On the evolution of speech representations for affective computing A brief history and critical overview

Sina Alisamir, Fabien Ringeval

On the evolution of speech representations for affective computing

A brief history and critical overview

Sina Alisamir & Fabien Ringeval

RECENT advances in the field of machine learning have shown great potential for the automatic recognition of apparent human emotions. In the era of Internet of Things (IoT) and big-data processing, where voice-based systems are well established, opportunities to leverage cutting-edge technologies to develop personalised and human-centered services are genuinely real, with a growing demand in many areas such as education, health, well-being and entertainment. Automatic emotion recognition from speech, which is a key element for developing personalised and human-centered services, has reached a degree of maturity that makes it of broad commercial interest today. However, there are still major limiting factors that prevent a broad applicability of emotion recognition technology. For example, one open challenge is the poor generalisation capabilities of currently used feature extraction techniques to interpret expressions of affect across different persons, contexts, cultures and languages.

Since speech and emotion involve interdependent cognitive processes, emotion can be observed both in the spoken words and in the acoustic properties of the speech signal, where many other factors such as gender, age, culture and personality come into play. Even though features derived from speech science have permitted to describe and predict some expressions of affect relatively well, these representations do not encompass all the perceptual cues that humans may sense during an emotional experience. With the advancement of machine (deep) learning, computational methods have been proposed for learning representations from raw speech data. Newly introduced deep representations, although not as easily interpretable as most descriptors from speech science, promise to disentangle many existing issues in affective computing research, such as lack of labelled data, robustness to noise, and domain mismatch [1, 2, 3].

In this contribution, we provide a brief history and critical overview of the different speech representations that have been used in automatic emotion recognition over the years (cf. Figure 1), focusing on how and why the new unsupervised representations in particular can provide major unprecedented benefits in affective computing. Thus, in here, we stay mainly on the topic of speech representations but also mention the new trend to integrate

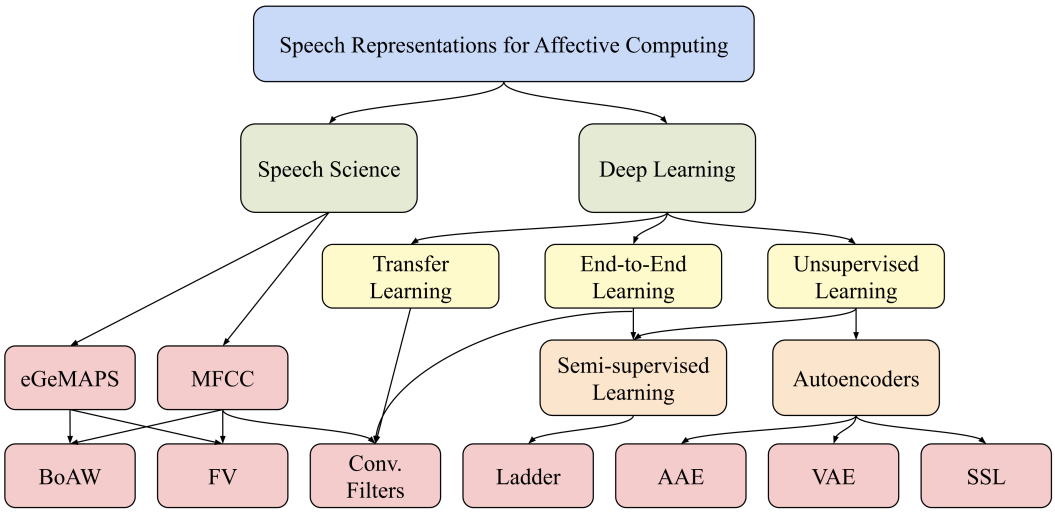

Figure 1: Overview of popular speech representations exploited for emotion recognition as covered in this article. Abbreviations used in the figure are GeMAPS: Geneva Minimalistic Acoustic Parameter Set, MFCC: Mel-Frequency Cepstral Coefficients, BoAW: Bag of Audio Word, FV: Fisher Vector, AAE: Advarsarial AutoEncoder, VAE: Variational AutoEncoder, SSL: Self-Supervised Learning

linguistic information with them. For more information on the models reaching emotion from speech representations and exemplary applications of representation learning for automatic emotion recognition, we refer the reader to the article "Deep Representation Learning for Affective Speech Signal Analysis and Processing" in this special issue [4].

The Usual Suspects: Representations derived from Speech Sciences

Mostly based on established procedures in speech sciences to measure different aspects of phonation, articulation and perception of different patterns of speech, affective computing researchers have been exploiting a large number of – rather hand-crafted – acoustic parameters to identify emotional cues in speech. First automatic emotion recognition systems relied on very-short term energy coefficients describing the frequencies contained in the speech signal according to the human's ear non-linear perception properties. Mel-Frequency Cepstral Coefficients (MFCCs) were later introduced as a deconvolution step between the glottal excitation and the vocal tract in order to preserve the signal variability coming from the source. Even though MFCCs date back to the 80s, they are still the most popular representation used today for emotion, music, and speech recognition. As acoustic correlates of emotion were fairly well documented in the early days of affective computing, researchers explored the acoustic space of speech in a comprehensive way, by combining relevant descriptors identified in speech science with statistical measures, to summarise the temporal trajectories into a vector of fixed-length usable for machine learning.

As many different techniques were proposed to extract speech descriptors such as prosodic (pitch, loudness,

and rhythm), and voice quality, in combination with several sets of statistical measures utilised to summarise their temporal trajectories, results achieved in the first emotion recognition studies were neither easily comparable nor really explainable. As a result, joint efforts have been undertaken to define a reduced set of acoustic descriptors based on expert knowledge, resulting for example in representations such as the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [5]. Even though such reduced set of acoustic descriptors based on expert knowledge has contributed in the standardisation of the feature extraction step, thereby increasing research reproducibility, it generally achieves relatively poor performance in emotion recognition tasks compared to more comprehensive feature sets.

Representations derived from speech science have been used for decades and are arguably the most dominant approach in affective computing so far. However, such descriptors have been engineered with our - rather limited - understanding of the human perception of speech, and therefore do not surely encompass all information that are perceived from the voice, especially in the context of affect.

As a first step in the integration of machine learning techniques for learning representations of speech, a clustering of the acoustic descriptors was proposed instead of their stochastic analysis. Bag of Audio Words (BoAWs) is one such approach where clusters of speech descriptors define a dictionary, which is further used to extract features as the distribution of clusters from the dictionary. Another popular approach is Fisher Vectors (FVs), where Gaussian Mixture Models (GMMs) are employed to estimate gradients of the log-likelihood of the data with respect to the GMM parameters, which are concatenated to form the feature set.

Meanwhile, the advent of deep learning models has paved the way towards the extraction of new representations of speech that can be obtained when solving a given machine learning task. Those approaches have incredibly changed the face of today's speech processing technology, where representations derived from speech science are gradually – and respectfully – told *Adieu* [6], to welcome deep representations that can be tailored to address issues such as data sparsity, robustness to noise, and domain mismatch, which are today's major limiting factors of affective computing.

Learning Deep Representations of Speech

Three main stages of data transformation are usually involved in emotion recognition from speech: (i) extraction of Low-Level Descriptors (LLDs) carrying relevant cues of emotion from the input signal, (ii) quantification and contextualisation of temporal patterns in the LLDs, and (iii) mapping of those patterns to high-level representations

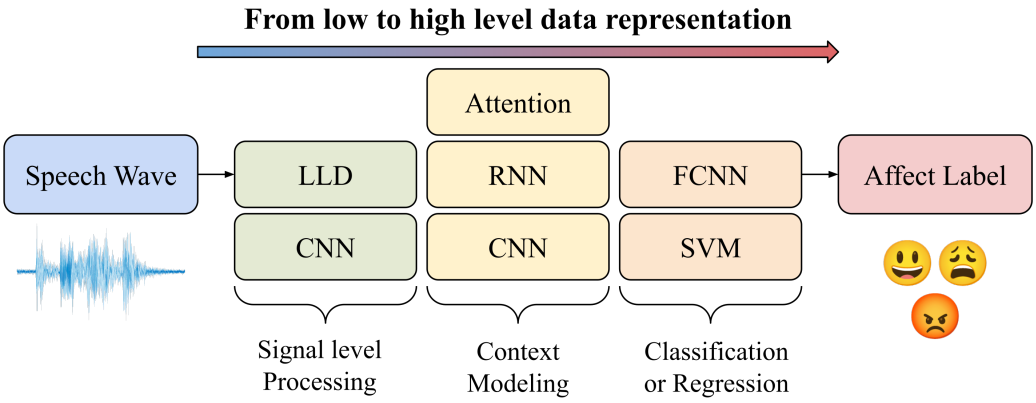

Figure 2: Flowchart of the different processing stages used in emotion recognition from speech; LLD: low-level descriptors derived from speech science; CNN: Convolutional Neural Networks; RNN: Recurrent Neural Networks; Attention: Models that use attention mechanism such as Transformers; FCNN: Fully Connected Neural Network; SVM: Support Vector Machine.

as contained in the labels, cf. Figure 2. Whereas first emotion recognition systems performed these processing stages separately, Deep Neural Networks (DNNs) have made it possible to address all of them jointly, using specific architectures where the level of abstraction of the representations extracted at different stages of processing increases progressively, from the raw waveform to the emotion labels.

DNNs, which are inspired by the hierarchical structure of the brain, are composed of hierarchical layers that perform non-linear transformations of a sequence of low-level representations of the input signal (e.g., MFCCs or the raw speech waveform), to predict a sequence of high-level representations such as emotion labels or dimensions. The last layer of this hierarchical structure would thus be generally less sensitive to most local variations of the input signal, while being more representative of abstract patterns that can be detected from speech and exploited for emotion recognition [7].

Learning Representations from End-to-End

One such paradigm, which learns speech representations at different levels of abstraction, is called end-to-end learning. It is mainly realised by convolutional neural networks (CNNs), which are composed of filters extracting relevant representations from the input signal using the convolution operation. Unlike fully connected neural networks (FCNNs), where each node in each layer is fully connected to all nodes in the next layer, CNNs reduce the number of parameters learned in the network by using only local connections between layers. They also preserve the temporal structure of the speech signal by sharing the parameters across the temporal dimension. When used as front-end, CNNs can thus be viewed as a data-driven feature extractor, where representations are obtained by convolutional

filter banks, and the optimal configuration of the filters is learnt through data instead of being pre-defined from human's auditory properties as with MFCCs. First experimental evaluations on exploiting the raw speech signal to recognise emotion from speech in an end-to-end manner have shown the superiority of this approach compared to hand-crafted representations [6].

However, the caveat of end-to-end learning is that it requires a large amount of labelled data. Indeed, the extraction of relevant descriptors being trained jointly with the prediction of labels, more parameters need to be optimised compared to the use of descriptors derived from speech science as inputs. This is problematic because collecting and annotating emotion data is expensive and we do not have access to large amounts of labelled data in emotion research. In addition, emotion labels are defined by a very small population of annotators who transfer their subjective perception of emotion into either discrete or continuous labels. Learning representations of speech for emotion recognition in an end-to-end manner therefore implies that the set of representations learnt from speech depends on the level of subjectivity present in the annotations of emotion. This complicates the development of emotion models that need to generalise well across people with different age, gender, personality, and culture, as each emotion data set involves specific populations of annotators, with the possible use of different psychological paradigms for describing emotion.

The Grass is Greener on the Other Side: Deep Spectrum

End-to-end systems are therefore limited by the amount of labelled data available to solve the emotion recognition task, which is problematic because emotion annotation is expensive and available data sets are scarce. However, as speech can be represented as an image through its time-frequency representation, one could take benefits of models trained to solve computer vision tasks such as object detection in images, and for which the number of labelled corpora is largely superior to the ones available for emotion recognition. Although describing a speech representation as the probability of identifying a wide variety of objects or animals may seem counter-intuitive, the presence of regular or irregular patterns in a spectrogram is exactly what phoneticians look for to characterize a given speech signal. Such representations are referred to as Deep Spectrum and have proven to match or outperform conventional representations derived from speech science for emotion recognition [8].

However, a major drawback of this approach is the difficulty in predicting the extent to which knowledge gained in solving computer vision tasks can be transferred to emotion recognition from speech using spectrograms. The accumulation of knowledge from various tasks into one can also be considered, however it poses its own challenges

[9]. Furthermore, representations of speech that are transferred from other domains might ignore some patterns that are specific to emotion.

Into the Unknown: Unsupervised Learning

The lack of emotion labels for a diverse range of data, when there is an abundance of unlabelled recordings from a plethora of individuals of different ages and cultures, and directly accessible at (almost) zero cost, has motivated the research in affective computing towards the definition of more agnostic approaches for extracting representations of speech, where human expert knowledge is exploited to define machine learning tasks for extracting abstract representations of speech.

Learning to Copy: Autoencoders

An autoencoder is a neural network that trains an encoder and a decoder model in a tandem in order to reconstruct the input signal from an intermediate representation. Assuming that the representation between the encoder and the decoder model is smaller in size than the input, the encoder model learns a mapping from the raw data to the intermediate representation, which can then be used to reproduce the original signal. Therefore, the encoder model, once trained, can be used as a feature extraction module, as the intermediate representation encompasses some generic information relevant to the reproduction of the original signal. Reducing the dimension of the data representation also implies that the encoder model discards information common to the training data. As a result, training a DNN model based on this representation generally achieves better performance in different domains for a reduced training time.

Recurrent neural networks (RNN), which are particularly useful for modeling the context of a speech signal, have also been investigated in an autoencoder architecture for emotion recognition [10]. In this approach, a latent representation of emotional speech is generated by an RNN-based autoencoder trained on a large amount of unlabelled data, thus improving emotion recognition. The success of RNNs for speech modeling is mainly due to the fact that these models take into account the order of the data, which is useful for speech since the meaning of a speech signal comes not only from the phonemes but also from the way they are ordered in time. The use of convolutional layers first to model low-level features, and then recurrent layers to model context has also proven to be a more generalisable approach for emotion recognition especially in cross-corpora settings [11]. These studies show the benefits of bypassing the lack of labelled data through autoencoders in the context of emotion recognition.

The best of both worlds: Semi-Supervised Learning

Although traditional autoencoders can learn generic high-level features using only unlabelled data, in many cases these features do not perform as well as a model trained in an end-to-end manner [1]. On the other hand, we generally do not have access to a sufficient number of labelled samples to obtain more affect-related features through end-to-end learning. Thus, many researchers use semi-supervised learning, which utilises both the ability of unsupervised learning in terms of learning representations from unlabelled data and also using labelled data to find more important features for the task at hand.

Semi-supervised learning can be done through separate steps in which a generic pattern is first reached in an unsupervised manner and then more important features are extracted towards a specific task by using labels. For example, CNNs can be used to first learn unsupervised representations and then the same layers can be used to reach more affect salient features through the use of labels [12]. To remove the unsupervised pre-training step, Ladder networks, which are a kind of denoising autoencoders, were extended in a way to minimise the supervised and unsupervised cost functions at the same time [13]. This method was later used to learn strong emotional representations for dimensional emotion prediction showing a better generalisation for the emotion prediction model [14]. In a similar approach, the idea of semi-supervised autoencoders, which can achieve state of the art performance for automatic emotion recognition using only a small number of labelled data, is introduced [15]. This method also relies on a joint loss function that minimises both the reconstruction error (similar to autoencoders), and the classification error (similar to supervised learning).

Encoding Meaning into Autoencoders

Traditional autoencoders focus primarily on reconstructing the data by removing common information and not detecting patterns that can explain it. While this has the advantage of reducing the dimensionality of the data, finding generic features, and thus making it easier to reach more complex targets such as emotion, it does not necessarily result in a higher-level representation that embodies the meaning of the data.

One popular way to encode the meaning of speech signals into a latent representation space reached by autoencoders is to enforce similarity between different samples in that space. It can be achieved through a regularisation term in the loss function, which assumes a probabilistic distribution of the latent space, e.g. normal distribution. This implies that small differences among samples in the latent space would be the variations observed between

similar data points and thus this approach is called Variational AutoEncoders (VAEs). It has also been shown that features reached through this method are able to learn latent representations of emotion from speech and reach state of the art results [16]. They show that VAEs can learn powerful features that, when combined with popular RNN models, yield better results than other mentioned techniques so far for both categorical and dimensional emotion recognition.

Another recently adopted approach to achieve better results is to generate additional samples for training the emotion recognition model using generative adversarial networks (GANs). GANs use a generative model to generate samples similar to the actual data, while a discriminative model in an adversarial process attempts to detect whether or not the generated samples have the same distribution as the data at hand. Thus, the generative model ideally contains patterns explaining the distribution of data. As GANs can generate samples that contain higher discriminative power compared to the original data, they achieve better performance than standard augmentation techniques for recognizing underrepresented emotions [17].

Apart from generating samples, the adversary process of GANs has gained much attention in recent years especially since it seems to be the element that enables us to encode patterns found in the data into the generative model. This pattern encoding part is what traditional autoencoders were lacking and thus by combining the two ideas, Advarsarial AutoEncoders (AAEs) were born. In [18], they show that not only can synthetic sample generation from AAEs improve emotion classification results, but also that the representations obtained using AAEs retain their discriminatory power across different emotion categories, meaning that AAEs are able to capture the underlying patterns related to different emotion expressions.

The Rise of Self-Supervised Learning

Another unsupervised learning technique that has recently become wildly popular is Self-Supervised Learning (SSL). This approach comes from methods used to build language models in the field of Natural Language Processing (NLP). These models used for text processing have shown great performance for downstream NLP tasks. Since both text and speech are first mapped into vectors in their first stage processing, the idea of applying language models to process speech has recently become an area of interest.

In SSL, the model learns a general purpose representation of data through training for a pre-defined task using only data. Most common used task in the literature is Contrastive Predictive Coding (CPC), which tries to distinguish the masked frame from another frame, which is usually randomly chosen from a proposed distribution. This way the

model can maximise mutual information over longer context instead of local ones [19]. The choice of CPC makes the model solve a classification task instead of a regression one and thus usage of cross-entropy based losses are usually considered. There are also approaches that solve a regression task, which seems to be more in-line with speech. One of them is Autoregressive Predictive Coding (APC), which tries to minimise an L1 loss for the prediction of the masked frame. It has also been argued that APC can gain better performance than CPC because it only gathers information sufficient for predicting the next frame rather than finding discriminatory factors between the next frame and another randomly chosen one [20].

Although, roughly speaking, SSL is similar to VAEs and AAEs in terms of the objective, which is to reach a representation from data that also considers the patterns seen on the data, there are differences between them. For example, in AAEs a separate generative process is used to model the distribution of data, whereas in SSL data distribution is learnt throughout the same model by an auxiliary task. Also, in VAEs a regularisation term has to be added to the loss function used for training the model, while in SSL, the loss function usually has only one term, which would not give rise to more complications during training the model such as optimising coefficients for different terms involved in the loss function. However, regarding the performance of each technique, all of these novel unsupervised approaches seem to be capable of reaching good representations in recent papers and a comparison between them highly depends on specific techniques and models used as well as the methods used for training them. Moreover, as far as comparing the results goes, given that these techniques are new, a comprehensive study on the performance differences between them especially in the context of affective computing from speech is yet to be done.

Although being novel, SSL has already been applied to many speech related tasks achieving state of the art results. For example, recently introduced Wav2Vec model [21], outperforms the best semi-supervised method for the task of speech recognition while using 100 times less labelled data for fine-tuning the representation learning model using transcribed speech. Although SSL approaches are new, they have also been used for emotion recognition. For example in [22], Problem-Agnostic Speech Encoder (PASE) is introduced, which tries to reach different low level features from raw wave form including MFCC, fundamental frequency, zero-crossing rate and energy (to account for prosody and emotional speech). They also compare their work to classical features such as mel-scaled filter bank, and show that their features can better classify emotions by a simple MultiLayer Perceptron (MLP) classifier. Later in [2], a contrastive representation learning approach has shown to reach better accuracies using a simple FCNN classifier compared to representations derived from speech science and PASE. They also show better performance across different data sets using different languages and show that performance can even be further improved by

fine-tuning the model on a small subset of targeted data. SSL techniques using APC has also been investigated in [23], reaching state of the art results for the task of emotion recognition. They show that by pre-training their model with an SSL approach, and then fine-tuning their model for the specific task, they can obtain a more general model that can be more easily transferred and used on other data sets, making it more practical for industrial applications.

Exploiting Linguistic Information

So far the focus was mainly on acoustic representations from speech. However, emotion is not only perceived by how something is said, but also by what was said. We can easily understand and differentiate words by simply hearing a speech and this plays a role in our perception of another person's emotion. However, for a machine, the verbal message can be understood by text much more easily because speech contains many other kinds of information such as ambient noise, different speakers and microphones. On the other hand, acoustic features contain helpful information for affective computing like prosody that can not always be found in text. Given that affect related behaviours can be found in both verbal and non-verbal communication, by using the two modalities, one expects to reach a representation from which emotion can be more easily recognised.

Textual Information

To understand how to incorporate linguistic information with acoustic ones, we first need to understand how each one of these modalities are processed. For processing text, we first need to tokenise it into recognisable units by a machine. Then an algorithm is used to reach embeddings (vectorised textual representations) from the tokens. For example, to reach word embeddings, the text is divided into words as sub-units of a sentence so that each word corresponds to a feature vector. Thus, for textual data, a token is mapped into and represented only by one unique deterministic embedding, which allows for having a finite set of targets. However, an embedding vector related to the speech wave of a spoken word is not unique because it is affected by different factors such as different speakers, microphones and environments.

This difference makes treating the two forms of signal different and thus the same exact method applied for text does not apply to speech signals. For example, training a self-supervised learning method on text can be done through a classification task which is usually achieved by minimising the negative log-likelihood over a sequence of tokens. On the other hand, self-supervised learning for speech is usually achieved either through a binary classification

task considering a contrastive loss or through a regression task by calculating some form of distance between an actual signal and its predicted version [20]. Thus, given that there are different methods of training required for each modality, one can not expect to simply fuse the two modalities through a straightforward multi-modal model. Nevertheless, to reach a representation that includes both modalities, researchers either try to align speech with textual embeddings or to reach a joint space from the embeddings of the two.

Joint representation of speech and text

Instead of learning a mapping between speech embeddings and textual ones, one can also reach a joint latent representation from both modalities. For example, AAEs have been used to reach a joint representation to recognise different categories of emotion with state of the art accuracy [3]. The results achieved by this representation also required a much simpler model (linear) versus using only the classical LLDs as features, which required a more complex model (SVM) to reach emotion categories. Fusion of two of the known self-supervised learning methods for both speech and text has also shown that one can use a very simple linear classifier and still reach better results when using the two representations for text and speech than using each of the modalities alone [24]. These studies are consistent with the hypothesis that unsupervised representations contain more high-level information, compared to handcrafted features, from which we can detect complex information such as emotion more easily.

Being Realistic: Performance of Different Representations

Ideal representations for affective computing are a set of comprehensive features that could best encompass the space of all possible latent representations of emotion. Towards reaching the ideal representation, many different methods have been proposed during the last decades. Hence, one may wonder how well these different techniques would actually perform on different affect related tasks. However, we should remember that the representation used is only one part of the equation and other parts consist of - but are not limited to - the model reaching the specific target like emotion from the features, the way models are trained, the data used, and the targeted labels.

Starting from more traditional features, we have looked at all the papers that have reported on the test set of RECOLA data set [25] with the same metric and for the same task of dimensional continuous emotion recognition. This data set has been worked on for years (mainly between 2015 to 2019) by different researchers across the world using different models and features and thus can provide us with a broad overview of representations ensued in the

Challenge	Task	Metric	Reference	Hand-crafted	End-to-end	Transfer Learning	Semi-Supervised	Unsupervised	Fusion with Linguistics
ComParE 2020	Elderly Emotion (Arousal)	UAR	Baseline	49.1		50.4		44.3	44.0
			Participants	54.3					63.7
	Elderly Emotion (Valence)	UAR	Baseline	41.7		40.3		33.8	49.0
			Participants	59.0					57.5
AVEC 2019	Depression Detection with AI	CCC	Baseline	.045		.108			
			Participants				.430		.403
ComParE 2019	Continuous Sleepiness	PC	Baseline	.314				.325	
			Participants	.383	.335				
	Baby Sound	UAR	Baseline	57.7				48.1	
			Participants	59.5				62.4	
AVEC 2018	Cross-cultural Emotion Recognition (Arousal)	CCC	Baseline	.236					
			Participants		.377				
	Cross-cultural Emotion Recognition (Valence)	CCC	Baseline	.217					
			Participants		.389				
	Gold-standard Emotion (Arousal)	CCC	Baseline	.651		.495			
			Participants						
	Gold-standard Emotion (Valence)	CCC	Baseline	.346		.158			
			Participants						
ComParE 2018	Atypical Affect	UAR	Baseline	43.1	28.0			35.6	
			Participants	41.1	47.8				
	Self-Assessed Affect	UAR	Baseline	65.2	46.6			57.3	
			Participants	67.0	48.3		48.9		68.4
	(Infant) Crying	UAR	Baseline	73.2	63.5			71.1	
			Participants	70.1					
ComParE 2017	Cold	UAR	Baseline	70.2	60.0		64.8		
			Participants	72.0	71.2				
ComParE 2016	Deception	UAR	Baseline	68.3					
			Participants	72.1		50.7	72.2		

Figure 3: Results achieved for recent affect related challenges according to different representation learning methods that they have used. UAR: Unweighted Average Recall, PC: Pearson's Correlation, CCC: Concordance Correlation Coefficient. The results are only for tasks that were affect related and did provide results using different audio only or acoustics fused with linguistics representations.

past years. By studying the results achieved from different laboratories on this data set, one can recognise that Mel-filter bank features are overally the features through which the best results were achieved for both dimensions of Arousal and Valence compared to other representations derived from speech science such as MFCC or clustering methods like BoAWs or even more complex features such as convolutional filters either learnt through end-to-end learning or used in a transfer learning paradigm.

To investigate the effectiveness of more recent representation learning techniques, we decided to also look into some of the recent affect related challenges such as Interspeech Computational Paralinguistics ChallengE (ComParE) [26] and audio/visual emotion challenge (AVEC) [8]. Since in these challenges, different representations are used for different tasks and different data, one may be able to get a broader overview of the effectiveness of different representation learning methods mentioned in this article. The detailed results of these challenges with respect to the type of representation used is presented in Figure 3. By looking at the results, one can notice all different forms of representations can be effective on almost all affect related tasks. Moreover, representations derived from speech science, which has been long used in different speech related tasks are still present today and can achieve comparable performances to more novel representations. This shows that the effect of other variables involved toward reaching

the results, such as different models used on top of the representations, can not be ignored. For example, one common theme seen throughout all of the challenges was fusion of the outputs or decisions reached by different representations independently, which often led to the best results.

Some of the other points that we have noticed is that by looking specifically at Compare 2017 and 2018, end-to-end learning seems to have mainly achieved worse results than utilising hand-crafted features, which can be explained by lack of labelled data available at these challenges. Also regarding the usage of textual information, the baseline results of Compare 2020 Challenge show that for continuous dimensional emotion, by incorporating linguistic information into acoustic ones, we can reach better performance especially for the Valence dimension, which is more susceptible to perception of meaning of the uttered words [26].

The way these representations are reached can also play an important role. For example, in ComParE 2020 challenge, unsupervised learning has the lowest results, however this can easily be due to the fact that these representations were achieved by using a limited amount of data available on the challenge. In fact, in all the recent challenges, unsupervised learning still remains highly unexplored. However, there are recent studies using unsupervised learning in the context of affective computing showing a better accuracy and transferability of knowledge over representations derived from speech science while requiring a more simple model to reach emotion [2, 24]. Even though many studies have been provided on this area to date, we still lack exhaustive comparisons over a wide range of different representations for different affect related tasks using different models and under different context, languages, and cultures. Thus, despite the progress made in recent years, there is still a long way to go to be able to confidently answer the most seemingly basic question, such as what kind of features to use.

Discussion

In Figure 4, we conceptualised our view of the link between the theory and application of emotion recognition from speech. The task of automatically detecting an observable emotional event from a given raw, possibly noisy, speech signal is currently achieved with machine learning techniques. Such techniques are designed to solve tasks that are clearly defined, i.e., wherein labels have a clear and objective definition providing almost no variations across different annotators. However, in the case of an emotion recognition system, these labels are subjective and represent the perception of a few different human annotators of the observed event, based on a psychological affect model. This approach induces a large amount of subjectivity in the labels because of the natural dependency of emotion with

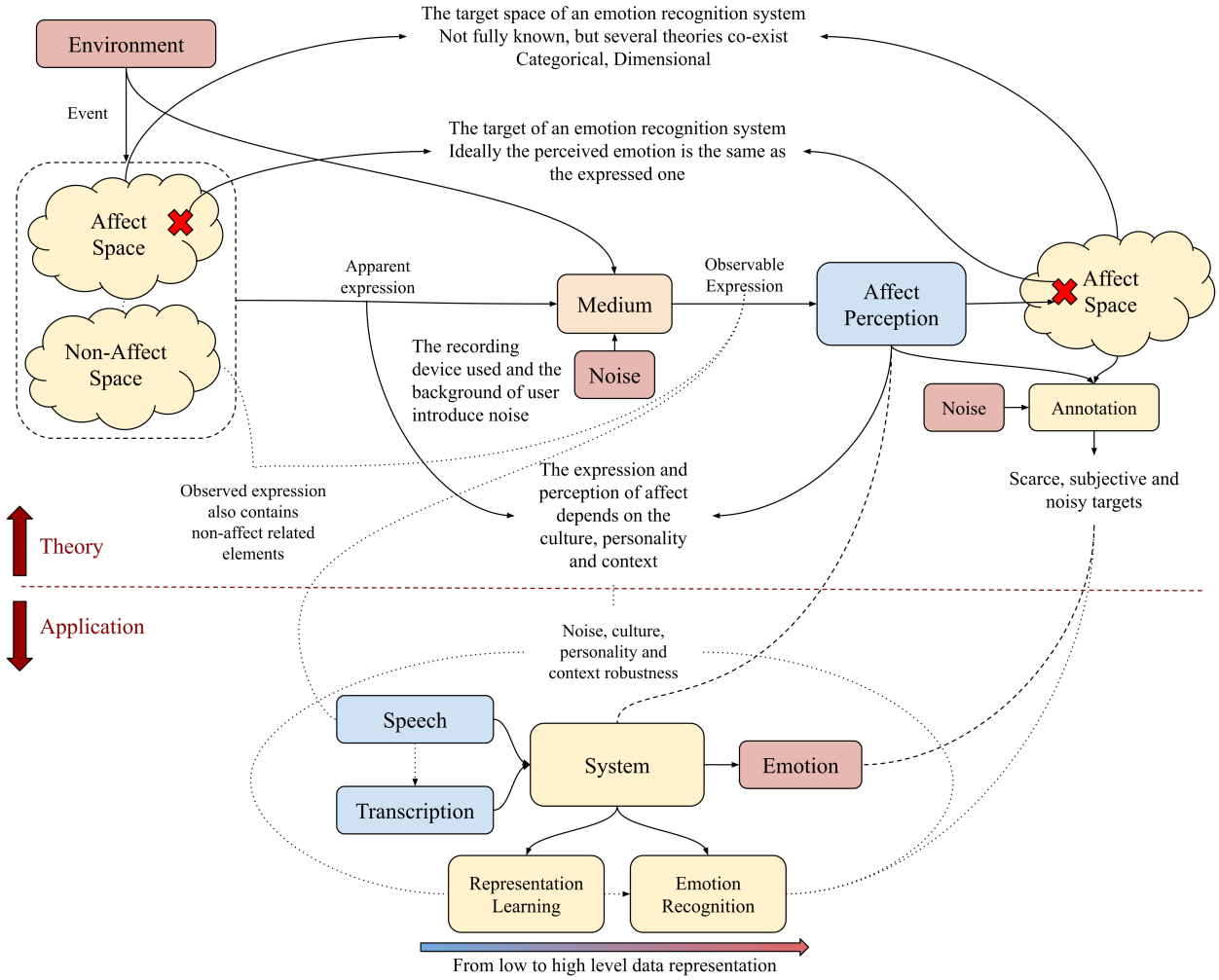

Figure 4: Emotion recognition from speech: overview and challenges.

a myriad of idiosyncratic contextual factors. Moreover, tools used to provide emotion annotations inherently add noise themselves. This means that, we cannot build truly objective emotion recognition models as we could achieve for automatic speech recognition. It also means that comparisons between different emotion recognition models are inherently vague as they can be defined in a myriad of ways. This vagueness in defining emotion targets also makes end-to-end learning methods not an ideal solution for reaching representations of speech relevant for emotion, as the error that is back-propagated in the model is directly quantified from the labels. Using speech representations that are generic and contain high-level information makes the use of subjective and scarce emotion labels less crucial for the emotion recognition task itself, even though we still need emotion labels at the end to guide the learning of the decision stage. Therefore, under this light, the trend toward the use of speech representations derived from an unsupervised machine learning model makes a lot of sense for affective computing, as we can expect the model to capture generic latent representation of speech for different speaker traits and states. This trend can therefore be seen as another approach to holistic speaker analysis, where various high level generic information are directly

captured in the representation of speech, instead of adopting a multi-target learning paradigm [27].

Emotion can be conveyed through both verbal and non-verbal communication. In theory, the verbal message is what can be written and the non-verbal part can only be found in the acoustic. However, there has been studies (mainly for text to speech applications) showing that textual and prosodic features are correlated. In addition, it has been recently shown that a text based SSL representation can outperform other methods for the detection of prosodic prominence from text with just ten percent of the training data [28].Those results suggest that it is not only the help of the verbal communication from textual representations that contains useful information to reach emotion but text can also contain information related to non-verbal communication, albeit the result of the natural correlation found between the two forms of verbal and non-verbal communication in a language.

Whereas most of hand-crafted representations of speech can be easily interpreted, we know that such representations have a large variability across speakers, which make emotion recognition models prone to generalisation errors when confronted to unknown speakers. On the other side, deep learning-based representations can be designed to be more robust for affect sensing, but they cannot provide acoustic and linguistic representations that can be directly explained in the light of emotion. As emotion recognition technology has many real-life applications with – either direct or indirect – educative or training purposes, it might be very much desired to not only provide accurate measurements of a given apparent emotion, but also some additional information on the insights of the taken decision. There are hopefully different possibilities to make deep learning representations more explainable for emotion recognition, such as directly identifying CNNs' activation functions that correlate well with specific hand-crafted representations [6], or using local attention that can explain which specific parts of the speech signal is emotionally salient.

Authors

Sina Alisamir (sina.alisamir@gmail.com) received his M.S. degree in digital electronic circuits from Amirkabir University of Technology, Iran. He is currently pursuing his Ph.D. degree at the Laboratoire d'Informatique de Grenoble at the Université Grenoble Alpes, Saint-Martin-d'Hères, 38400, France and Atos, Échirolles, 38130, France. His research is focused on deep learning algorithms applied to speech signals mainly for prediction of human affective behaviors.

Fabien Ringeval (fabien.ringeval@imag.fr) received his Ph.D. degree for his research on the automatic recognition

of acted and spontaneous emotions from speech from the Université Pierre et Marie Cure, Paris, France. He is currently an associated professor at the Université Grenoble Alpes, Saint-Martin-d'Hères, 38400, France. His research interests include speech processing and artificial intelligence, with applications on the machine sensing of human behaviors (e.g., emotions, health) from multimodal conversational data. He is an associate editor of *IEEE Transactions on Affective Computing* and serves as an area chair for the *Association for Computing Machinery Multimedia* (emotion and social signals) and as a senior program committee member for the *International Conference on Affective Computing and Intelligent Interaction*.

Conclusions

Automatically sensing emotion from speech waves requires multiple stages of data transformation. After each stage, we expect to reach a set of representation of the input data that is more informative of the emotion experienced by the subject at the time of affect expression. However, many challenges still limit how emotion can be robustly detected from speech using acoustic and linguistic representations that vary significantly across a large amount of inter-dependent factors, such as age, gender, personality, social role, health condition, language, and culture.

Recently, a new machine learning trend has emerged and shown great promises toward solving the aforementioned issues. Research has shown that newly introduced unsupervised techniques, especially representations reached by self-supervised learning methods, can deal with many of the issues in affect related tasks much better than traditional approaches based on hand-crafted representations. This is due to the fact that these techniques can recognise high level patterns from only unlabelled data without any supervision or assumption defined by limited human knowledge. These high level abstractions can later be related to the subjective emotion annotations with simple models. Even though these approaches are still in their infancy, they have already been investigated rather exhaustively in many closely related domains such as speech recognition. It has been shown that by using representations reached by these techniques, we require much less labelled data and a simpler model that can generalise better and be less susceptible to low level changes of the signal and other issues like noise and domain mismatch.

In our view, future works in this domain will continue to use representations derived from speech sciences, as they are still the most comprehensive and widely used features today. As many new unsupervised learning methods have been recently introduced, we expect more investigations of these techniques to be conducted in the coming years. Since SSL methods show very promising results so far, they will certainly be the subject of further studies in this area.

Interestingly, predicting unseen frames, which is the basis of SSL methods, has strong roots in neuroscience, and it has been shown that each brain has its own model constantly making predictions of the world at different levels of abstraction [19]. Thus, in view of current trends, the gap between the perception of the world, being emotional or not, by humans and by machines, seems to have already begun to narrow.

References

[1] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W Schuller. "Deep representation learning in speech processing: Challenges, recent advances, and future trends". In: *arXiv preprint arXiv:2001.00378* (2020).

[2] Apoorv Nandan and Jithendra Vepa. "Language Agnostic Speech Embeddings for Emotion Classification". In: *Proc. ICML Workshop SAS* (2020), pp. 1–6.

[3] Eesung Kim, Hyungchan Song, and Jong Won Shin. "Affective Latent Representation of Acoustic and Lexical Features for Emotion Recognition". In: *Sensors* 20.9 (2020), p. 2614.

[4] Chi-Chun Lee, Kusha Sridhar, Jeng-Lin Li, Wei-Cheng Lin, Bo-Hao Su, and Carlos Busso. "Deep Representation Learning for Affective Speech Signal Analysis and Processing". In: *IEEE Signal Processing Magazine* Submitted (02/15/2021) (2021).

[5] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE transactions on affective computing* 7.2 (2015), pp. 190–202.

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

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[8] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019, pp. 3–12.

[9] Lisa Torrey and Jude Shavlik. "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[10] Michael Neumann and Ngoc Thang Vu. "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 7390–7394.

[11] Vipula Dissanayake, Haimo Zhang, Mark Billinghurst, and Suranga Nanayakkara. "Speech Emotion Recognition 'in the wild'Using an Autoencoder". In: *Proc. Interspeech* (2020), pp. 526–530.

[12] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. "Speech emotion recognition using CNN". In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 801–804.

[13] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. "Semi-supervised learning with ladder networks". In: *Advances in neural information processing systems*. 2015, pp. 3546–3554.

[14] Srinivas Parthasarathy and Carlos Busso. "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes". In: *arXiv preprint arXiv:1804.10816* (2018).

[15] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, and Björn Schuller. "Semisupervised autoencoders for speech emotion recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.1 (2017), pp. 31–43.

[16] Siddique Latif, Rajib Rana, Junaid Qadir, and Julien Epps. "Variational autoencoders for learning latent representations of speech emotion: A preliminary study". In: *arXiv preprint arXiv:1712.08708* (2017).

[17] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan. "Data Augmentation Using GANs for Speech Emotion Recognition." In: *INTERSPEECH*. 2019, pp. 171–175.

[18] Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Wael AbdAlmageed, and Carol Espy-Wilson. "Adversarial auto-encoders for speech based emotion recognition". In: *arXiv preprint arXiv:1806.02146* (2018).

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

[20] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. "An unsupervised autoregressive model for speech representation learning". In: *arXiv preprint arXiv:1904.03240* (2019).

[21] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: *arXiv preprint arXiv:2006.11477* (2020).

[22] Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. "Learning problem-agnostic speech representations from multiple self-supervised tasks". In: *arXiv preprint arXiv:1904.03416* (2019).

[23] Ruixiong Zhang, Haiwei Wu, Wubo Li, Dongwei Jiang, Wei Zou, and Xiangang Li. "Transformer based unsupervised pre-training for acoustic representation learning". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 6933–6937.

[24] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. "Jointly Fine-Tuning" BERT-like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition". In: *arXiv preprint arXiv:2008.06682* (2020).

[25] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE. 2013, pp. 1–8.

[26] Björn W Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, et al. "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks". In: *Proceedings INTERSPEECH. Shanghai, China: ISCA* (2020), pp. 2042–2046.

[27] Björn W Schuller, Yue Zhang, and Felix Weninger. "Three recent trends in paralinguistics on the way to omniscient machine intelligence". In: *Journal on Multimodal User Interfaces* 12.4 (2018), pp. 273–283.

[28] Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. "Predicting prosodic prominence from text with pre-trained contextualized word representations". In: *arXiv preprint arXiv:1908.02262* (2019).