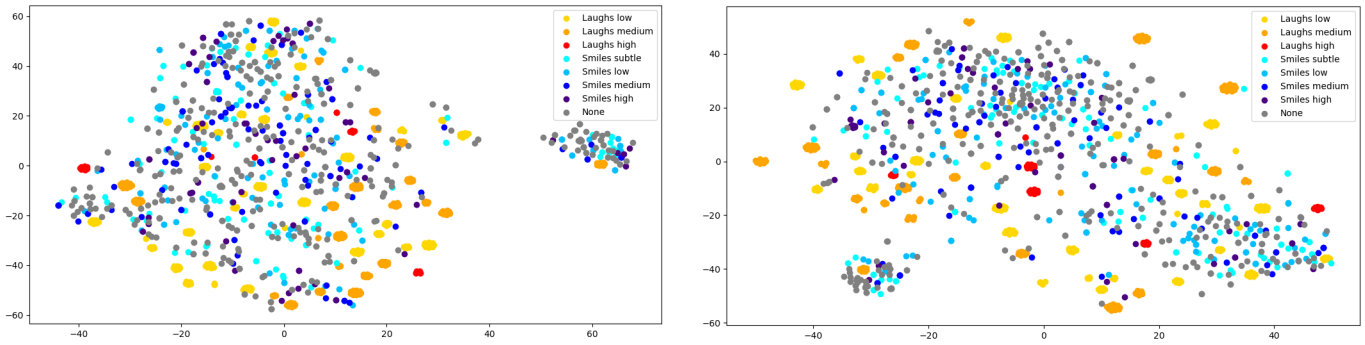
We can therefore mainly conclude that:

1) Fine-tuning is beneficial for performance and generalization in most cases and should be considered instead of training a model from scratch.
2) Given all the observations and analyses made regarding the intensity levels, we can safely conclude that the relationship between smiles and laughs is not as simple as a binary or single class relationship. A more complex relationship should therefore be considered when dealing with these expressions,

Finally, the readers should note that we suspect some aspects of the dataset to have probably influenced negatively some of the results. A first aspect is that some files contain speech coming from the interlocutor of the concerned subject, overlapping the subjects laughter. More accurate detection could be achieved by removing those artefacts from the dataset. A second drawback is that some of the annotations contain subjectivity due to the limited number of annotators and this can make the annotations more sensitive to human error.

VI. Conclusion and Future Works

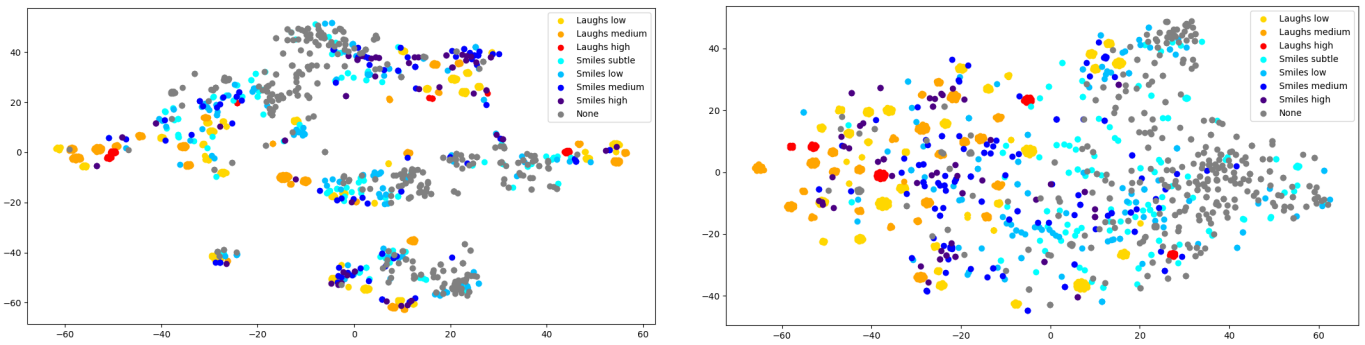
In this work, we presented a study on deep-learning based S&L classifiers discriminating laughs and smiles as two separate entities. We investigated models applied to word



(a) Model trained from scratch.

(b) *Last-layers* fine-tuned model.

Fig. 4: Audio models outputs using a 2D t-SNE representation.



(a) Model trained from scratch.

(b) *Fully* fine-tuned model.

Fig. 5: Visual models outputs using a 2D t-SNE representation.

recognition and lip reading applications for audio and visual-based S&L classification and we fused both modalities with a simple network of fully connected layers. We showed that the fusion system achieves better overall performance than single modalities. We also highlighted the influence that fine-tuning has on the generalization of all models to different datasets. We analysed the behavior of the models with respect to the S&L intensity levels and concluded that these levels should be considered in S&L detection as they have an impact on the models trained in all modalities considered here, even though no prior knowledge was given about them to the model during training.

We intend to investigate in future works other fusion approaches and their effect on classification, trying to improve the fusion results and focus on improving the S&L detection system's efficiency. The observations regarding the behavior of the models with respect to the intensity levels suggested that more complicated relationships between smiles and laughs should be considered rather than a simple binary one or grouping them in a single class. We will therefore also study different grouping methods in different contexts to optimize the use of the datasets. Finally we will integrate intensity level knowledge in the training process and analyse their effect on the knowledge acquired by the models per modality and the impact on the performances.

Ethical Impact Statement

Several aspects should be considered when using data representing emotional states or that could be used as biometric information. In this section we attempt to draw the attention of the readers to a non-exhaustive set of aspects to take into account.

S&L are socio-culturally dependant expressions. Therefore, using them in applications without taking the context into account (user gender, socio-cultural background, interaction topic, etc.) could have unintended impacts, for example by changing the message intended or by extracting false information in a user profiling app.

The resources available in S&L related work in general are limited due to the difficulty of collecting and annotating these expressions and their context. This makes it difficult to generalise the results obtained. Even in this work, although efforts were put in this to balance our datasets for the purposes of our experiments, improvements are still needed and even though the results suggested here provide good directions for future work, a particular attention should be put on the diversity of the subjects in the data and not only its quantity.

S&L have been considered in previous work for biometric identification purposes [33], [34]. It is therefore important that the field of S&L detection in general should evolve in a direction preserving the users privacy.

REFERENCES

[1] P. Glenn, *Laughter in interaction*. Cambridge University Press, 2003, vol. 18.

[2] E. Holt, "The last laugh: Shared laughter and topic termination," *Journal of Pragmatics*, vol. 42, no. 6, pp. 1513–1525, 2010.

[3] R. R. Provine, "Laughter punctuates speech: Linguistic, social and gender contexts of laughter," *Ethology*, vol. 95, no. 4, pp. 291–298, 1993.

[4] D. A. Hayward, E. J. Pereira, A. R. Otto, and J. Ristic, "Smile! social reward drives attention." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 44, no. 2, p. 206, 2018.

[5] S. K. Scott, N. Lavan, S. Chen, and C. McGettigan, "The social life of laughter," *Trends in cognitive sciences*, vol. 18, no. 12, pp. 618–620, 2014.

[6] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting corpus," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 1. IEEE, 2003, pp. I–I.

[7] E. Gilmartin, F. Bonin, N. Campbell, and C. Vogel, "Exploring the role of laughter in multiparty conversation," *Proceedings of the SemDial*, pp. 191–193, 2013.

[8] K. Laskowski and S. Burger, "Analysis of the occurrence of laughter in meetings." in *INTERSPEECH*, 2007, pp. 1258–1261.

[9] N. Chovil, "Discourse-oriented facial displays in conversation," *Research on Language & Social Interaction*, vol. 25, no. 1-4, pp. 163–194, 1991.

[10] X. Guo, L. Polania, and K. Barner, "Smile detection in the wild based on transfer learning," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 679–686.

[11] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Toward practical smile detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2106–2111, 2009.

[12] D. Cui, G.-B. Huang, and T. Liu, "Elm based smile detection using distance vector," *Pattern Recognition*, vol. 79, pp. 356–369, 2018.

[13] V. C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper registers," *The Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101–2107, 1994.

[14] P. Arias, P. Pascal Belin, and J.-J. Aucouturier, "Hearing smiles and smiling back," *Proceedings of Laughter Workshop 2018*, p. 12, 2018.

[15] P. Arias, P. Belin, and J.-J. Aucouturier, "Auditory smiles trigger unconscious facial imitation," *Current Biology*, vol. 28, no. 14, pp. R782–R783, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982218307528

[16] R. B. Kantharaju, F. Ringeval, and L. Besacier, "Automatic recognition of affective laughter in spontaneous dyadic interactions from audiovisual signals," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 220–228.

[17] R. Niewiadomski, M. Mancini, G. Varni, G. Volpe, and A. Camurri, "Automated laughter detection from full-body movements," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 113–123, 2016.

[18] B. Berker Turker, Y. Yemez, T. M. Sezgin, and E. Erzin, "Audio-facial laughter detection in naturalistic dyadic conversations," *IEEE Transactions on Affective Computing*, vol. 8, no. 4, pp. 534–545, 2017.

[19] F. Yang, M. A. Sehili, C. Barras, and L. Devillers, "Smile and laughter detection for elderly people-robot interaction," in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, Eds. Cham: Springer International Publishing, 2015, pp. 694–703.

[20] A. Ito, X. Wang, M. Suzuki, and S. Makino, "Smile and laughter recognition using speech processing and face recognition from conversation video," in *2005 International Conference on Cyberworlds (CW'05)*, 2005, pp. 8 pp.–444.

[21] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotions, qualia, and consciousness*. World Scientific, 2001, pp. 426–443.

[22] J. Trouvain, "Phonetic aspects of "speech-laughs"," in *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: L'Harmattan*. Citeseer, 2001, pp. 634–639.

[23] M. Davila-Ross and G. Dezecache, "The complexity and phylogenetic continuity of laughter and smiles in hominids," *Frontiers in Psychology*, vol. 12, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2021.648497

[24] L. Heron, J. Kim, M. Lee, K. El Haddad, S. Dupont, T. Dutoit, and K. Truong, "A dyadic conversation dataset on moral emotions," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 687–691.

[25] R. J. Van Son, D. Binnenpoorte, H. v. d. Heuvel, and L. Pols, "The ifa corpus: a phonemically segmented dutch" open source" speech database," 2001.

[26] K. El Haddad, S. N. Chakravarthula, and J. Kennedy, "Smile and laugh dynamics in naturalistic dyadic interactions: Intensity levels, sequences and roles," in *2019 International Conference on Multimodal Interaction*, ser. ICMI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 259–263. [Online]. Available: https://doi.org/10.1145/3340555.3353764

[27] T. L. A. Max Planck Institute for Psycholinguistics, "Elan (version 6.3) [computer software]," Retrieved from https://archive.mpi.nl/tla/elan, 2022.

[28] K. El Haddad, "Interaction behavior database," May 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3820510

[29] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6319–6323.

[30] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7608–7612.

[31] X. Guo, L. Polania, and K. Barner, "Smile detection in the wild based on transfer learning," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 679–686.

[32] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[33] C. O. Folorunso, O. P. Popoola, and O. S. Asaolu, "Laughter signature, a new approach to gender recognition," *Engineering Reports*, vol. 2, no. 11, p. e12267, 2020.

[34] A. Harshavardhan, T. Archana, M. Sridevi, and H. Bhukya, "Face smile determination using face and smile detection for perceptual user interfaces (puis) for real-time interaction," *Materials Today: Proceedings*, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785320386612