DOI:10.1145/3129340

Tracing 20 years of progress in making machines hear our emotions based on speech signal properties.

BY BJÖRN W. SCHULLER

Speech Emotion Recognition

Two Decades in a Nutshell, Benchmarks, and Ongoing Trends

COMMUNICATION WITH COMPUTING machinery has become increasingly 'chatty' these days: Alexa, Cortana, Siri, and many more dialogue systems have hit the consumer market on a broader basis than ever, but do any of them truly notice our emotions and react to them like a human conversational partner would? In fact, the discipline of automatically recognizing human emotion and affective states from speech, usually referred to as *Speech Emotion Recognition* or SER for short, has by now surpassed the "age of majority," celebrating the 22nd anniversary after the seminal work of Daellert et al. in 1996[10]—arguably the first research paper on the topic. However, the idea has existed even longer, as the first patent dates back to the late 1970s.[41]

Previously, a series of studies rooted in psychology rather than in computer science investigated the role of acoustics of human emotion (see, for example, references[8,16,21,34]). Blanton,[4] for example, wrote that "*the effect of emotions upon the voice is recognized by all people. Even the most primitive can recognize the tones of love and fear and anger; and this knowledge is shared by the animals. The dog, the horse, and many other animals can understand the meaning of the human voice. The language of the tones is the oldest and most universal of all our means of communication.*" It appears the time has come for computing machinery to understand it as well.[28] This holds true for the entire field of *affective computing*—Picard's field-coining book by the same name appeared around the same time[29] as SER, describing the broader idea of lending machines emotional intelligence able to recognize human emotion and to synthesize emotion and emotional behavior.

Until now, the broader public has experienced surprisingly little *automatic recognition* of emotion in everyday life. In fact, only few related commercial products have found their way to the market, including the first-ever hardware product—the "Handy Truster"—which appeared around the turn of the millennium and claimed to be able to sense human stress-level and

» key insights

- **Automatic speech recognition helps enrich next-gen AI with emotional intelligence abilities by grasping the emotion from voice and words.**
- **After more than two decades of research, the field has matured to the point where it can be the "next big thing" in speech user interfaces, spoken language processing, and speech analysis for health, retrieval, robotics, security, and a plethora of further applications. This is also shown in the benchmarks of more than a dozen research competitions held in the field to date.**
- **While deep learning started in this field a decade ago, it recently pushed to end-to-end learning from raw speech data—just one of a couple of current breakthroughs.**

ILLUSTRATION BY VAULT49

deception contained in speech. Approximately 10 years later, the first broad-consumer market video game appeared. "Truth or Lies" (THQ) was equipped with a disc and a microphone for players to bring the popular "Spin the Bottle" game to the digital age. Unfortunately, the meta-review service metacritic.com reported only a score of 28 out of 100 based on only six reviews from professional critics. The tech side seemed premature: reviewers complained about "unstable tech" and "faulty software" that failed to achieve what it promised—detect lies from human speech. However, the first success stories can be observed at this time; including the European ASC-Inclusion project[a] that reports encouraging observations in open trials across three countries for a serious video game that teaches autistic children in a playful way how to best show emotions. Interestingly, a recent study shows that voice-only as modality seems best for humans' empathic accuracy as compared to video-only or audiovisual communication.[22]

Here, I aim to provide a snapshot of the state-of-the-art and remaining challenges in this field. Of course, over the years further overviews have been published that the reader may find of interest, such as references[2,6,15,20,38] or on the broader field of affective computing[17,43] where one finds an overview also on further modalities such as facial expression, body posture, or a range of bio-sensors and brain waves for the recognition of human emotion. These surveys cover progress up to 2013, but quite a bit has happened since then. Further, this short survey is the first to provide an overview on all open competitive challenges in this field to date. Finally, it distills a number of future tendencies discussed here for the first time.

The Traditional Approach

Let's start off by looking at the conventional way to build up an engine able to recognize emotion from speech.

Modeling. First things first: approaching the automatic recognition of emotion requires an appropriate emotion representation model. This raises two main questions: How to represent emotion per se, and how to optimally quantify the time axis.[32] Starting with representing emotion in an adequate way to ensure proper fit with the psychology literature while choosing a representation that can well be handled by a machine, two models are usually found in practice. The first model is discrete classes, such as the Ekman "big six" emotion categories, including anger, disgust, fear, happiness, and sadness—often added by a "neutral" rest-class as opposed to a *value* "continuous" dimension approach that appears to be the favored approach today.[17] In this second approach, the two axes *arousal* or activation (known to be well accessible in particular by acoustic features) and *valence* or positivity (known to be well accessible by linguistic features[17]) prevail alongside others such as power or expectation. One can translate between the categories and dimensions such as 'anger'→{*negative_valence*, *high_arousal*} in a coarse quantization. Other aspects of modeling include the temporal resolution[17] and the quality and masking of emotion, such as acted, elicited, naturalistic, pretended, and regulated.

Approaching the automatic recognition of emotion requires an appropriate emotion representation.

Annotation. Once a model is decided upon, the next crucial issue is usually the acquisition of labeled data for training and testing that suits the according emotion representation model.[13] A particularity of the field is the relatively high subjectivity and uncertainty in the target labels. Not surprising, even humans usually disagree to some degree as to what the emotion should be expressed in the speech of others—or any other modality accessible to humans.[13] Self-assessment could be an option, and is often used when no information to annotators is available or easily accessible, such as for physiological data. Suitable tools exist, such as the widely used PANAS, allowing for self-report assessment of positive and negative affect.[39] Yet, self-reported affect can be tricky as well, as no one has exact knowledge or memory of the emotion experienced at a moment in time. Further, observer rating can be a more appropriate label in the case of automatic emotion recognition that today largely targets assessment of the expressed emotion, rather than the felt emotion.

Likewise, external annotation may be more focused on the emotion ob-

a http://www.asc-inclusion.eu

served and being indeed observable. Likewise, usually five or more external raters' annotations—particularly in the case of crowdsourcing—form the basis of the construction of target labels, for example, by majority vote, or average in the case of a value continuous emotion representation.[17] Further, elimination of outliers or weighting of raters by their agreement/disagreement with the majority of raters can be applied, for example, by the *evaluator weighted estimator* (for example, Schuller and Batliner[32]). Such weighting becomes particularly relevant when crowdsourcing the labels, such as in a gamified way, for example, by the iHEARu-PLAY platform.[b] In the case of value continuous label and time representation, for example, for continuous arousal assessment, raters often move a joystick or slider in real time per emotion dimension while listening to the material to rate. This poses a challenge to time align different raters' annotations, as delays and speed variations in reaction time coin the annotation tracks. Such delays can be around four seconds,[37] and time warping enabled alignment algorithms should be preferred. In the case of discretized time, that is, judgment per larger segment of speech, pairwise comparisons leading to a ranking have recently emerged as an interesting alternative, as it may be easier for a rater to compare two or more stimuli rather than find an absolute value assignment for any stimulus.[17]

To avoid needs of annotation, past works often used acting (out an experience) or (targeted) elicitation of emotions. This comes at a disadvantage because the emotion may not be realistic or it may be questionable whether the right data collection protocol was followed such that the assumptions made on which emotion is finally collected would hold. In the present big data era, simply waiting for the emotion sought to become part of the collected data seems more feasible aiming at collection of emotion "from the wild" rather than from the lab.

Audio features. With labeled data at hand, one traditionally needs characteristic audio and textual features before feeding data into a suited machine-learning algorithm. This is an ongoing active subfield of research in the SER domain—the design of ideal features that best reflect the emotional content and should be robust against environmental noises, varying languages, or even cultural influences. Most of the established ones are rather low level, such as energy or spectral information, as these can be robustly determined. Yet, in the *synthesis* of emotion, there is a strong focus on prosodic features, that is, describing the intonation, intensity, and rhythm of the speech next to voice quality features. The automatic *analysis* of emotional speech often adds or even focuses entirely on spectral features, such as formants or selected band-energies, center of gravity, or roll-off points and cepstral features such as MFCC or mel-frequency bands as well as linear prediction coefficients.[2,15,38] Based on frame-by-frame extraction, one usually derives statistics by applying functionals that map a time series of frames with varying length onto a scalar value per segment of choice.[2,6] The length may vary with the *unit of analysis*, such as voiced or unvoiced sound, phoneme, syllable, or word. A second of audio material or shorter can be recommended considering the trade-off of having more information at hand versus higher parameter variability if the length of the analysis window is further increased. A high number of functionals is often used such as moments, extremes, segments, percentiles, or spectral functionals, for example, as offered by the openSMILE toolkit[c] that provides predefined feature sets that often serve as baseline reference in the research competitions in the field. The current trend is to increase the number of features up to some several thousands of brute-forced features that was often in stark contrast to the sparse amount of training material available in this field.[17,38,43]

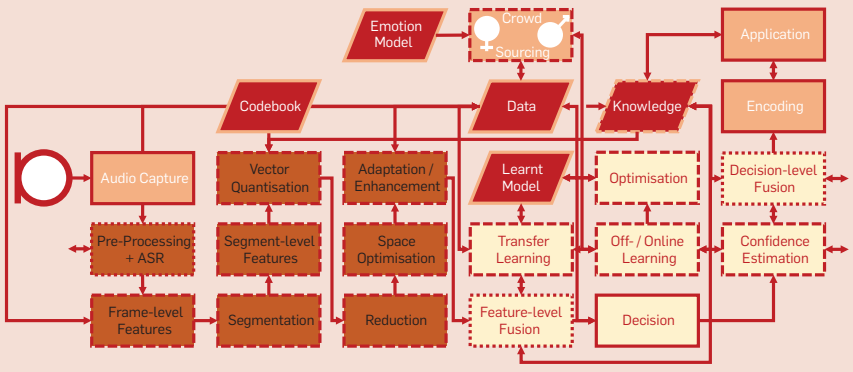
Textual features. Going beyond how something is said, *textual features* as derived from the automatic speech recognition engine's output are mostly looking at individual words or sequences of these such as *n*-grams and their posterior probability to estimate a particular emotion class or value.[23] Alternatively, bag-of-word approaches are highly popular, where each textual entity in the vocabulary of all meaningful entities—from now on referred to as words—seen during vocabulary construction usually forms a textual feature.[18] Then, the frequency of occurrence of the words is used as actual feature value. It is possibly normalized to the number of occurrences in the training material, or to the current string of interest, length of the current string, or represented by logarithm, in binary format, and so on. Linguistically moti-

b https://ihearu-play.eu

c http://audeering.com/technology/opensmile

Figure 1. A current speech emotion recognition engine.

The chain of processing follows from the microphone (left) via the signal processing side of preprocessing and feature extraction (dark orange boxes) via the machine learning blocks (light orange) to encoding of information to feed into an application. Dashed boxes indicate optional steps. Five databases are shown in red. Crowdsourcing serves labeling efforts in the first place. ASR = Automatic Speech Recognition.



vated clustering of word variants may be applied, such as by stemming or representing morphological variants like different tenses. Also, "stopping," or the elimination of entities that do not occur sufficiently or frequently or seem irrelevant from a linguistic or expert's point of view, can be considered. However, in the recent years of increasingly big textual and further data resources to train from, the representation type of the word frequencies, as well as stemming and stopping, seem to have become increasingly irrelevant.[33] Rather, the retagging by word classes, such as part-of-speech tagging, for example, by groups such as noun, verb, or adjective, semantic word groups such as standard linguistic dimensions, psychological processes, personal concerns, and spoken categories as in the LIWC toolkit[d] or even the translation to affect categories or values by linguistic resources such as SenticNet[e] and others, or via relationships in ConceptNet,[f] General Inquirer,[g] WordNet,[h] and alike, can help to add further meaningful representations.

A promising recent trend is to use either soft clustering, that is, not assigning an observed word to a single word in the vocabulary or more general consideration of embedding words such as by word2vec approaches or convolutional neural networks. Alternatively, recurrent neural networks—possibly enhanced by long short-term memory[33]—and other forms of representation of longer contexts seem promising.

It should be noted that the traditional field of *sentiment analysis* is highly related to the recognition of emotion from text, albeit traditionally rather dealing with written and often longer passages of text.[33] This field offers a multiplicity of further approaches. A major difference is given by the uncertainty one has to deal with in spoken language—ideally, by incorporating confidence measures or *n*-best alternative hypotheses from the speech recognizer. Also, spoken language naturally differs from written text by lower emphasis on grammatical correctness, frequent use of word fragments, and so on. In particular, non-verbal vocalizations such as laughter, hesitations, consent, breathing, and sighing frequently occur, and should best be recognized as well, as they are often highly informative as to the emotional content. Once recognized, they can be embedded in a string.

Acoustic and linguistic feature information can be fused directly by concatenation into one single feature vector if both operate on the same time level, or by late fusion, that is, after coming to predictions per feature stream.[23] The latter also allows for representation of different acoustic or linguistic feature types on different time levels. As an example, one can combine bags-of-phonemes per fixed-length chunk of audio with turn-level word histograms in a late(r) fusion manner.

Peeking under the engine's hood. Now, let us look under the hood of an entire emotion recognition engine in Figure 1. There, one can see the features described here are the most characteristic part of a *speech* emotion recognizer —the rest of the processing chain is mainly a conventional pattern recognition system, and will thus not be further explored here. Some blocks in the figure will be mentioned in more detail later. Others, such as the learning part or, which classifier or regressor is popular in the field, will be illustrated by the practical examples from research competitions' results shown below.

En Vogue: The Ongoing Trends

Here, I outline a number of promising avenues that have recently seen increasing interest by the community. Obviously, this selection can only represent a subset, and many others exist.

Holistic speaker modeling. An important aspect of increased robustness is to consider other states and traits that temporarily impact on the voice production. In other words, one is not only emotional, but also potentially tired, having a cold, is alcohol intoxicated, or, sounds differently because being in a certain mood. Likewise, modern emotion recognition engines should see the larger picture of a speaker's states and traits beyond the emotion of interest to best recognize it independent of such co-influencing factors. As training a holistic model is difficult due to the almost entire absence of such richly annotated speech data resources that encompass a wide variety of states and traits, weakly supervised cross-task labeling offers an alternative to relabel databases of emotional speech in a richer way.

Efficient data collection. An ever-present if not main bottle neck since the beginning is the scarcity of speech data labeled by emotion. Not surprising, a major effort has been made over the last years to render data collection and annotation as efficient as possible.[17,38,43]

Weakly supervised learning. *Semi-supervised learning* approaches could prove successful in exploiting additional unlabeled data, once an initial engine was trained.[12,25] The idea is to have the machine itself label new previously unseen data—ideally only if a meaningful confidence measure is exceeded. However, it seems reasonable to keep human labeling in the loop to ensure a sufficient amount of quality labels. *Active learning* can help to reduce such human labeling requirements significantly. The machine preselects only those unlabeled instances for human labeling, which seem of particular interest. Such interest can be determined, for example, based upon whether a sample is likely to be from a class or interval on a continuous dimension that has previously been seen less than others. Further, the expected change in model parameters of the learned model can be the basis—if knowing the label would not change the model, there is no interest in spending human-labeling efforts. An extension can be to decide on how many and which humans to ask about a data point.

As mentioned earlier, emotion is often subjective and ambiguous. One usually must acquire several opinions. However, the machine can gradually learn "whom to trust when" and start with the most reliable labeler, for example, measured by the individual's average agreement with the average labeler population. If the label deviates from what the machine expects, a next opinion can be crowdsourced—ideally from the labeler who in such case would be most reliable. Putting these two ideas—semi-supervised and active learning —together, leads to the par-

d http://liwc.wpengine.com
e http://sentic.net
f http://conceptnet.io
g http://www.wjh.harvard.edu/ inquirer
h http://wordnet.princeton.edu

ticularly efficient *cooperative learning* of machines with human help.[24] In this approach, the machine decides based upon its confidence in its estimate whether it can label the data itself, such as in case of high confidence. If it is not sufficiently confident, it evaluates whether asking a human for aid is worth it. The overall process can be executed iteratively, that is, once newly labeled data either by the machine or a human is obtained, the model can be retrained, which will mostly increase its reliability and confidence. Then, the data that had not been labeled in a previous iteration might now be labeled by the machine or considered as worth labeling by a human. Monitoring improvements on test data is mandatory to avoid decreasing reliability.

If no initial data exists to start the iterative loop of weakly supervised learning, but similar related data is at hand, transfer learning may be an option.[11] To give an example, one may want to recognize the emotion of child speakers, but has only adult emotional speech data at hand. In such case, the features, the trained model, or even the representation, and further aspects can be transferred by learning from the data to the new domain. A broad number of transfer learning and domain adaptation algorithms exists and have been applied in this field, such as in Abdelwahab and Busso.[1] An interesting option of data enrichment can be to include other non-speech audio: as perception of certain emotional aspects such as arousal or valence seem to hold across audio types including music and general sound, one can seemingly train a speech emotion recognizer even on music or sound, as long as it is labeled accordingly.[40]

Obviously, transfer learning can help to make the types of signal more reusable to train emotion recognition engines across these audio types. Even image pretrained deep networks have recently been used to classify emotion in speech based on spectral representations at very impressive performance by the auDeep toolkit.[i] Should collecting and/or labeling of speech data not be an option, also *synthesized speech* can be considered for training of acoustic emotion models—either using synthesis of emotional speech, or simply to enrich the model of neutral speech by using non-emotional synthesized speech.[26] This can be beneficial, as one can generate arbitrary amounts of speech material at little extra cost varying the phonetic content, the speaker characteristics, and alike. Ideally, one could even ad-hoc render a phonetically matched speech sample in different emotions to find the closest match. A similar thought is followed by the recent use of generative adversarial network topologies, where a first neural network learns to synthesize training material, and another to recognize real from synthesized material and the task of interest.[5] Obviously, transfer learning can bridge the gap between artificial and real speech. In future efforts, a closer and immediate coupling between synthesis and analysis of emotional speech could help render this process more efficient.

An important aspect of increased robustness is to consider other states and traits that temporarily impact on voice production.

If no *annotated data* is available, and no emotional speech synthesizer is at hand, *unsupervised learning* could help if the knowledge of the emotion is not needed explicitly and in human-interpretable ways. An example is the integration of information on emotion in a spoken dialogue system: if features that bear information on the emotional content are used during unsupervised clustering of emotionally unlabeled speech material, one may expect the clusters to represent information related to emotion. The dialogue system could then learn—best reinforced—how to use the information on the current cluster in a dialogue situation to decide on its reaction based on observations of human-to-human dialogue. Likewise, at no point would someone know exactly what the clusters represent beyond designing the initial feature set for clustering to reflect, say, emotion; yet, the information could be used. Should no speech data be available, rule-based approaches could be used, which exploit the knowledge existing in the literature. A basis will usually be a speaker normalization. Then, one measures if the speech should, for example, be faster, higher pitched, or louder to assume a joyful state. Yet, given the oversimplification of a high-dimensional non-linear mapping problem, such an approach would, unfortunately, have limits.

i https://github.com/auDeep/auDeep

Data-learned features. As the quest for the optimal features has dominated the field similarly as the ever-lacking large and naturalistic databases, it is not surprising that with increased availability of the latter the first can be targeted in a whole new way, that is, *learn* features from data. This bears the charm that features should be optimally fitted to the data. Further, higher-level features could be learned. On the downside, one may wonder about potentially decreased generalization ability across databases. Below, two currently popular ways of learning feature representations are introduced.

The idea to cluster chunks of audio into words to then be able to treat these just like textual words during further feature extraction, for example, by histogram representation as "bag of audio words" was first used in sound recognition, but has found its way into recognition of emotion in speech.[31] Interestingly, these form some kind of modeling in between acoustic and linguistic representation depending on the low-level features that are used as basis.[31] As an example, one may use wavelet or cepstral coefficients and cluster these to obtain the audio words and the vocabulary built up by all found audio words. An even simpler, yet often similarly effective way is random sampling k vectors as audio words, that is, executing only the initialization of k-means. Then, the actual feature could be frequency of occurrence per audio word in a larger time window such as a second, a turn, or alike, for example, by the openXBOW tool.[j]

Split-vector quantization allows you to group the basis features to derive several histograms, for example, one for prosodic features and one for spectral features. The construction of this vocabulary is the actual data-injection step during feature learning, as speech data will be needed to reasonably build it up. There exists a huge potential of unexploited, more elaborate forms of audio words, such as variable length audio-words by clustering with dynamic time warping, soft-assignments of words during histogram calculation, audio-word embeddings, audio-word retagging or hierarchical clustering, such as the part-of-speech tagging in textual word handling, or speech component audio words by executing non-negative matrix factorization or alike, and creating audio words from components of audio.

The "neuro"-naissance or renaissance of neural networks has not stopped at revolutionizing automatic speech recognition.

The "neuro"-naissance or renaissance of neural networks has not stopped at revolutionizing automatic speech recognition. Since the first publications on deep learning for speech emotion recognition (in Wöllmer et al.,[42] a long-short term memory recurrent neural network (LSTM RNN) is used, and in Stuhlsatz et al.[35] a restricted Boltzman machines-based feed-forward deep net learns features), several authors followed this idea to learn the feature representation with a deep neural network, for example, Cibau[7] and Kim et al.[19] Convolutional neural networks (CNN) were also successfully employed to learn emotional feature representations.[27] The first end-to-end learning system for speech emotion recognition was recently presented by a sequence of two CNN layers operating at different time resolutions: 5ms first, then 500ms followed by a LSTM RNN at highly impressive performance.[37,k] In future topologies, one may consider stacking neural layers with different purposes such as speech denoising, feature extraction, feature enhancement, feature bundling, for example, by use of a bottleneck layer, and classification/regression with memory.[33]

Confidence measures. Given the higher degree of subjectivity of the task and imperfect recognition results, the provision of *confidence measures* of an emotion estimate seems mandatory in any application context.[17] However, the estimation of meaningful independent confidence measures beyond direct measures coming from the machine-learning algorithm, for example, distance to the hyperplane in kernel machines, softmax functions at the output layer of neural networks, or alike, has hardly been researched in SER.[17] Four main directions seem promising: 1) Automatic estimation of human labelers' agreement on unseen data: instead of training the emotion as a target, one can train a classifier on the number of raters that agreed on

j https://github.com/openXBOW/openXBOW

k A recent toolkit is found at https://github.com/end2you/end2you.

the label or, the standard deviation or alike in case of a regression task. Then, by automatically estimating human agreement on *novel* data, one obtains an impression on the difficulty of the current emotion prediction. In other words, one learns to estimate for new data if humans would agree or likely disagree on its emotion. Ideally, this can be targeted as a multitask problem learning the emotion and human agreement in parallel. 2) One can train a second learning algorithm to predict errors of the emotion recognition engine. To this end, one needs to run the trained emotion recognizer versus the development data to then train the confidence estimator on the errors or non-errors of the SER engine observed on that data. In case of a regression task, the linear error or other suited measures can be used as target. 3) Estimating the similarity of the data to the training data can be another option. A possible solution is training a compression autoencoder (a neural network that maps the feature space input onto itself to learn for example a compact representation of the data) on the data the emotion recognition was trained upon. Then, the new data to be handled can be run through the autoencoder. If the deviation between input and output of the autoencoder is high, for example, measured by Euclidean distance, one can assume low confidence in the emotion recognition results as the data is likely to be highly dissimilar. 4) Estimating acoustic degradation or word error rate. On a final note, reliable confidence measures are also the heart-piece of efficient weakly supervised learning.

Coming Clean: The Benchmarks

But how reliable are SER engines? This can partially be answered looking at the research challenges held in the field up to now. While the first official competition event with properly defined train and test sets and labels unknown to the participants—the Interspeech 2009 Emotion Challenge[l]—dates back nine

l http://compare.openaudio.eu.

Benchmark results of the SER challenge events.

Databases = the basis of data used in the competitions. Note that sometimes only subsets have been used. Only challenges are listed that provided audio only results (thus excluding, for example, AVEC 2014 and EmotiW since 2015). Some abbreviations here are obvious, others include lng=language (by country code ISO 3166 ALPHA-2 where "–" indicates an artificial language). hrs/spks/# = hours/speakers/number of data points. Task gives the number of classes or the dimensions =(A)rousal, (V)alence, (P)ower, (E)xpectation. "·2"= a binary classification per dimension. oS = openSMILE (feature extractor with standardized feature sets). EC = Interspeech Emotion Challenge. CRNN = CNN followed by a recurrent neural network with LSTM. RF = Random Forests. SVM/R = Support Vector Machines/Regression. BoAW = Bag-of-Audio-Words. UA = Unweighted Accuracy. WA = Weighted Accuracy. MAP = Macro Average Precision. PCC = Pearson's Correlation Coefficient. CCC = Concordance Correlation Coefficient. Baseline results follow the order under each "task."

Challenge	Database	lng	Quality	hrs/spks/#	task	unit	# feat	model	baseline
EC 09	FAU AEC	DE	lab	9.1/51/18216	2/5	chunk	384 oS	SVM	.677/.382 UA
ComParE 13	GEMEP	–	lab	~.6/10/1260	AV·2/12	turn	6373 oS	SVM	.750/.616/.409 UA
AVEC 11	SEMAINE	UK	lab	3.7/24/50350	AVPE·2	word	1941 oS	SVM	.412/.558/.527/.592 WA
AVEC 12	SEMAINE	UK	lab	3.7/24/50350	AVPE	word	1841 oS	SVR	.014/.040/.016/.038 PCC
AVEC 13	AViD	DE	lab	240/292/864k*	AV	sgmt	2268 oS	SVR	.090/.089 PCC
AVEC 15	RECOLA	FR	VoIP	2.3/27/202527	AV	sgmt	102 oS	SVR	.228/.068 CCC
AVEC 16	RECOLA	FR	VoIP	2.3/27/202527	AV	sgmt	88 oS	SVR	.648/.375 CCC
							–	CRNN	.686/.261 CCC
							BoAW	SVR	.753/.430 CCC
AVEC 17	SEWA	DE	VoIP	3/64/106896	AV	sgmt	BoAW	SVR	.225/.244 CCC
EmotiW 13	AFEW 3.0	US	film	~.8/315/1088	7	clip	1582 oS	SVM	.2244 WA
EmotiW 14	AFEW 4.0	US	film	~1.0/428/1368	7	clip	1582 oS	SVM	.2678 WA
MEC 16	CHEAVD	CN	film/TV	2.3/238/2852	8	clip	88 oS	RF	.2402 MAP/.2436 WA
MEC 17	CHEAVD 2.0	CN	film/TV	7.9/527/7030	8	clip	88 oS	SVM	.392 MAP/.405 WA

years by now, several further followed. In 2011, the first AudioVisual Emotion Challenge (AVEC 2011) took place, which also featured a speech-only track. By now, seven annual AVEC challenges took place[m]—in 2015 physiological signal information was added for the first time. The Interspeech Computational Paralinguistics challengE (Interspeech ComParE) series revisited SER as task in 2013. Meanwhile, challenges considering media-material such as clips of films appeared, namely the annual (since 2013) Emotion in the Wild Challenge (EmotiW[14]) run, and the new Multimodal Emotion Challenge (MEC 2016 and MEC 2017[n]). A loser relation to emotion in speech is given in further challenges such as MediaEval[o] ("affective (2015) /emotional (2016) impact of movies" task).

The accompanying table presents an overview on the challenges and their results to date that focused on SER. Interestingly, all challenges used the same feature extractor for the baselines. For comparison, the AVEC 2016 results for end-to-end learning[37] and Bags-of-Audio-Words[31] are further given, which are no official baselines. At press time, the series MEC is rerun, and the series ComParE is calling for participation for their 2018 reinstantiations offering novel affect tasks on atypical and self-assessed affect.

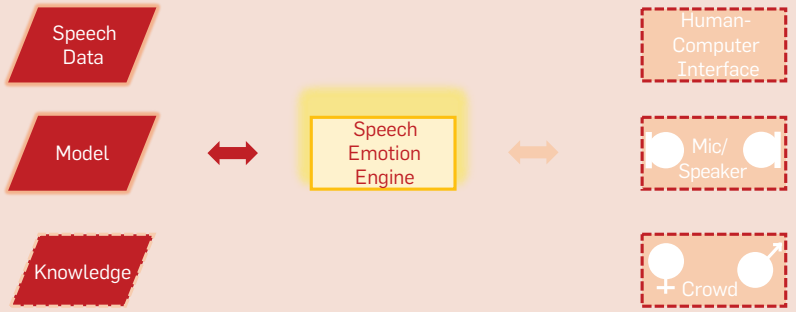
One would wish to compare these challenges in terms of technical or chronological improvements over the years. However, as the table indicates, the same database was used only once in two challenges with the same task definition (AVEC 15/16). There, one notices a striking improvement in the baseline of this challenge in the more recent edition. It seems desirable to rerun former tasks more often for a better comparability across years rather than having a mere provision of snapshots. However, the table shows that the task attempted was becoming increasingly challenging, going from lab to voice over IP to material from films with potential audio overlay.

Further, one would want to see the results of these events set into relation with human emotion perception benchmarks. Again, this is not straightforward for the following reasons: the ground truth does not exist in a reliable way—the data was labeled by a small number of humans in the first place. Comparing it to the perception of other humans on the test data would thus not be entirely fair, as they would likely have a different perception from those who labeled the training and test data. Further, there simply is no perception study available on these sets, indicating another white spot in the tradition of challenge culture in the field. Perhaps the more important question would be how these results relate to acceptable rates for human-machine applications. Such numbers are unfortunately also largely missing and would need to be provided by application developers stemming from according usability studies. To provide a statement non-the-less, the technology can already be used in a range of applications as outlined above, and seems to improve over time, reaching closer to human performance.

m http://sspnet.eu/avec201x, with $x \in [1–7]$
n http://www.chineseldc.org/htdocsEn/emotion.html
o http://multimediaeval.org

Figure 2. A modern speech emotion recognition engine.

Ideally, this engine experiences life-long learning 24/7 to analyze and synthesize emotion across languages and cultures. It gains feedback from the crowd in cooperative, gamified, and reinforced ways learning end-to-end and transferring gained knowledge. For a holistic understanding, it integrates contextual knowledge such as other speaker states and traits.



Moonshot Challenges?

Seeing the ongoing trends in the field, one may wonder what is left as high hanging fruit, grand challenge, or even a moonshot challenge. Certainly, several further steps must be taken before SER can be considered ready for broad consumer usage "in the wild." These include robustness across cultures and languages as one of the major white spots in the literature. A number of studies show the downgrades one may expect when going cross-language in terms of acoustic emotion recognition.[3] As to cross-cultural studies, these are still particularly sparse, and there exists practically no engine that is adaptive to cultural differences at the time. Beyond cross-cultural robustness, such against atypicality must be further investigated. For example, a few studies deal with emotion portrayal of individuals on the autism spectrum.[30] Further, the assessment of emotion of speaker groups has hardly been targeted. In a first step, this requires dealing with far-field acoustics, but it also must ideally isolate speakers' voices to analyze overlapping speech in search of emotional cues to then come to a conclusion regarding a groups' emotion. A potentially more challenging task may then be the recognition of irony or sarcasm as well as regulation of emotion. Differences between the acoustic and the linguistic channels may be indicative, but the research up to this point is limited. Next, there is little work to be found on speaker long-term adaptation, albeit being highly promising.

A genuine moonshot challenge, however, may be to target the *actual* emotion of an individual sensed by speech analysis. Up to this point, the gold standard is to use other human raters' assessment, that is, ratings or

annotations, as an "outer emotion", as perceived by others, as learning target. Obviously, this can be highly different from the "inner emotion" of an individual. To assess it, one will first need a ground truth measurement method, for example, by deeper insight into the cognitive processes as measured by EEG or other suited means. Then, one will also have to develop models that are robust against differences between expressed emotion and the experienced one—potentially by deriving further information from the voice which is usually not accessible to humans such as the heart rate, skin conductance, current facial expression, body posture, or eye contact,[32] and many further bio-signals.

Obviously, one can think of many further interesting challenges such as emotion recognition "from a chips bag" by high-speed camera capture of the vibrations induced by the acoustic waves,[9] in space, under water, and, of course, in animal vocalizations.

Conclusion

In this article, I elaborated on making machines hear our emotions from end to end—from the early studies on acoustic correlates of emotion[8,16,21,34] to the first patent[41] in 1978, the first seminal paper in the field,[10] to the first end-to-end learning system.[37] We are still learning. Based on this evolution, an abstracted summary is shown in Figure 2 presenting the main features of a modern engine. Hopefully, current dead-ends, such as the lack of rich amounts of spontaneous data that allow for coping with speaker variation, can be overcome. After more than 20 years into automatic recognition of emotion in the speech signal, we are currently witnessing exciting times of change: data learned features, synthesized training material, holistic architectures, and learning in an increasingly autonomous way—all of which can be expected to soon lead to the rise of broad day-to-day usage in many health, retrieval, security, and further beneficial use-cases alongside—after years of waiting[36]—the advent of emotionally intelligent speech interfaces.

Acknowledgments

The research leading to these results has received funding from the European Union's HORIZON 2020 Framework Programme under the Grant Agreement No. 645378.

References

1. Abdelwahab, M. and Busso, C. Supervised domain adaptation for emotion recognition from speech. In *Proceedings of ICASSP*. (Brisbane, Australia, 2015). IEEE, 5058–5062.
2. Anagnostopoulos, C.-N., Iliou, T. and Giannoukos, I. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review 43*, 2 (2015), 155–177.
3. Bhaykar, M., Yadav, J. and Rao, K.S. Speaker dependent, speaker independent and cross language emotion recognition from speech using GMM and HMM. In *Proceedings of the National Conference on Communications*. (Delhi, India, 2013). IEEE, 1–5.
4. Blanton, S. The voice and the emotions. *Q. Journal of Speech 1*, 2 (1915), 154–172.
5. Chang, J. and Scherer, S. Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks. *arxiv.org*, (arXiv:1705.02394), 2017.
6. Chen, L., Mao, X., Xue, Y. and Cheng, L.L. Speech emotion recognition: Features and classification models. *Digital Signal Processing 22*, 6 (2012), 1154–1160.
7. Cibau, N.E., Albornoz. E.M., and Rufiner, H.L. Speech emotion recognition using a deep autoencoder. San Carlos de Bariloche, Argentina, 2013, 934–939.
8. Darwin, C. *The Expression of Emotion in Man and Animals*. Watts, 1948.
9. Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G. J., Durand, F. and Freeman, W.T. The visual microphone: Passive recovery of sound from video. *ACM Trans. Graphics 33*, 4 (2014), 1–10.
10. Dellaert, F., Polzin, T. and Waibel, A. Recognizing emotion in speech. In *Proceedings of ICSLP 3*, (Philadelphia, PA, 1996). IEEE, 1970–1973.
11. Deng, J. Feature Transfer Learning for Speech Emotion Recognition. PhD thesis, Dissertation, Technische Universität München, Germany, 2016.
12. Deng, J., Xu, X., Zhang, Z., Frühholz, S., and Schuller, B. Semisupervised Autoencoders for Speech Emotion Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 26*, 1 (2018), 31–43.
13. Devillers, L., Vidrascu, L. and Lamel, L. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks 18*, 4 (2005), 407–422.
14. Dhall, A., Goecke, R., Joshi, J., Sikka, K. and Gedeon, T. Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of ICMI* (Istanbul, Turkey, 2014). ACM, 461–466.
15. El Ayadi, M., Kamel, M.S., and Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition 44*, 3 (2011), 572–587.
16. Fairbanks, G. and Pronovost, W. Vocal pitch during simulated emotion. *Science 88*, 2286 (1938), 382–383.
17. Gunes, H. and Schuller, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing 31*, 2 (2013), 120–136.
18. Joachims, T. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers, 2002.
19. Kim, Y., Lee, H. and Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of ICASSP*, (Vancouver, Canada, 2013). IEEE, 3687–3691.
20. Koolagudi, S.G. and Rao, K.S. Emotion recognition from speech: A review. *Intern. J. of Speech Technology 15*, 2 (2012), 99–117.
21. Kramer, E. Elimination of verbal cues in judgments of emotion from voice. *The J. Abnormal and Social Psychology 68*, 4 (1964), 390.
22. Kraus, M.W. Voice-only communication enhances empathic accuracy. *American Psychologist 72*, 7 (2017), 644.
23. Lee, C.M., Narayanan, S.S., and Pieraccini, R. Combining acoustic and language information for emotion recognition. In *Proceedings of INTERSPEECH*, (Denver, CO, 2002). ISCA, 873–876.
24. Leng, Y., Xu, X., and Qi, G. Combining active learning and semi-supervised learning to construct SVM classifier. *Knowledge-Based Systems 44* (2013), 121–131.
25. Liu, J., Chen, C., Bu, J., You, M. and Tao, J. Speech emotion recognition using an enhanced co-training algorithm. In *Proceedings ICME*. (Beijing, P.R. China, 2007). IEEE, 999–1002.
26. Lotfian, R. and Busso, C. Emotion recognition using synthetic speech as neutral reference. In *Proceedings of ICASSP*. (Brisbane, Australia, 2015). IEEE, 4759–4763.
27. Mao, Q., Dong, M., Huang, Z. and Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimedia 16*, 8 (2014), 2203–2213.
28. Marsella, S. and Gratch, J. Computationally modeling human emotion. *Commun. ACM 57*, 12 (Dec. 2014), 56–67.
29. Picard, R.W. and Picard, R. *Affective Computing*, vol. 252. MIT Press Cambridge, MA, 1997.
30. Ram, C.S. and Ponnusamy, R. Assessment on speech emotion recognition for autism spectrum disorder children using support vector machine. *World Applied Sciences J. 34*, 1 (2016), 94–102.
31. Schmitt, M., Ringeval, F. and Schuller, B. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Proceedings of INTERSPEECH*. (San Francisco, CA, 2016). ISCA, 495–499.
32. Schuller, B. and Batliner, A. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2013.
33. Schuller, B, Mousa, A. E.-D., and Vasileios, V. Sentiment analysis and opinion mining: On optimal parameters and performances. *WIREs Data Mining and Knowledge Discovery* (2015), 5:255–5:263.
34. Soskin, W.F. and Kauffman, P.E. Judgment of emotion in word-free voice samples. *J. of Commun. 11*, 2 (1961), 73–80.
35. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G. and Schuller, B. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *Proceedings of ICASSP*. (Prague, Czech Republic, 2011). IEEE,5688–5691.
36. Tosa, N. and Nakatsu, R. Life-like communication agent-emotion sensing character 'MIC' and feeling session character 'MUSE.' In *Proceedings of the 3rd International Conference on Multimedia Computing and Systems*. (Hiroshima, Japan, 1996). IEEE, 12–19.
37. Trigeorgis, G., Ringeval, F., Brückner, R., Marchi, E., Nicolaou, M., Schuller, B. and Zafeiriou, S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of ICASSP*. (Shanghai, P.R. China, 2016). IEEE, 5200–5204.
38. Ververidis, D. and Kotropoulos, C. Emotional speech recognition: Resources, features, and methods. *Speech Commun. 48*, 9 (2006), 1162–1181.
39. Watson, D., Clark, L.A., and Tellegen, A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. of Personality and Social Psychology 54*, 6 (1988), 1063.
40. Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M., and Scherer, K.R. On the acoustics of emotion in audio: What speech, music and sound have in common. *Frontiers in Psychology 4*, Article ID 292 (2013), 1–12.
41. Williamson, J. Speech analyzer for analyzing pitch or frequency perturbations in individual speech pattern to determine the emotional state of the person. U.S. Patent 4,093,821, 1978.
42. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E. and Cowie, R. Abandoning emotion classes—Towards continuous emotion recognition with modeling of long-range dependencies. In *Proceedings of INTERSPEECH*. (Brisbane, Australia, 2008). ISCA, 597–600.
43. Zeng, Z., Pantic, M., Roisman, G.I., and Huang, T.S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence 31*, 1 (2009), 39–58.

Björn W. Schuller (schuller@tum.de) is a professor and head of the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at the the University of Augsburg, Germany.

Watch the author discuss his work in this exclusive *Communications* video. https://cacm.acm.org/videos/speech-emotion-recognition