Using AI to predict service agent stress from emotion patterns in service interactions

Citation for published version (APA):

Bromuri, S., Henkel, A. P., Iren, D., & Urovi, V. (2021). Using AI to predict service agent stress from emotion patterns in service interactions. *Journal of Service Management*, *32*(4), 581-611. Advance online publication. https://doi.org/10.1108/josm-06-2019-0163

DOI:
10.1108/josm-06-2019-0163

Document status and date:
Published: 10/09/2021

Document Version:
Publisher's PDF, also known as Version of record

Using AI to predict service agent stress from emotion patterns in service interactions

Stefano Bromuri
Computer Science, Open University of the Netherlands, Heerlen, The Netherlands
Alexander P. Henkel
Organization, Open University of the Netherlands, Heerlen, The Netherlands
Deniz Iren
Information Systems, Open University of the Netherlands, Heerlen, The Netherlands, and
Visara Urovi
Institute of Data Science, Maastricht University, Maastricht, The Netherlands

Received 1 June 2019
Revised 1 November 2019
13 May 2020
24 July 2020
Accepted 28 August 2020

Abstract

Purpose – A vast body of literature has documented the negative consequences of stress on employee performance and well-being. These deleterious effects are particularly pronounced for service agents who need to constantly endure and manage customer emotions. The purpose of this paper is to introduce and describe a deep learning model to predict in real-time service agent stress from emotion patterns in voice-to-voice service interactions.

Design/methodology/approach – A deep learning model was developed to identify emotion patterns in call center interactions based on 363 recorded service interactions, subdivided in 27,889 manually expert-labeled three-second audio snippets. In a second step, the deep learning model was deployed in a call center for a period of one month to be further trained by the data collected from 40 service agents in another 4,672 service interactions.

Findings – The deep learning emotion classifier reached a balanced accuracy of 68% in predicting discrete emotions in service interactions. Integrating this model in a binary classification model, it was able to predict service agent stress with a balanced accuracy of 80%.

Practical implications – Service managers can benefit from employing the deep learning model to continuously and unobtrusively monitor the stress level of their service agents with numerous practical applications, including real-time early warning systems for service agents, customized training and automatically linking stress to customer-related outcomes.

Originality/value – The present study is the first to document an artificial intelligence (AI)-based model that is able to identify emotions in natural (i.e. nonstaged) interactions. It is further a pioneer in developing a smart emotion-based stress measure for service agents. Finally, the study contributes to the literature on the role of emotions in service interactions and employee stress.

Keywords Customer service employees, Call center service interactions, Speech emotion recognition, Stress detection, Deep learning, Artificial intelligence

Paper type Research paper

Work stress is widely regarded as the central cause of a myriad of deleterious critical organizational and personal health consequences (Cropanzano *et al.*, 2003; Ganster and Rosen, 2013; Halbesleben and Bowler, 2007; Wang, 2005). For instance, it is associated with absenteeism (Grandey *et al.*, 2004), under-performance and turnover (Wright and

The authors would like to thank the participating companies for providing access to their data as well as all service agents for their dedication in coding the customer service interactions.

Funding: This research was supported by the Province of Limburg, The Netherlands, under grant number SAS-2019-00247.

Journal of Service Management
Vol. 32 No. 4, 2021
pp. 581-611
© Emerald Publishing Limited
1757-5818
DOI 10.1108/JOSM-06-2019-0163

Cropanzano, 1998), burnout (Maslach and Leiter, 2016) and negative cardiovascular responses (Vrijkotte *et al.*, 2000). Customer service agents (hereafter "service agents" for brevity) in particular must pay a high toll for constant exposure to emotionally demanding customers and required adherence to organizational display rules in the form of heightened stress levels (Dormann and Zapf, 2004; Grandey *et al.*, 2007). A plethora of research has investigated the cause-and-effect relationships underlying work stress (e.g. Alarcon, 2011; Eatough *et al.*, 2011), yet the field has repeatedly reiterated calls for more forward-looking and innovative studies creating actionable insights for managers and employees in order to avoid the negative consequences of stress (Bliese *et al.*, 2017).

With the advent of artificial intelligence (AI) and its capacity to learn to interpret human behavior, new opportunities have surfaced to help organizations prevent the undesirable effects of work stress, rather than merely to treat its symptoms. The first step toward preventing the escalation of stress consequences lies in periodically monitoring it. At the moment, the most universally accepted work stress assessment is performed using an employee self-assessment (Stanton *et al.*, 2001). This measure was developed to circumvent the shortcomings of other self-reported, behavioral or biological measures in terms of impracticality and intrusiveness. Yet, a self-report measure still bears a range of limitations, as it only provides cross-sectional information in a relatively obtrusive and highly subjective form.

To overcome these limitations, the underlying paper developed a deep neural network (DNN) (LeCun *et al.*, 2015) model as a real-time, early stress detection tool in voice-to-voice service interactions. A DNN mimics the structure of a human brain and is composed of multiple layers of artificial neurons. The multilayer structure ensures that neurons at higher layers will activate when a combination of neurons at lower layers are activated, providing the neural network the ability to recognize complex combinations of properties in a signal. This study trained a DNN based on the most important stressors identified in prior service management work: interpersonal stressors emanating from the constant exposure of service agents to customer emotions and the related emotion regulation requirements (Baranik *et al.*, 2017; Grandey, 2003; Grandey *et al.*, 2007; Lewig and Dollard, 2003). Getting insights into employee stress levels in real-time before they escalate would create a wealth of opportunities for managers to combine this information with prevention and coping strategies, and create better early stress detection alerting systems, customized employee training and cognitive-behavioral interventions (Richardson and Rothstein, 2008).

Guided by the motivation to innovate and improve the assessment of service agent stress to achieve the aforementioned benefits from a service research perspective and to broaden the methodological repertoire of service researchers, this study addresses two main methodological research questions. The first question relates to the modeling of a DNN to handle sequences of emotions in an interaction. The second question concerns the further use of the emotion patterns extracted from a recorded interaction to predict service agent stress. The study describes how to use a neural network trained on speech emotion recognition (SER) to perform stress prediction during a service interaction. Answering these questions would also extend theories in the field of emotion analytics (Burkhardt *et al.*, 2005; Busso *et al.*, 2008; Livingstone and Russo, 2018). To the best of our knowledge, no research so far has attempted to develop and evaluate emotion recognition and stress detection deep learning models that (1) are based on real human interactions, (2) take into consideration the sequences of emotions in an interaction and (3) relate emotions to stress.

The remainder of the paper is structured as follows. It first provides a succinct review of stress-related studies in call center service work, before discussing relevant related work on SER and stress detection. It then describes the data and methodology and continues with evaluation of the AI-based machine learning architectures. Finally, it links back the findings to theory and service management practice. The list of acronyms used in this paper is provided in the Appendix, in Table A1.

1. Theoretical background

A large yet fragmented body of work has examined service agent stress in a call center context, mainly related to emotional labor and interpersonal stressors, and role stress. Call center work is defined by constantly changing interactions with customers and tight organizational display rules (Grandey *et al.*, 2010). The lack of autonomy paired with frequent incidents of interpersonal stressors in the form of customer incivility has been identified as a particularly toxic combination for the performance and well-being of service agents (e.g. Gabriel and Dieffendorf, 2015; Ganster and Rosen, 2013; Rafaeli *et al.*, 2012). Similarly detrimental consequences have been reported for experiencing role stress (e.g. De Ruyter *et al.*, 2001).

Some effective organizational measures to counter or prevent service agent stress identified in service management research include: providing display autonomy, such that agents do not have to suppress their true emotions (Goldberg and Grandey, 2007), adopting an empowered leadership style (De Ruyter *et al.*, 2001) and ensuring fair supervisor treatment (van Jaarsveld *et al.*, 2019). Another theoretically promising, yet practically challenging provision is hiring service agents with a high level of trait optimism (Tuten and Neidermeyer, 2004). Also service management practice has replied to the challenge of service agent stress using several technology-based innovations, including: adaptive hygiene factors (e.g. adapting room temperature, lighting, canceling white noise), automatically filtering out fraudulent incoming calls, scripting software and personalized dashboards to help track the own progress (Call Centre Helper, 2017).

Irrespective of the precautions taken, eventually every service agent will encounter interpersonal stressors in the form of difficult or uncivil customers. On the level of service agents, engaging in emotion-focused coping is a frequently applied strategy to manage the job stressors they encounter. While such reactions (e.g. seeking social support with coworkers, avoidance or venting) may prove effective, some of them may ironically endanger service quality if service agents behaviorally disengage from customers or infect them with negativity (Goussinsky, 2012). In summary, the causes of and coping strategies for stress both revolve around emotions, and reducing and managing stress is of central concern to service agents and managers. Yet, even the latest technological innovations in service practice do not focus on the detection of and personalized intervention for service agent stress. This leaves ample potential for the deployment of state-of-the-art technology to help monitor service agent stress in real time, while there is still time to counter its negative consequences (Bliese *et al.*, 2017). The next section reflects on relevant work in the field of SER and stress detection, with the aim of providing the theoretical basis for such an innovative intelligent tool.

2. Speech emotion recognition and stress detection models

The main challenge of SER is to categorize a voice signal into an emotion with the help of machine learning methods. Machine learning is a discipline aimed at exploiting data to create a model of a process that a human cannot characterize otherwise. The main advantage of machine learning with respect to standard engineering modeling is that it substitutes the step of acquiring specialized domain knowledge with obtaining data describing the process or system behavior of interest.

A machine learning data set may consist of vectors of numerical, ordinal and categorical data which together comprise “features” to predict a target variable. If the target variable is present in a machine learning process, the learning algorithms for the prediction are called “supervised,” otherwise they are called “unsupervised.” The underlying study considers the special case of supervised tasks dealing with predicting categorical target variables. Such a supervised task is also known as a “classification task.” A learning algorithm is trained by means of a training data set and an optimization objective suitable for the selected target.

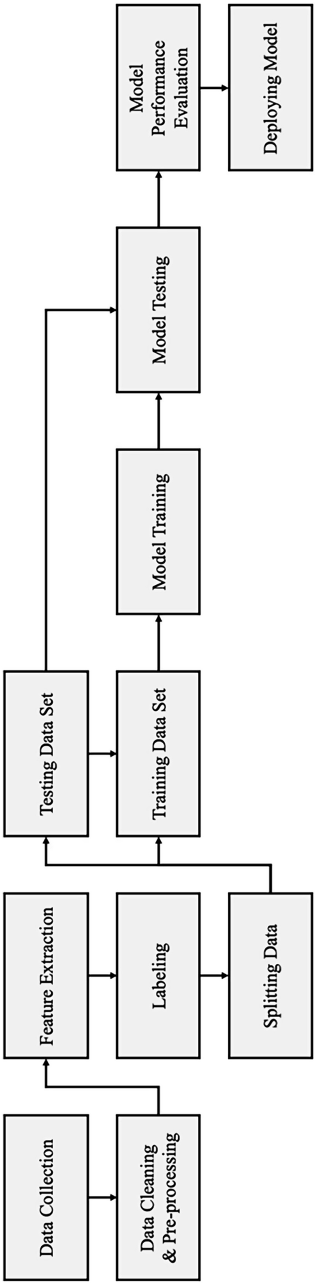


Figure 1.
Machine learning process

Typically, in addition to a training set, it is also necessary to specify a test set to discover how well the algorithm performs with respect to unseen data. Thus, a machine learning process implies the stages shown in Figure 1.

Features may represent basic properties associated with an item to be categorized (e.g. columns in a table), or a combination of properties (e.g. a weighted average of the columns) which may highlight the category of the item better than the basic properties in isolation. In the latter case, a feature extraction process is applied on the data before the training of the classification algorithm.

The most basic feature of a voice signal is the numerical value of the sequence representing the encoding with which the voice is numerically recorded. In the case of telephony and digital computers, the standard approach is to encode the analog voice signal through a sampling rate and a bit depth. The sampling rate comprises the number of samples recorded per second and the bit depth the number of possible digital values that each sample can take. In addition to its encoding, a voice signal can be studied in the time domain, in the frequency domain, or in both domains combined by decomposing the signal in its frequency components (see Appendix, Figure A1). Extracting features from either domain may be useful for the classification process (Masui and Wada, 2004). Research on emotion recognition in speech can be broadly divided into traditional and more recent approaches. Within the former, time-frequency features (e.g. Kleijn, 1991) are manually engineered by a domain expert before being fed to a machine learning algorithm (Casale *et al.*, 2008). The latter, in contrast, rely on deep learning to automatically identify an optimal combination of time-frequency features for the categorization task (e.g. Huang *et al.*, 2019).

2.1 Speech emotion recognition

SER deals with the challenges associated with detecting emotions conveyed in speech. SER unites two interrelated fields of inquiry; one pertaining to the features used to classify the emotions, and one focusing on machine learning models. As with any machine learning task, the use of rich features can help produce a superior result. Advanced machine learning models, such as deep learning, may automate part or all of the feature extraction. From a features perspective, SER evolved from considering only basic time-based features (e.g. frequency of spoken words, amount of silences; Lehiste and Lass, 1976) to complex time-frequency features such as the spectrogram of the speech signal (i.e. Huang *et al.*, 2019). From the machine learning model perspective, until 2014 the focus has been on using standard machine learning models (e.g. Casale *et al.*, 2008), and afterward it increasingly shifted toward using deep learning models, thanks to the introduction of graphical learning units that allowed to speed up the training of such models (Chen and Lin, 2014).

Deep learning models of choice have been long short-term memories (LSTMs, see Appendix, Section 3) (Hochreiter, 1998), for their ability to deal with long numerical sequences and convolutional neural networks (CNN; see Appendix, Section 3.1), where numerical filters in the form of matrices, or vectors, are multiplied by the original signal to retain features that are relevant to the categorization to be performed. An interesting aspect of both LSTMs and CNNs is that, once trained, the network's intermediate layers can be transferred to related machine learning tasks using similar data, but different targets. Transfer learning (i.e. reusing a network on a different target) is in all effects an additional feature extraction approach beyond the time-frequency human engineered features.

Most work in SER is based on staged conversations, where one or more actors are given a set of sentences to reproduce with a specific scripted emotion. Following this schema, several data sets have been defined to serve as benchmarks for emotion recognition performance, such as IEMOCAP, EMO-DB and RAVDESS (Burkhardt *et al.*, 2005; Busso *et al.*, 2008; Livingstone and Russo, 2018). IEMOCAP consists of 12 h of audiovisual data, in dyadic sessions performed by actors who improvise scripted scenarios and sentences of variable lengths, each containing emotional responses such as anger, happiness, sadness and neutrality. EMO-DB consists of 500 three-second utterances of actors emulating an emotional tone (e.g. happy, angry, anxious, fearful, bored and disgusted). Finally, RAVDESS is composed of 24 recordings of actors vocalizing two lexically matched statements of variable lengths in order to produce calm, happy, sad, angry, fearful, surprised and disgusted emotion expressions of different intensity (i.e. normal and strong).

While these data sets have advanced the SER field considerably, they present two limitations. First, acted emotions may not present the same features as spontaneous emotions. Second, the acted emotions are evenly distributed in the data set, which is an unrealistic assumption in nonstaged conversations. More recently, first attempts at classifying emotions have been undertaken to counter the main criticism toward staged data sets, by focusing on data collected in spontaneous and natural settings (e.g. Huang *et al.*, 2019). Table 1 provides an overview of the most important articles related to the underlying study and visualizes its methodological contribution in the field of SER.

The feature extraction methods in Table 1 constitute time, frequency and transfer learning to reflect the evolution of the SER field concerning the approach toward feature extraction. The method column reflects the use of standard machine learning, deep learning or both, and the data type column reflects the use of staged data versus realistic data.

2.2 Stress detection

Stress detection, like SER, is a field that is gaining momentum (Slavich *et al.*, 2019), in relation to observable physiological signals (e.g. (Pisanski *et al.*, 2018), or specifically to speech (Szaszak *et al.*, 2018). For instance, Paulmann *et al.* (2016) explore how induced psychological stress affects the production and recognition of emotions in speech. They further document that stress is recognizable in a speaker's voice even by non-native listeners, but that negative emotions are recognized more poorly when produced by stressed speakers than by nonstressed speakers. As the underlying reason, they refer to the variability of the voice of the stressed speaker, which may differ from the expectations of the receiver in terms of frequency and pitch.

The work of Szaszak *et al.* (2018) focuses on modeling stress as related to prosodic features (i.e. tones, pauses, pitch and tempo) in speech. In this sense their work is close to the underlying study, as part of the features extracted with the audio-signal analysis software library that were used in here, also take into consideration prosodic features (i.e. PyAudio Analysis; Giannakopoulos, 2015). The main difference though resides in the fact that this study reused the output of a network pretrained on emotion recognition for the purpose of classifying stress. These emotion features allow to consider long sequences of speech as they occur in customer service interactions and thus classify stress through the full interaction. Another study relevant to the underlying context investigated the relationship between hormones and stress response in a trial with 80 individuals. Results suggest that stress is associated with a systematic increase in voice pitch, cortisol hormone levels and a decrease in skin temperature in both genders (Pisanski *et al.*, 2018).

From a technical perspective, also the usage of machine learning models to detect events in speech is gaining momentum. Han *et al.* (2018) use a deep learning model to detect stress in speech features. Yet, they do so, applying it on short sequences whereas the underlying study uses bottleneck features to handle long sequences as they unfold in service interactions. Mun *et al.* (2016) and Xia *et al.* (2018) apply a similar approach to this by using intermediate outputs of a neural network to detect events in speech, although not focusing on stress detection in particular. The next section describes the construction of the data sets that are employed for building the model.

Table 1. Speech emotion recognition-related work

Author (Year)	Feature extraction Time domain features	Frequency domain features	Transfer learning	Method Standard machine learning	Deep learning models	Data type Staged data	Realistic data	Key findings	Implications and intuition
James (1976)	x						x	Establishes the problem of collecting prosodic features (such as fundamental frequency, rhythm and stress of the voice) for the purpose of automatic evaluation of speech related tasks	Fundamental frequency is a feature that concerns the pitch of the person, where the rhythm can be associated with excitement. The stress in the voice can be associate with how hard the words are pronounced, distinguishing between positive and negative emotions
Lehiste and Lass (1976)	x						x	Identifies suprasegmental features, a set of features consisting of pitch, stress and quantity, as a fundamental part of speech. This work was important at observing that speech can be represented as a string of its subparts representing phonetic units, which if characterized can lead us to a better interpretation of the speech signal	The relationship between the phonetic units has a time effect that can help distinguishing between emotions
Imai (1983)		x					x	Discusses how to create a spectral representation of speech signal, by representing the spectral envelope of the speech signal in terms of its log magnitude on the nonlinear frequency scale (known as mel-log-spectrum). The main finding is that the mel-cepstrum is a good representation of a speech signal	A spectral representation is in essence a subdivision of a signal in its composing frequencies. The amplitude of the signal associated with the frequencies composing the voice have been shown to be a good feature for classification purposes in multiple voice classification tasks
Kleijn (1991)	x							Discusses linear predictive coding (LPC) as an effective time-based technique to extract features in speech signals	LPC is based on linear regression. The regression is fit to the audio signal and the coefficients of the regression become a highly descriptive feature for the machine learning process

(*continued*)

Table 1.

	Feature extraction			Method		Data type			
Author (Year)	Time domain features	Frequency domain features	Transfer learning	Standard machine learning	Deep learning models	Staged data	Realistic data	Key findings	Implications and intuition
New *et al.* (2003)	x	x		x		x		Assesses the viability of hidden Markov models (HMM) to recognize emotional speech in 6 classes of emotions, using Mel-cepstrum coefficients and LPC for feature extraction. The developed model is trained on staged data for the Mandarin and Burmese languages and it shows an average accuracy of 78% on the staged data	HMM model sequences by assuming that there are states emitting the sequence in exam. The training algorithm for HMM identifies the relationships between the states. In a machine learning process the final state is the class of the signal
Wu *et al.* (2011)	x	x	x	x		x	x	Fuses several feature extractions such as LPC, mel frequency cepstrum coefficients (MFCC), modulation spectral features (MSF) and prosodic features. The data used include. The fused features are then fed to a support vector machine classifier. The key finding is that MSF features obtain the best results in the recognition of emotions, with an accuracy above 90%, and that the results can be transferred in another data set	Time and frequency features both contribute to the classification task. It also shows that more complex combinations of the features (like MSF) may also contribute to a better classification result. Finally, it applies transfer learning, training an algorithm in one data set and then applying it in another data set
Pan *et al.* (2012)	x	x		x		x		Complements Wu *et al.* (2011) by studying the use of MFCC features on a Berlin emotion corpus and Chinese emotion corpus, obtaining accuracy rates above 90%	MFCC features are effective features to identify emotions in multiple languages

(*continued*)

Table 1.

Author (Year)	Feature extraction Time domain features	 Frequency domain features	 Transfer learning	Method Standard machine learning	 Deep learning models	Data type Staged data	 Realistic data	Key findings	Implications and intuition
Mao *et al.* (2014)	x				x	x		Uses a two stages approach to learn local invariant features (LIF) out of sparse autoencoders neural networks, concerning speech data containing emotions. After this stage, the LIF features are further refined to represent affective speech, by means of a salient features discriminant analysis (SFDA) approach. Such features are then compared with traditional MFCC and LPC features showing that they are superior at classifying emotions	A neural network can be used to extract features without the need to apply another time (LPC) or frequency-based method (MFCC). This implies that features obtained from the intermediate layers of a neural network are highly descriptive summarization of the signal and can be effectively used for classification tasks
Oord *et al.* (2016)	x	x			x		x	Defines a new model of neural networks called WaveNet which can generate realistic speech in text to speech tasks and produce precise classification of phonemes when used as discriminative models. Shows that neural network architectures are currently the state of the art in handling sound-related tasks thanks to their ability to handle autoregressive data with nonlinear filters	CNNs and LSTMs are very effective concerning representation tasks in audio applications, given that they produce superior quality results in audio generation tasks

(*continued*)

Table 1.

Author (Year)	Feature extraction			Method		Data type		Key findings	Implications and intuition
	Time domain features	Frequency domain features	Transfer learning	Standard machine learning	Deep learning models	Staged data	Realistic data		
Satt *et al.* (2017)	x	x			x	x		Demonstrates the use of spectrograms in combination with convolutional neural networks to classify emotions in snippets of three seconds of speech extracted from the IEMOCAP data set. Spectrograms in combination with produce a comparable performance to MFCC features with low latency	Neural networks are able to handle unprocessed representations of the signals, such as spectrograms, in a way that is comparable to performing a feature extraction on the signal and then applying standard machine learning models. The main implication is that neural networks can learn to automatically extract rich features from the data
Liu *et al.* (2018)	x	x			x		x	Combines speaker related and speaker unrelated features (calculated using MFCC) to perform the discrete classification of emotional speech. As a classification method, it uses extreme machine learning decision trees. The main finding is that, when possible, personalizing the algorithm to the speaker can hold good results on staged data	Including information on the speaker, when and if available, and personalizing the classification, improves the results concerning emotion analytics tasks
Zhao *et al.* (2018)	x	x			x	x		Evaluates the use of advanced neural network architectures including 1D and 2D convolutional blocks on spectrograms and mel-cepstrum coefficients. The main finding is that 2D convolutional networks combined with LSTM models outperform traditional architectures when trained on staged emotion data sets	2D and 1D CNNs combined with LSTMs produce good results in identifying emotions in audio data, which provides a direction to the current study toward a family of neural network architectures to be used

(*continued*)

	Feature extraction			Method		Data type			
Author (Year)	Time domain features	Frequency domain features	Transfer learning	Standard machine learning	Deep learning models	Staged data	Realistic data	Key findings	Implications and intuition
Guo *et al.* (2018)	x	x			x			Complementing Zhao *et al.* (2018), discusses the use of 2D CNN in spectrogram of the speech data, but in addition extreme learning machines (ELM) are used to learn additional features which are then fused with the spectrogram for the final classification. The performance, calculated on a staged data set, shows that such a combination holds better results than only using spectrograms	The CNN/LSTM approach also perform well for feature extraction purposes as they can be combined with ELM to improve the classification accuracy
Badshah *et al.* (2019)	x	x			x	x		Proposes a CNN-based model in which the filter of the CNN uses a rectangular shape, rather than the typical squared shape (as for example used by Guo *et al.* (2018) and Zhao *et al.* (2018)). Varying the size and dimension of the filter shows high performance in two staged data sets, the EmoDB and the Korean Speech Corpus	Modifying the neural network architectures may produce improvement on baseline neural architectures, which means that multiple architectures should be attempted in a study to identify the best performing one

(*continued*)

Table 1.

Table 1.

	Feature extraction			Method		Data type			
Author (Year)	Time domain features	Frequency domain features	Transfer learning	Standard machine learning	Deep learning models	Staged data	Realistic data	Key findings	Implications and intuition
Huang *et al.* (2019)	x	x			x		x	Verbal and nonverbal parts of speech are used in two parallel neural networks with attention, whose output is then combined to produce a classification of the emotion in spontaneous speech data from the NMIME corpus. The key finding is that combining verbal and nonverbal parts of speech increases the accuracy of the emotion classifier (37% precision and recall, when considering full match in a time window) and that spontaneous speech signals pose more challenges than staged speech signals	Realistic data are more challenging to model than staged data, due to the presence of nonverbal communication, such as pauses and sighing. This implies that models trained on staged data present limitations when used in real scenarios due to the simplifying assumption with which the data set has been created
The underlying study	x	x	x		x		x	Uses both time and frequency properties of the speech to predict emotions and stress. The emotions are modeled as a sequence related with time-frequency features of the speech. In addition, the stress network makes use of transfer learning through the intermediate layers of the emotion network, allowing the prediction of stress over long conversations	Networks trained to recognize emotions can be reused for additional higher-level tasks, such as detecting service agent stress in real time

3. Method

The development of an emotion recognition algorithm for a customer service context implies privacy constraints. In traditional emotion databases, the transcript of a recording and the recording itself are fully available to the researcher. In this case, data from actual customer service interactions can be processed only in a completely anonymized fashion. To this end, the content and the recording have to pass through a feature extraction process that ensures that the original signal (i.e. the content of the conversation) cannot be reconstructed. These are common challenges encountered in analyses of natural speech data from real-world social contexts (Devillers and Vidrascu, 2007). Next follows a presentation of the data sets used in this study and a motivation of the modeling choices to deal with the challenges specified above.

3.1 Description of data sets

Relying on the voices of actors staging an emotion presents several methodological problems for SER. In addition to altering both the distributions and the features associated with the emotions, the audio snippets are short and present no relationship with each other. This fragmented incoherence of the audio snippets used for training impairs the predictive power of the machine learning models in real-world applications. Rather, emotions only become interpretable and meaningful in the context in which they are expressed. To address these shortcomings, this study works with anonymized conversation data collected from the call centers of two major pension service providers in the Netherlands, with audio signals of real customer service interactions. Building a machine learning model to classify an audio snippet of a service interaction into emotion classes presents a number of challenges. First, the emotion patterns are difficult to characterize and annotate, as emotions can be expressed in a variety of ways. Second, as emotions represent sensitive information, only those service agents who have conducted the call are allowed to annotate it. Third, emotions are context dependent and imbalanced. Fourth, a distinguishable emotion may not be present or may not be explicitly expressed during a specific point in time.

The data set is composed of two parts. The first part is based on recordings of service interactions that were expert coded by professional service agents based on six basic emotions (Ekman and Friesen, 1969). These data were used as predeployment training data. Subsequently, the emotion-labeled conversation data were used to train an automated emotion recognition model that uses a DNN architecture. These postdeployment data also contain the associated subjective stress scores of a service interaction as indicated by the service agents. More detailed information on the data sets is presented below.

Predeployment training data. In the predeployment phase, 363 recordings of call center conversations with an average duration of 183 s were collected (see Appendix, Figure A2). These conversations took place between service agents of the two pension fund service providers and customers who contacted the service provider about pension-related topics, such as the status of their current pension portfolio. The sensitivity and the seriousness of the topic gave rise to a multitude of emotions reflected in the tone of voice. Twenty experienced senior service agents annotated the conversations by associating speech-related features of voice signals with emotions reflected in the tone of the voice. In order to safeguard privacy, service agents exclusively listened to recordings they participated in. Audio files were split into sequences of three-second snippets and were presented to the annotators in sequential order.

Although emotions constitute an important component of human communication, there is no fully-agreed-upon definition and typology of emotions today (Ekman and Cordaro, 2011; Izard, 1992; Plutchik, 1982; Tomkins, 1963). The probably most universally accepted approach is provided by Ekman's theory of basic emotions (see Ekman and Cordaro, 2011).

Ekman postulates the existence of only six basic emotions (i.e. sadness, happiness, anger, surprise, disgust and fear). These basic emotions are universally exhibited in facial expression and perceived in the same way across cultures (Ekman and Friesen, 1969). Hence, this study took these emotions as a starting point for the automated categorization. The initial label list thus comprised six discrete emotions and an additional neutral class to indicate the lack of any distinguishable emotion. The annotation software did not allow assigning a neutral label together with an emotion label or no label at all. A total of 27,889 snippets of three second duration were annotated through this method.

In many instances, the annotators were not able to assign a snippet to an emotion class (e.g. neutral voice, silence) resulting in a majority of neutral labels during the training of the algorithm. Further, annotators struggled to distinguish between the approach-motivated emotions of anger and disgust and the avoidance-motivated emotions of fear and surprise, respectively (Elliot and Thrash, 2002). Finally, since annotators had difficulties identifying sadness, and since sadness only represented 0.6% of annotated emotions, this study took a conservative approach and collapsed those instances with the other instances of nonspecified emotions (i.e. neutral).

This categorization into four emotion categories is in line with recent research challenging the long-standing theory of six basic emotions (Ekman and Cordaro, 2011). Rather, following an evolutionary account, anger and disgust, as well as fear and surprise may be rooted in the same basic emotion, respectively (Jack *et al.*, 2016). Following a bottom-up approach in conjunction with an evolutionary perspective, this study hence grouped emotions into four different categories as suggested by the data and in line with Jack *et al.* (2016). To summarize, this study used the following aggregated four labels to train the DNN emotion classifier: (1) happiness (4.2% of the snippets), (2) surprise and fear (4.5% of the snippets), (3) anger and disgust (3.44% of the snippets), (4) neutral and sadness (87.86% of the snippets). The next section discusses how these data were used for training the machine learning model.

Postdeployment data. For a period of one month after deploying the emotion recognition model in the two call centers, data were collected comprising 4,672 service interactions for which the voice features and related emotion predictions were recorded. Each recorded call comprises both the vocal features of the customer and the service agent and the emotion labels provided by the models as explained in the previous section. Therefore, a total of

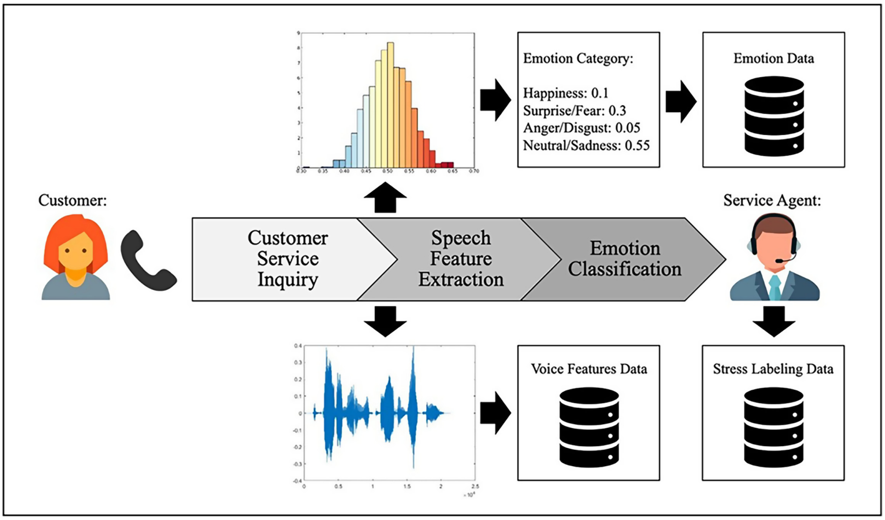


Figure 2.
Data collection procedure

519,370 voice feature vectors and associated emotion scores were collected concerning the 4,672 recordings. Figure 2 depicts the protocol used for the recognition and storage of the emotions in the second phase of the project.

Immediately after each conversation, one of 40 service agents was asked to indicate the perceived stressfulness of the conversation on a 7-point scale from not at all to very much after Grandey *et al.* (2004). Possibly due to social desirability, service agents mainly (above 94%) used the lower end of the scale to only report nuances in increases of perceived stress (i.e. 1–3 on the scale). This trait of the data provided an additional challenge for the training of the algorithm as it needed to identify these nuances. In total, about one-third of all calls (i.e. 28.5%) was annotated with at least some level of stress and subsequently used as input for the deep learning model in the category of stressful calls (i.e. 2–7 on the scale). A limitation that was imposed by the participating companies is that the 40 agents annotated an unequal amount of calls. This uneven distribution implies that the categorization of stressful calls was dominated by those 15 service agents who annotated most of the calls. The entire data set was anonymized before further processing.

3.2 From data collection to feature extraction

The collected predeployment data consist of audio recordings of conversations and associated emotion annotations that need further refinement to serve for the training of a machine learning algorithm. First, each of the collected audio recordings is split into three-second snippets. Second, each of the three-second snippets is sampled in order to extract features from the audio signals (see a comprehensive summary of the features used in Table 2) by using Py Audio Analysis (Giannakopoulos, 2015). The sampling of the features, for each three-second snippet, is performed with a 25 ms window and a step of 50 ms.

3.3 Sequence representation

In the particular case of emotion recognition, the usual approach is to label each audio snippet with an emotion without explicitly considering a dependency between the emotions. This study, instead, considers a sequential dependency between the emotions (see Appendix, Section 5). In the speech recognition model, each of the extracted features refers three-second audio snippets, appended to each other in batches of five, resulting in 15 s snippets. Each single three-seconds interval contains as target the emotion annotation provided by a service agent.

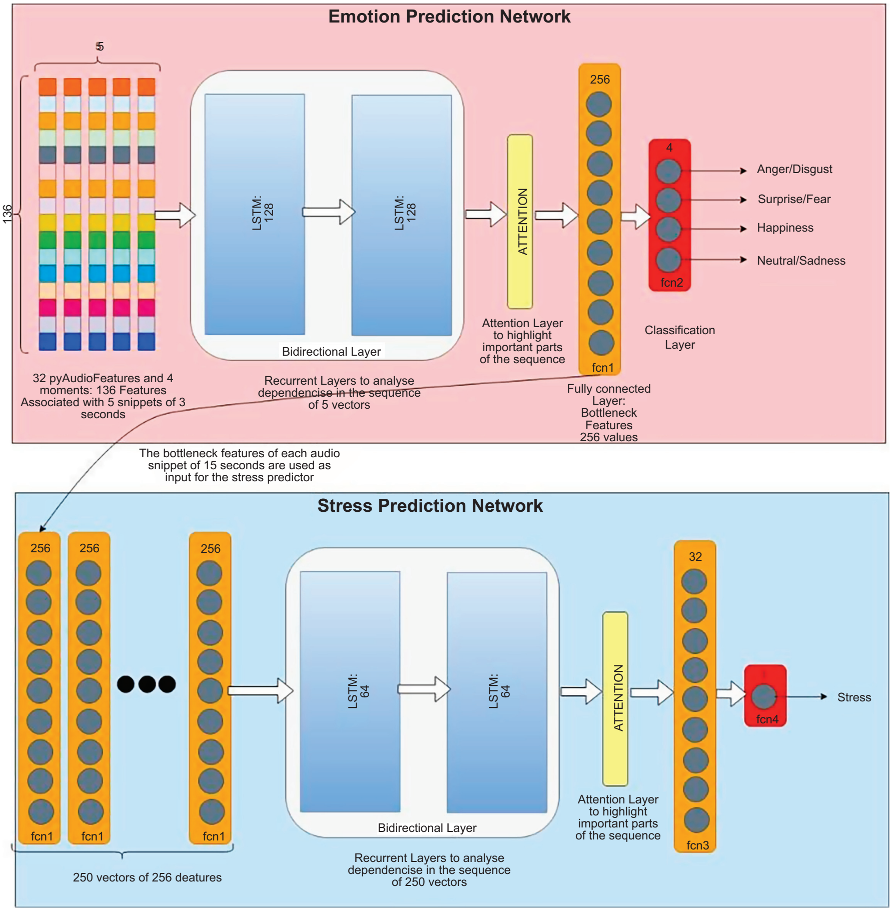
3.4 Deep learning model: long short-term memory (LSTM) networks

Deep learning (LeCun *et al.*, 2015) is a subfield of machine learning that focuses on the creation of networks as multiple interconnected layers of artificial neurons (see Appendix, Section 6). This multilayer structure provides the neural network the ability to recognize complex combinations of properties in a signal and to produce a categorization of the input signal. Recurrent neural networks (RNN) are neural networks whose output does not only depend on the input but also on the previous states of the network. Thus, RNNs have a memory, as there is a feedback loop that feeds the output of the network back into its input (see Appendix, Section 3). This study makes use of LSTM networks, a type of RNN network that can deal with data represented by long sequences as in the case of speech signals.

The LSTM-based architectures used in this study are shown in Figure 3. The bidirectional LSTM model (bi-LSTM) considers recurrent relationships that take place in the sequence from left to right and from right to left and as such, improve the results in speech recognition tasks (Graves *et al.*, 2013). A key part of the models is the attention layer (Bahdanau *et al.*, 2016) constructed on top of the bi-LSTM. Attention is able to focus on prior sequence states

Index	Name	Description	Intuition
1	Zero-crossing rate	Rate of sign changes of the signal during the duration of a particular frame	Represents the rate at which the signal switches sign; in speech signals it offers a good feature to distinguish between periods of silence and periods of speech
2	Energy	Sum of squares of signal values, normalized by the respective frame length	Represents the energy of the signals; in a speech signal it represents the sound volume across the frequencies
3	Entropy of energy	The entropy of subframes' normalized energies. It can be interpreted as a measure of abrupt changes	Represents the rate of abrupt changes that can be associated with the excitement in the speech signal
4	Spectral centroid	The center of gravity of the spectrum	Characterizes areas of voiced signals when the vocal cords vibrate versus unvoiced signals when the vocal cords do not vibrate. The vibration of the vocal cords changes depending on the emotion expressed, therefore, this feature can help characterizing the emotions category
5	Spectral spread	The second central moment of the spectrum	Specifies the spread of the speech signal around a frequency and thus, characterizes the dominance of a tone. The sequences of tones used can help identify emotions
6	Spectral entropy	Entropy of the normalized spectral energies for a set of subframes	Is used with spectral centroid (see 4) to identify areas of voiced speech versus unvoiced speech, with similar implications concerning emotions
7	Spectral flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames	Calculates the changes of power in the spectrum, therefore, it can indicate how fast the exchange between two speakers takes place, or how the speaker is modulating the voice. A big variability implies presence of emotions whereas no variability implies a neutral voice
8	Spectral rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated	Emotions present different distributions of the spectrum given the same uttered word, so the spectral rolloff helps distinguish between emotions
9–21	Mel-cepstral coefficients	Mel-frequency cepstral coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale	The frequency representation of the audio signal can help highlight fundamental frequencies to distinguish an emotion, similarly to the spectral rolloff (see 8)
22–33	Chroma vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of Western-type music (semitone spacing)	Chroma vectors have been used to perform musical emotion recognition. Similar to how the minor or major chords change the mood of a music piece, in a speech signal, the speech tonality affects the perception and emotion associated with the spoken words
34	Chroma deviation	The SD of the 12 chroma coefficients	Provides a measure of how variable the tonality of the spoken words is. A very narrow chroma deviation may imply a very neutral speech signal

Table 2.
Spectral and prosodic features extracted from the speech signal of the call center


Emotion Prediction Network
5
136
32 pyAudioFeatures and 4 moments: 136 Features Associated with 5 snippets of 3 seconds
LSTM: 128
LSTM: 128
Bidirectional Layer
Recurrent Layers to analyse dependencise in the sequence of 5 vectors
ATTENTION
Attention Layer to highlight important parts of the sequence
256
fcn1
Fully connected Layer: Bottleneck Features 256 values
4
fcn2
Classification Layer
Anger/Disgust
Surprise/Fear
Happiness
Neutral/Sadness
The bottleneck features of each audio snippet of 15 seconds are used as input for the stress predictor
Stress Prediction Network
256
256
256
fcn1
fcn1
fcn1
250 vectors of 256 deatures
LSTM: 64
LSTM: 64
Bidirectional Layer
Recurrent Layers to analyse dependencise in the sequence of 250 vectors
ATTENTION
Attention Layer to highlight important parts of the sequence
32
fcn3
fcn4
Stress

Note(s): The first network uses speech features calculated with PyAudioAnalysis to perform a prediction of the 4 emotion labels. The penultimate layer of the emotion network (fc1) is implemented with RELU neurons and is connected with then final layer of the network (fc2), that is a fully connected softmax layer outputting the probability of an emotion. The stress prediction network uses the output of fc1 of the first network to calculate the probability of stress in the full call

Figure 3.
LSTM-based architecture for speech emotion recognition

and combine this information with the output of the LSTM. This results in certain parts of the input sequence having more or less impact on the outcome of the model. In the case of the emotion prediction network, the attention model learns to assign more importance to those speech features that contribute the most toward a correct emotion classification. In the case of the stress prediction network, the attention model learns to weigh as more important the part of the phone call that is perceived as most stressful.

3.5 Using the emotion recognition model features

In addition to the task of predicting the emotions in a service interaction, the most relevant methodological contribution of this study is to use the emotion stream and the weights of the

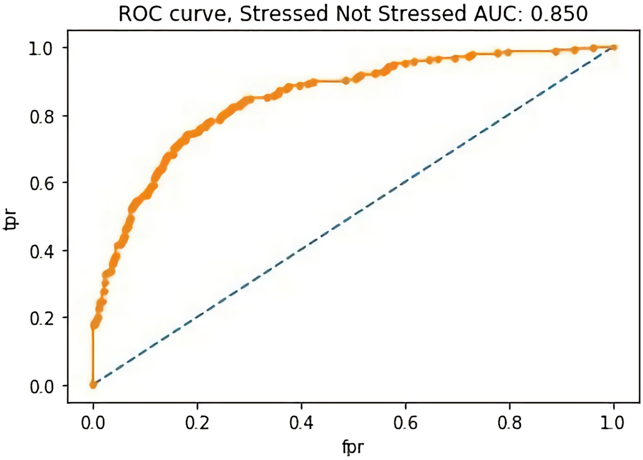


Figure 4. ROC curve and area under the curve (AUC) of the stress prediction network

Note(s): The x-axis denotes the false positive rate (fpr) and the y-axis denotes the true positive rate (tpr). The areaunder the curve (AUC) represents the probability for which a sample labelled as stressful is ranked higher thanan example labelled as non-stressful. Hence, an AUC of 0.85 indicates a strong classifier for stress

network to further classify the calls by predicting if a call has been stressful or not for a service agent. For this purpose, the output of the emotion prediction network is used to train additional classifiers. The features extracted by the emotion recognition model (see FC1 in Figure 5) serve as input to a second neural network to further classify the calls as stressful or not stressful. The usage of intermediate layers outputs of neural networks trained for other targets is an approach to extract compressed meaningful features from complex data (e.g. Xia *et al.*, 2018). Intermediate layers features taken before the last layer of the network are known as “bottleneck features” since they reduce the dimensions of the data coming from the previous layers of the network before being categorized in the last network layer.

In this study, such bottleneck features of the emotion model are optimized for the purpose of categorizing a voice signal in emotion categories given a sequence of low-level features (e.g. time-frequency features). They are further used as input to the stress neural network to further classify the signal in containing stress or not as depicted in Figure 5. This approach allows to model longer sequences than with the original speech features used to classify emotions that are too heavy to handle for the stress prediction neural network in terms of both computational load and memory consumption. The network weights are optimized such that the predictions of the network output approximate the target variable as closely as possible (e.g. using categorical cross-entropy; see Appendix Section 3.2 for a more detailed explanation).

4. Evaluation and results

This section reports the results of the emotion recognition and stress recognition models along two dimensions. The first dimension pertains to predicting emotions, where it shows the performance of the emotion recognition model to classify emotions in four classes. The second dimension comprises the prediction of the stressfulness of a call. As evaluation metrics, this study uses precision, recall, *f*1 score and micro-average, all of which are commonly used performance measurement metrics in the fields of information retrieval and

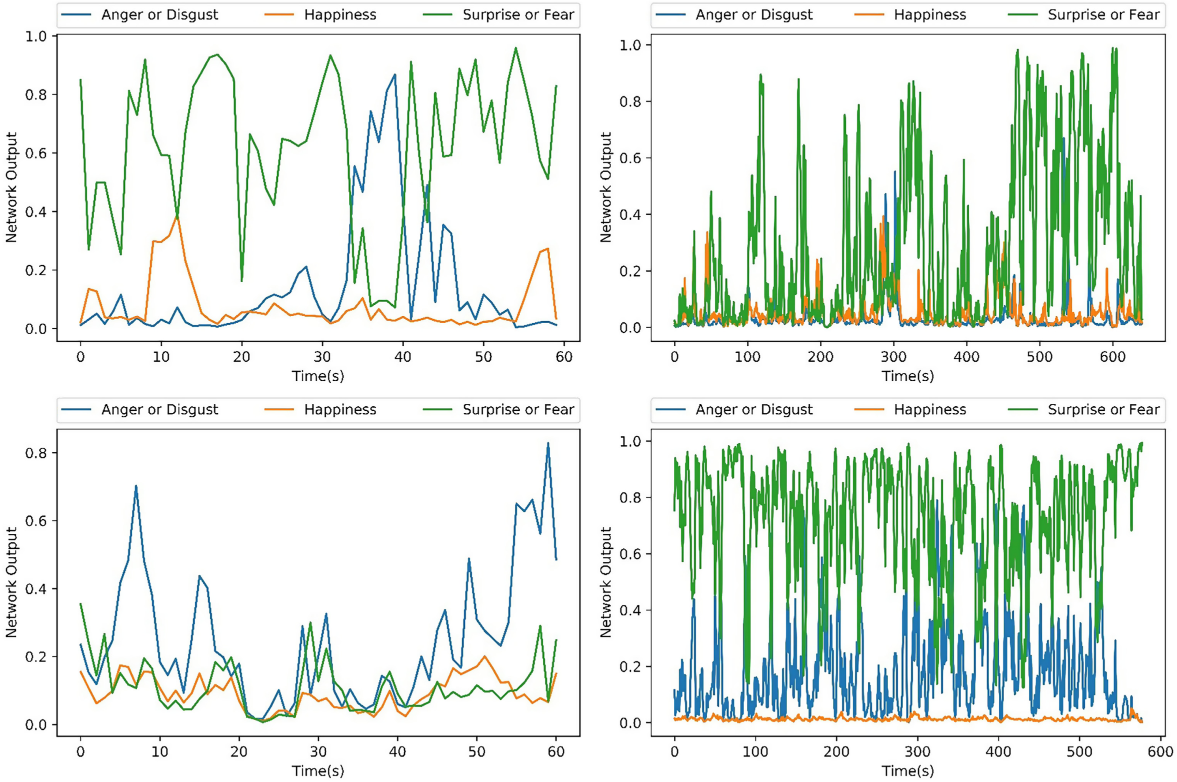


Figure 5.
Emotion content of high/low stress calls

Note(s): Top left: short call without stress, emotion breakdown. Top right: long call without stress, emotion breakdown. Bottom left: short call with stress, emotion breakdown. Bottom right: long call with stress, emotion breakdown. The y-axis enotes the network output, representing the confidence that the network has concerning the presence of an emotion at a specific point in time machine learning (Buckland and Gey, 1994). Precision specifies the ability of a machine learning model to avoid misclassifying negative elements as positive ones. The formula of precision is expressed as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP indicates the true positive rate and FP the false positive rate. Recall represents a ratio between the positive elements that have been correctly identified, and the sum of the total positive elements, including those misclassified as negative elements. As such, it represents the percentage of relevant patterns that the model could recall from the data, which can be expressed with the following formula:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where FN indicates the rate of false negatives. The $f1$ score or $f1$ measure calculates the harmonic mean of precision and recall, weighting the two of them with the same importance:

$$f1\,\text{score} = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

Finally, the micro-average calculates the weighted average of the three metrics expressed above with respect to the (imbalanced) classes in the data set.

4.1 Predicting emotions

The 363 predeployment calls are analyzed using the aggregated labeling proposed in the previous section. The following architecture variations are attempted:

(1) DLA: this network uses 1D convolutions, bidirectional LSTMs, attention, optimized with categorical cross-entropy.

(2) LA: a bidirectional LSTM network with one objective function and attention model.

(3) BIL: a bidirectional LSTM model.

(4) Frequency classifier: this is a baseline random classifier based on the frequencies of the data points.

For the analysis, 70% of the data is used for training purposes, and 30% of the data for testing. Table 3 reports the results of the testing procedure concerning the four selected classifiers, reporting also the 95% confidence interval.

As reported in Table 3, the three attempted models all remain significantly predictive with respect to a baseline frequency classifier, indicating that a pattern can be identified that connects the labels with the speech features. On the one hand, the neutral class precision and recall is higher than the other classes, meaning that the algorithm can distinguish audio snippets with emotional response versus audio snippets in which there is no emotional response. On the other hand, the precision and recall concerning the emotions all revolve around 20%. This means that the neural network model may rank the emotion labeled by the service agent as less predominant than the other emotions. This can happen for several reasons. First, multiple emotions may occur in the same audio snippet: anger and surprise or happiness and surprise can often occur at the same time and the service agent has labeled the most salient or context-relevant emotion and ignored the others. Second, emotions that are far apart like happiness and anger, are sometimes both expressed with rising voice volume levels. Third, though the content may suggest the presence of emotions, the latter may not be expressed in the voice of the customer. Such results are consistent with

Table 3. Emotion recognition results

	Prec	1DLA Recall	*f1*	Prec	LA Recall	*f1*	Prec	BIL Recall	*f1*	Prec	Frequency classifier Recall	*f1*
Anger/ Disgust	0.21 (0.04)	0.12 (0.04)	0.17 (0.04)	0.21 (0.04)	0.17 (0.03)	0.19 (0.03)	0.27 (0.05)	0.18 (0.03)	0.22 (0.03)	0.08 (0.03)	0.09 (0.03)	0.08 (0.03)
Fear/ Surprise	0.23 (0.04)	0.25 (0.04)	0.19 (0.04)	0.19 (0.03)	0.19 (0.03)	0.19 (0.03)	0.21 (0.03)	0.21 (0.03)	0.21 (0.03)	0.12 (0.03)	0.15 (0.03)	0.13 (0.03)
Happiness	0.25 (0.04)	0.07 (0.03)	0.1 (0.03)	0.24 (0.04)	0.26 (0.05)	0.27 (0.05)	0.28 (0.05)	0.17 (0.04)	0.21 (0.04)	0.1 (0.03)	0.09 (0.03)	0.09 (0.03)
Sadness/ Neutral	0.78 (0.01)	0.88 (0.01)	0.83 (0.01)	0.8 (0.01)	0.81 (0.01)	0.81 (0.01)	0.8 (0.01)	0.86 (0.01)	0.83 (0.01)	0.73 (0.01)	0.68 (0.01)	0.71 (0.01)
Micro-avg	0.67 (0.01)	0.67 (0.01)	0.67 (0.01)	0.68 (0.01)	0.68 (0.01)	0.68 (0.01)	0.68 (0.01)	0.68 (0.01)	0.68 (0.01)	0.53 (0.01)	0.53 (0.01)	0.53 (0.01)

Note(s): The upper statistic represents the mean and the lower statistic in-between brackets represents the 95%-confidence interval ($p < 0.05$)

the results obtained by Huang *et al.* (2019) who also use spontaneous speech to train their deep learning models.

4.2 Predicting stressfulness of a call

Given that LA architecture (LSTM model with attention) has better precision and recall than the other architectures for emotion recognition, this study focuses on this architecture to extract global features from a call and produce a stress prediction based on the stress prediction network depicted in Figure 5. After the analysis with the LA architecture, the intermediate output is a set of 4,672 time series of different lengths, which are each composed of four emotion classes. In order to classify the calls into stressful or nonstressful, the stream of emotions associated with each of the calls are transformed into vector format using one of the intermediate layers of the emotion recognition neural network. For this purpose, as the study is dealing with sequences, it applies a similar network as the one presented in Figure 5.

After transforming the calls in vector format, each conversation is represented as a sequence of 250 vectors of 256 dimensions which are analyzed using a stress predictor LSTM-based network. One limitation of this approach, given that LSTMs accept only sequences of regular size, is that sequences that are shorter than 250 audio snippets have to be padded (a standard approach is to use vectors filled with zeros) and sequences that are longer than 250 audio snippets have to be truncated. The 250-snippet length threshold is chosen to cover the length of 95% of the calls in the training data. The data are split into 80% for training and 20% for testing, after which a second network is trained to perform a binary prediction (stressed/not stressed) on sequences of audio snippets. Table 4 depicts the results of the prediction.

The data set stress labels are unbalanced, showing a predominance of calls in which the service agents selected no stress. Only one-third of the calls are labeled as stressful. Despite this imbalance, the algorithm manages to predict the positive class (stressful call) with a precision of 0.68 and a recall of 0.7, with an average precision of 0.8 for and an average recall of 0.8 for the two classes combined. The ROC curve in Figure 4 confirms that a strong classifier can be trained, presenting an area under the ROC curve (AUC) of 0.85 (Pepe, 2000).

The stress prediction becomes more interpretable when observing the emotion patterns. For this purpose, Figure 5 breaks down short and long calls with high and low stress, respectively. As depicted in the figure, for both short and long calls with low stress, the dominant emotion that is captured is surprise (the single peak of anger in the middle of the short call likely represents an artifact rather than a true expression of anger). The stressful calls, in contrast, contain clearly discernible patterns of anger, while happiness is absent.

5. Discussion

This study provides an emotion analytics approach to analyzing customer service interactions in a call center context. Two deep learning models are presented: the first model analyzes emotions in customer service interactions with an average accuracy of 68%

Table 4.
Stress prediction network performance

Stress network predictor	Precision	Recall	f1-score
No-stress	0.86 (0.03)	0.84 (0.03)	0.85 (0.03)
Stress	0.68 (0.06)	0.7 (0.06)	0.69 (0.06)
Micro-avg	0.8 (0.03)	0.8 (0.03)	0.8 (0.03)

Note(s): The upper statistic represents the mean and the lower statistic in-between brackets represents The 95%-confidence interval ($p < 0.05$)

in a four-class problem; the second model builds on the first model to predict the perceived stressfulness of a call with an average accuracy of 80%. This study makes important contributions to three different research fields: emotion analytics, organizational work stress and service management.

5.1 Implications for emotion analytics

This study contributes to the state-of-the-art of emotion analytics by considering real interactions taking place between service agents and customers. This is an important advancement, as currently, most speech emotion analytics research utilizes data models that have been created on nondyadic staged conversations (e.g. Casale *et al.*, 2008; Zhao *et al.*, 2019). The staged conversation data sets consist of unnaturally balanced emotion distributions. A related issue is that emotions appear in isolation in staged conversations whereas, in real conversations they are interrelated. Moreover, staged data sets lack the context of actual conversations where emotions are manifested. These three attributes hinder the performance of emotion classification on real-life conversations, and make the results difficult to interpret due to the difference in contexts between staged and real conversations. Thus, this study contributes to the literature on emotion analytics as it trains models on emotions as they unfold in their actual environment and within the context of real service interactions, keeping the emotion distribution natural for both training and testing phases.

Yet, the underlying work also cautions the use of more subtle emotions in practice, as both AI predictions and the human expert labeling struggled to identify the six discrete basic emotions from voice (Ekman and Cordaro, 2011). The data support a simpler categorization that is in line with more recent advancements of an evolutionary theory (Jack *et al.*, 2016). This evolutionary theory suggests that anger and disgust, as well as fear and surprise, are rooted in the same basic emotion, respectively. Thus, at least in a voice-based real-life context, where emotions are expressed naturally by customers rather staged by actors (Burkhardt *et al.*, 2005; Busso *et al.*, 2008), the informational value of adding further emotion categories to predict emotion-based phenomena such as stress might be poor. Moreover, this is the first study in which information is derived from emotions (i.e. stress) that extends beyond their labeling (Huang *et al.*, 2019; Liu *et al.*, 2018). Using the bottleneck features of an LSTM network trained on discrete customer emotions to predict the employee perceived stress of a service interaction reached an accuracy of 80%.

5.2 Implications for organizational work stress

The newly developed AI model makes stress quantifiable in near-real-time and in a reliable and entirely unobtrusive fashion. Thus, this study contributes to prior work on occupational stress (Bliese *et al.*, 2017; Maslach and Leiter, 2016; Zapf *et al.*, 2001), where the gold standard of stress assessment relies on periodic employee self-assessments (Stanton *et al.*, 2001). While this self-assessment measure was developed to circumvent the shortcomings of other self-report (Cavanaugh *et al.*, 2000) and physiological measures (Vrijkotte *et al.*, 2000) in terms of impracticality and intrusiveness, the AI model takes this development one step further. In doing so, the study directly answers frequently echoed calls to develop a more forward-looking and innovative approach to providing actionable insights on work stress (Bliese *et al.*, 2017), moving beyond traditional cause-and-effect relationships (e.g. Alarcon, 2011; Eatough *et al.*, 2011). Relying on the latest technological advancements in the area of deep learning (Zhao *et al.*, 2019), the model allows the continuous monitoring of service agent stress without the need for physical (e.g. wearing patches on the skin) or psychological (e.g. filling in surveys) interventions.

5.3 Implications for service management

This study also contributes to the field of service management. It directly answers to a recent call for service research to put a stronger attention on an employee perspective, with a particular focus on occupational stress (Subramony *et al.*, 2017). The nature of the service frontline is changing and the integration of AI in frontline service interactions is a core priority (Huang and Rust, 2018). Service agents will partly be replaced by AI (e.g. humanoid service robots replacing service agents at information points) and they will also have to learn to collaborate with AI (e.g. humanoid service robots waiting tables at a restaurant). The focus of the stress detection algorithm developed in this study lies in the latter – the integration of AI and service agents (Henkel *et al.*, 2020; Marinova *et al.*, 2017; Wilson and Daugherty, 2018). However, the tool also offers benefits beyond the mere augmentation of service agents (Larivière *et al.*, 2017). On a meta-level, it can offer insights to service managers to the benefit of service agents and customers.

The stress detection model carries several actionable practical applications in a service agent context and as such carries manifold opportunities for future research. First, the stress detection model can be used for managers to observe and intervene between calls, before a particular service agent's stress level escalates beyond a predefined threshold where additional stress might be detrimental to performance and well-being (e.g. Cropanzano *et al.*, 2003; Ganster and Rosen, 2013; Halbesleben and Bowler, 2007; Wang, 2005). The granularity of the observation period is subject to future research, yet, it is likely that in this context the algorithm will be more effective in a tool that has a more long-term focus (Schneiderman *et al.*, 2005). Second, the stress algorithm could be integrated into another intelligent algorithm that automatically allocates customers to service agents through a routing system (Gans and Zhou, 2007), based on a real-time assessment of service agent stress and customer emotions. That way highly stressed service agents could be shielded from stressful, aggressive complaining customers (Rupp and Spencer, 2006). Third, it could be deployed as an early warning system for service agents, such that the latter can pay attention to the building of stress throughout the workday over an extended period. An extended version of the AI tool could directly indicate to service agents appropriate cognitive-behavioral intervention strategies to avoid the negative consequences of stress (Richardson and Rothstein, 2008).

Beyond predicting stress, the newly developed AI model allows the assessment of the presence and variation of emotions during a service interaction (Hareli and Rafaeli, 2008) by looking at the output of the emotion prediction network in time. Since neural network outputs can take a probabilistic interpretation, this feature is independent of the classification accuracy. It opens up various opportunities for service managers. For instance, given the high turnover rate of call center employees and its direct link with emotional exhaustion (Kraemer and Gouthier, 2014), the algorithm may be deployed to predict turnover intentions and employee resilience (Britt *et al.*, 2016), based on emotion interaction patterns in customer service calls.

Other applications extend to the customer's side, where uncovering patterns in service interactions may be input to improving customer outcomes. Knowing the emotional journey of a customer and complementing this information with, for instance, other CRM data can be a powerful tool to identify service bottlenecks to further improve the customer experience throughout the customer journey (Lemon and Verhoef, 2016). The algorithm developed in the underlying study may also serve as the basis for an automated emotion-based prediction of customer satisfaction (Farhadloo *et al.*, 2016) on an individual basis via mapping emotions across services and topics to predicting future customer behavior. We submit it to future research to build on our work and explore these promising lines of inquiry linked to service interactions and customer outcomes. It may also be conceivable to deploy the tool in a face-to-face service context. With the advent of service robots in the organizational frontline (Wirtz *et al.*, 2018), a particularly promising area of application also lies in the integration of emotion detection from human voice in human robot interaction (Hudlicka, 2003). With the help of the stress algorithm, service robots might be trained to sense the stress level of their interaction partners, which may be customers or human colleagues.

5.4 Limitations and future research

As with every research, also the underlying study does not come without its limitations. First, the analysis highlights that in a call center service context, the emotions observed in a service interaction are highly unbalanced. Moreover, the task of labeling this type of data introduces a degree of subjectivity that makes the classification and prediction a difficult task at the level of the discrete emotions associated with an audio snippet emanating from a real-life service interaction. The main hurdle is the subjective interpretation of emotions associated with a specific call, varying based on the emotion recognition experience and the emotional intelligence of the listener (Salovey and Mayer, 1990). This explains why some discrete emotions tend to be more easily confused by the deep learning model than others. Notwithstanding these challenges, the underlying study demonstrates that basic discrete emotions can still be input to a powerful interpretation tool when predicting global call attributes, such as service agent stress. Also, it is important to remember that this study has been applied in the specific context of pension service providers.

In addition to the already mentioned opportunities, future work might assess how the findings hold throughout different contexts, demographic groups and cultures. For instance, the stress algorithm might need fine-tuning or retraining for service interactions underlying different display rules across different cultural dimensions (Grandey *et al.*, 2010). It is conceivable that the model works even better in cultures that license emotional expression, since more emotion cues are discernible from a customer's voice (Pennebaker *et al.*, 1996). On the other hand, service agents in these cultures may also experience and interpret customer emotions differently in terms of the stress they produce.

It is conceivable that the emotion-based stress detection algorithm that was developed in a call center context may also be applied in other types of service encounters. First, the tool could potentially be deployed in traditional face-to-face settings with high customer traffic to help select and train service agents, but also to monitor their stress levels to help them avoid the negative consequences for their well-being and the firm's bottom-line. Second, the algorithm may also find an application in digital services where customers solely interact with technology. For instance, it may be relevant for inhabitants of a smart home which consumers control with voice commands that the latter caters to their momentary emotion- and stress-related needs (e.g. the atmosphere of the light).

Irrespective of the type of service interaction in which the stress algorithm is deployed, it is crucial to mention that first a diligent assessment of its ethical and privacy implications needs to be performed. For the scope of this study, the participating companies collected the consent from both service agents and customers, ensuring that the collected data could only be used for its original purpose – the training of service agents to better handle customer interactions. However, as is the case with many smart technologies, the stress algorithm can work hidden in the background, which depending on the context may be unethical or even illegal. Even if deployed with the best intentions, service managers should ensure that the resulting data are not abused. Whatever direction the further development and practical application of the stress algorithm developed in this study may bring, we hope that it will stimulate initiatives that have the well-being of service agents at the center (Anderson and Ostrom, 2015).

References

Alarcon, G.M. (2011), "A meta-analysis of burnout with job demands, resources, and attitudes", *Journal of Vocational Behavior*, Vol. 79 No. 2, pp. 549-562.

Anderson, L. and Ostrom, A.L. (2015), "Transformative service research: advancing our knowledge about service and well-being", *Journal of Service Research*, Vol. 18 No. 3, pp. 243-249.

Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M.Y., Kwon, S. and Baik, S.W. (2019), "Deep features-based speech emotion recognition for smart affective services", *Multimedia Tools Applications*, Vol. 78 No. 5, pp. 5571-5589.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y. (2016), "End-to-end attention-based large vocabulary speech recognition", *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, Shanghai, China, March 20-25, 2016, IEEE, pp. 4945-4949.

Baranik, L.E., Wang, M., Gong, Y. and Shi, J. (2017), "Customer mistreatment, employee health, and job performance: cognitive rumination and social sharing as mediating mechanisms", *Journal of Management*, Vol. 43 No. 4, pp. 1261-1282.

Bliese, P.D., Edwards, J.R. and Sonnentag, S. (2017), "Stress and well-being at work: a century of empirical trends reflecting theoretical and societal influences", *Journal of Applied Psychology*, Vol. 102 No. 3, pp. 389-402.

Britt, T.W., Shen, W., Sinclair, R.R., Grossman, M.R. and Klieger, D.M. (2016), "How much do we really know about employee resilience?", *Industrial and Organizational Psychology*, Vol. 9 No. 2, pp. 378-404.

Buckland, M. and Gey, F. (1994), "The relationship between recall and precision", *Journal of the American Society for Information Science*, Vol. 45 No. 1, pp. 12-19.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. and Weiss, B. (2005), "A database of German emotional speech", *Ninth European Conference on Speech Communication and Technology*, Lisbon, Portugal, September 4-8, 2005, ISCA, pp. 1517-1520.

Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S. and Narayanan, S.S. (2008), "IEMOCAP: interactive emotional dyadic motion capture database", *Language Resources and Evaluation*, Vol. 42 No. 4, pp. 335-339.

Call Centre Helper (2017), "13 ways technology can reduce agent stress", available at: https://www.callcentrehelper.com/13-ways-technology-can-reduce-agent-stress-90198.htm (accessed 16 July 2020).

Casale, S., Russo, A., Scebba, G. and Serrano, S. (2008), "Speech emotion classification using machine learning algorithms", *Proceedings of the 2nd IEEE International Conference on Semantic Computing (ICSC 2008)*, August 4-7, 2008, Santa Clara, California, IEEE Computer Society, pp. 158-165.

Cavanaugh, M.A., Boswell, W.R., Roehling, M.V. and Boudreau, J.W. (2000), "An empirical examination of self-reported work stress among US managers", *Journal of Applied Psychology*, Vol. 85 No. 1, pp. 65-74.

Chen, X.W. and Lin, X. (2014), "Big data deep learning: challenges and perspectives", *IEEE Access*, Vol. 2, pp. 514-525.

Cropanzano, R., Rupp, D.E. and Byrne, Z.S. (2003), "The relationship of emotional exhaustion to work attitudes, job performance, and organizational citizenship behaviors", *Journal of Applied Psychology*, Vol. 88 No. 1, pp. 160-169.

De Ruyter, K.O., Wetzels, M. and Feinberg, R. (2001), "Role stress in call centers: its effects on employee performance and satisfaction", *Journal of Interactive Marketing*, Vol. 15 No. 2, pp. 23-35.

Devillers, L. and Vidrascu, L. (2007), "Real-life emotion recognition in speech", *Speaker Classification II*, Springer, pp. 34-42.

Dormann, C. and Zapf, D. (2004), "Customer-related social stressors and burnout", *Journal of Occupational Health Psychology*, Vol. 9 No. 1, pp. 61-82.

Eatough, E.M., Chang, C.H., Miloslavic, S.A. and Johnson, R.E. (2011), "Relationships of role stressors with organizational citizenship behavior: a meta-analysis", *Journal of Applied Psychology*, Vol. 96 No. 3, pp. 619-632.

Ekman, P. and Cordaro, D. (2011), "What is meant by calling emotions basic", *Emotion Review*, Vol. 3 No. 4, pp. 364-370.

Ekman, P. and Friesen, W.V. (1969), "The repertoire of nonverbal behavior: categories, origins, usage, and coding", *Semiotica*, Vol. 1 No. 1, pp. 49-98.

Elliot, A.J. and Thrash, T.M. (2002), "Approach-avoidance motivation in personality: approach and avoidance temperaments and goals", *Journal of Personality and Social Psychology*, Vol. 82 No. 5, pp. 804-818.

Farhadloo, M., Patterson, R.A. and Rolland, E. (2016), "Modeling customer satisfaction from unstructured data using a Bayesian approach", *Decision Support Systems*, Vol. 90 No. 10, pp. 1-11.

Gabriel, A.S. and Diefendorff, J.M. (2015), "Emotional labor dynamics: a momentary approach", *Academy of Management Journal*, Vol. 58 No. 6, pp. 1804-1825.

Gans, N. and Zhou, Y.P. (2007), "Call-routing schemes for call-center outsourcing", *Manufacturing and Service Operations Management*, Vol. 9 No. 1, pp. 33-50.

Ganster, D.C. and Rosen, C.C. (2013), "Work stress and employee health: a multidisciplinary review", *Journal of Management*, Vol. 39 No. 5, pp. 1085-1122.

Giannakopoulos, T. (2015), "Pyaudioanalysis: an open-source python library for audio signal analysis", *PloS One*, Vol. 10 No. 12, pp. 1-17.

Goldberg, L.S. and Grandey, A.A. (2007), "Display rules versus display autonomy: emotion regulation, emotional exhaustion, and task performance in a call center simulation", *Journal of Occupational Health Psychology*, Vol. 12 No. 3, pp. 301-318.

Goussinsky, R. (2012), "Coping with customer aggression", *Journal of Service Management*, Vol. 23 No. 2, pp. 170-196.

Grandey, A. (2003), "When 'the show must go on': surface acting and deep acting as determinants of emotional exhaustion and peer-rated service delivery", *Academy of Management Journal*, Vol. 46 No. 1, pp. 86-96.

Grandey, A., Dickter, D. and Sin, H. (2004), "The customer is not always right: customer aggression and emotion regulation of service employees", *Journal of Organizational Behavior*, Vol. 25 No. 3, pp. 397-418.

Grandey, A., Kern, J.H. and Frone, M.R. (2007), "Verbal abuse from outsiders versus insiders: comparing frequency, impact on emotional exhaustion, and the role of emotional labor", *Journal of Occupational Health Psychology*, Vol. 12 No. 1, pp. 63-79.

Grandey, A.A., Rafaeli, A., Ravid, S., Wirtz, J. and Steiner, D.D. (2010), "Emotion display rules at work in the global service economy: the special case of the customer", *Journal of Service Management*, Vol. 21 No. 3, pp. 388-412.

Graves, A., Jaitly, N. and Mohamed, A. (2013), "Hybrid speech recognition with deep bidirectional LSTM", *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, Olomouc, Czech Republic, December 8-12 2013, IEEE, pp. 273-278.

Guo, L., Wang, L., Dang, J., Zhang, L. and Guan, H. (2018), "A feature fusion method based on extreme learning machine for speech emotion recognition", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB*, Canada, April 15-20, 2018, IEEE, pp. 2666-2670.

Halbesleben, J.R. and Bowler, W.M. (2007), "Emotional exhaustion and job performance: the mediating role of motivation", *Journal of Applied Psychology*, Vol. 92 No. 1, pp. 93-106.

Han, H., Byun, K. and Kang, H.G. (2018), "A deep learning-based stress detection algorithm with speech signal", *Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, AVSU'18*, Seoul, Republic of Korea, ACM, pp. 11-15.

Hareli, S. and Rafaeli, A. (2008), "Emotion cycles: on the social influence of emotion in organizations", *Research in Organizational Behavior*, Vol. 28 No. 1, pp. 35-59.

Henkel, A.P., Bromuri, S., Iren, D. and Urovi, V. (2020), "Half human, half machine – augmenting service employees with AI for interpersonal emotion regulation", *Journal of Service Management*, Vol. 31 No. 2, pp. 247-265.

Hochreiter, S. (1998), "The vanishing gradient problem during learning recurrent neural nets and problem solutions", *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 6 No. 2, pp. 107-116.

Huang, M.H. and Rust, R.T. (2018), "Artificial intelligence in service", *Journal of Service Research*, Vol. 21 No. 2, pp. 155-172.

Huang, K.Y., Wu, C.H., Hong, Q.B., Su, M.H. and Chen, Y.H. (2019), "Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds", *44th International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019*, Brighton, UK, 2019, pp. 5866-5870.

Hudlicka, E. (2003), "To feel or not to feel: the role of affect in human–computer interaction", *International Journal of Human-Computer Studies*, Vol. 59 Nos 1-2, pp. 1-32.

Imai, S. (1983), "Cepstral analysis synthesis on the mel frequency scale", *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 83*, Boston, Massachusetts, USA, April 14-16, 1983, IEEE, pp. 93-96.

Izard, C.E. (1992), "Basic emotions, relations among emotions, and emotion-cognition relations", *Psychology Review*, Vol. 99 No. 3, pp. 561-565.

Jack, R.E., Sun, W., Delis, I., Garrod, O.G. and Schyns, P.G. (2016), "Four not six: revealing culturally common facial expressions of emotion", *Journal of Experimental Psychology: General*, Vol. 145 No. 6, pp. 708-730.

James, E. (1976), "The acquisition of prosodic features of speech using a speech visualizer", *International Review of Applied Linguistics in Language Teaching*, Vol. 14 No. 3, pp. 227-244.

Kleijn, W.B. (1991), "Continuous representations in linear predictive coding", *ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Ontario, Canada, pp. 201-204.

Kraemer, T. and Gouthier, M.H.J. (2014), "How organizational pride and emotional exhaustion explain turnover intentions in call centers: a multi-group analysis with gender and organizational tenure", *Journal of Service Management*, Vol. 25 No. 1, pp. 125-148.

Larivière, B., Bowen, D., Andreassen, T.W., Kunz, W., Sirianni, N.J., Voss, C., Wünderlich, N.V. and De Keyser, A. (2017), "'Service Encounter 2.0': an investigation into the roles of technology, employees and customers", *Journal of Business Research*, Vol. 79, pp. 238-246.

LeCun, Y., Bengio, Y. and Hinton, G. (2015), "Deep learning", *Nature*, Vol. 521, pp. 436-444.

Lehiste, I. and Lass, N.J. (1976), "Suprasegmental features of speech", *Contemporary Issues in Experimental Phonetics*, Vol. 225, p. 239.

Lemon, K.N. and Verhoef, P.C. (2016), "Understanding customer experience throughout the customer journey", *Journal of Marketing*, Vol. 80 No. 6, pp. 69-96.

Lewig, K.A. and Dollard, M.F. (2003), "Emotional dissonance, emotional exhaustion and job satisfaction in call centre workers", *European Journal of Work and Organizational Psychology*, Vol. 12 No. 4, pp. 366-392.

Liu, Z.T., Wu, M., Cao, W.H., Mao, J.W., Xu, J.P. and Tan, G.Z. (2018), "Speech emotion recognition based on feature selection and extreme learning machine decision tree", *Neurocomputing*, Vol. 273, pp. 271-280.

Livingstone, S.R. and Russo, F.A. (2018), "The Ryerson audio-visual database of emotional speech and song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English", *PloS One*, Vol. 13 No. 5, pp. 1-35.

Mao, Q., Dong, M., Huang, Z. and Zhan, Y. (2014), "Learning salient features for speech emotion recognition using convolutional neural networks", *IEEE Transactions on Multimedia*, Vol. 16 No. 8, pp. 2203-2213.

Marinova, D., de Ruyter, K., Huang, M.H., Meuter, M.L. and Challagalla, G. (2017), "Getting smart: learning from technology-empowered frontline interactions", *Journal of Service Research*, Vol. 20 No. 1, pp. 29-42.

Maslach, C. and Leiter, M.P. (2016), "Understanding the burnout experience: recent research and its implications for psychiatry", *World Psychiatry*, Vol. 15 No. 2, pp. 103-111.

Masui, Y. and Wada, S. (2004), "Analysis and synthesis of emotional voice by time-frequency method", *Proceedings of 7th International Conference on Signal Processing, ICSP'04*, Beijing, China, IEEE, Vol. 1, pp. 626-629.

Mun, S., Shon, S., Kim, W. and Ko, H. (2016), "Deep neural network bottleneck features for acoustic event recognition", in Morgan, N. (Ed.) *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, San Francisco, CA, USA, September 8-12, 2016, ISCA, pp. 2954-2957.

Nwe, T.L., Foo, S.W. and De Silva, L.C. (2003), "Speech emotion recognition using hidden Markov models", *Speech Communication*, Vol. 41 No. 4, pp. 603-623.

Oord, A.V.D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K. (2016), "WaveNet: a generative model for raw audio", *The 9th ISCA Speech Synthesis Workshop*, Sunnyvale, CA, USA, 13-15 September 2016, ISCA, pp. 125-139.

Pan, Y., Shen, P. and Shen, L. (2012), "Speech emotion recognition using support vector machine", *International Journal of Smart Home*, Vol. 6 No. 2, pp. 101-108.

Paulmann, S., Furnes, D., Bøkenes, A.M. and Cozzolino, P.J. (2016), "How psychological stress affects emotional prosody", *PloS One*, Vol. 11 No. 11, pp. 1-21.

Pennebaker, J.W., Rimé, B. and Blankenship, V.E. (1996), "Stereotypes of emotional expressiveness of Northerners and Southerners: a cross-cultural test of Montesquieu's hypotheses", *Journal of Personality and Social Psychology*, Vol. 70 No. 2, pp. 372-380.

Pepe, M.S. (2000), "An interpretation for the ROC curve and inference using GLM procedures", *Biometrics*, Vol. 56 No. 2, pp. 352-359.

Pisanski, K., Kobylarek, A., Jakubowska, L., Nowak, J., Walter, A., Błaszczyński, K., Kasprzyk, M., Łysenko, K., Sukiennik, I., Piątek, K. and Frackowiak, T. (2018), "Multimodal stress detection: testing for covariation in vocal, hormonal and physiological responses to Trier Social Stress Test", *Hormones and Behavior*, Vol. 106 No. 1, pp. 52-61.

Plutchik, R. (1982), "A psychoevolutionary theory of emotions", *Social Science Information*, Vol. 21 Nos 4-5, pp. 529-553.

Rafaeli, A., Erez, A., Ravid, S., Derfler-Rozin, R., Treister, D.E. and Scheyer, R. (2012), "When customers exhibit verbal aggression, employees pay cognitive costs", *Journal of Applied Psychology*, Vol. 97 No. 5, pp. 931-950.

Richardson, K.M. and Rothstein, H.R. (2008), "Effects of occupational stress management intervention programs: a meta-analysis", *Journal of Occupational Health Psychology*, Vol. 13 No. 1, pp. 69-93.

Rupp, D.E. and Spencer, S. (2006), "When customers lash out: the effects of customer interactional injustice on emotional labor and the mediating role of discrete emotions", *Journal of Applied Psychology*, Vol. 91 No. 4, pp. 971-978.

Salovey, P. and Mayer, J.D. (1990), "Emotional intelligence", *Imagination, Cognition and Personality*, Vol. 9 No. 3, pp. 185-211.

Satt, A., Rozenberg, S. and Hoory, R. (2017), "Efficient emotion recognition from speech using deep learning on spectrograms", in Lacerda, F. (Ed.) *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, August 20-24, 2017, ISCA, pp. 1089-1093.

Schneiderman, N., Ironson, G. and Siegel, S.D. (2005), "Stress and health: psychological, behavioral, and biological determinants", *Annual Review of Clinical Psychology*, Vol. 1 No. 1, pp. 607-628.

Slavich, G.M., Taylor, S. and Picard, R.W. (2019), "Stress measurement using speech: recent advancements, validation issues, and ethical and privacy considerations", *Stress: The International Journal on the Biology of Stress*, Vol. 22 No. 4, pp. 408-413.

Stanton, J.M., Balzer, W.K., Smith, P.C., Parra, L.F. and Ironson, G. (2001), "A general measure of work stress: the stress in general scale", *Educational and Psychological Measurement*, Vol. 61 No. 5, pp. 866-888.

Subramony, M., Ehrhart, K., Groth, M., Holtom, B.C., van Jaarsveld, D.D., Yagil, D., Darabi, T., Walker, D., Bowen, D.E., Fisk, R.P., Grönroos, C. and Wirtz, J. (2017), "Accelerating employee-related scholarship in service management: research streams, propositions, and commentaries", *Journal of Service Management*, Vol. 28 No. 5, pp. 837-865.

Szaszák, G., Tündik, M.Á. and Gerazov, B. (2018), "Prosodic stress detection for fixed stress languages using formal atom decomposition and a statistical hidden Markov hybrid", *Speech Communication*, Vol. 102, pp. 14-26.

Tomkins, S.S. (1963), "Illuminating and stimulating", *Science*, Vol. 139, pp. 400-401.

Tuten, T.L. and Neidermeyer, P.E. (2004), "Performance, satisfaction and turnover in call centers: the effects of stress and optimism", *Journal of Business Research*, Vol. 57 No. 1, pp. 26-34.

van Jaarsveld, D.D., Walker, D.D., Restubog, S.L.D., Skarlicki, D., Chen, Y. and Frické, P.H. (2019), "Unpacking the relationship between customer (in)justice and employee turnover outcomes: can fair supervisor treatment reduce employees' emotional turmoil?", *Journal of Service Research*, October, doi: 10.1177/1094670519883949.

Vrijkotte, T.G., van Doornen, L.J. and de Geus, E.J. (2000), "Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability", *Hypertension*, Vol. 35 No. 4, pp. 880-886.

Wang, J. (2005), "Work stress as a risk factor for major depressive episode(s)", *Psychological Medicine*, Vol. 35 No. 6, pp. 865-871.

Wilson, H.J. and Daugherty, P.R. (2018), "Collaborative intelligence: humans and AI are joining forces", *Harvard Business Review*, Vol. 96 No. 4, pp. 114-123.

Wirtz, J., Patterson, P.G., Kunz, W.H., Gruber, T., Lu, V.N., Paluch, S. and Martins, A. (2018), "Brave new world: service robots in the frontline", *Journal of Service Management*, Vol. 29 No. 5, pp. 907-931.

Wright, T.A. and Cropanzano, R. (1998), "Emotional exhaustion as a predictor of job performance and voluntary turnover", *Journal of Applied Psychology*, Vol. 83 No. 3, pp. 486-493.

Wu, S., Falk, T.H. and Chan, W.Y. (2011), "Automatic speech emotion recognition using modulation spectral features", *Speech Communication*, Vol. 53 No. 5, pp. 768-785.

Xia, X., Togneri, R., Sohel, F. and Huang, D. (2018), "Random forest classification based acoustic event detection utilizing contextual-information and bottleneck features", *Pattern Recognition*, Vol. 81, pp. 1-13.

Zapf, D., Seifert, C., Schmutte, B., Mertini, H. and Holz, M. (2001), "Emotion work and job stressors and their effects on burnout", *Psychology and Health*, Vol. 16 No. 5, pp. 527-545.

Zhao, Z., Zhao, Y., Bao, Z., Wang, H., Zhang, Z. and Li, C. (2018), "Deep spectrum feature representations for speech emotion recognition", *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data, ASMMC-MMAC'18*, Seoul, Republic of Korea, ACM, pp. 27-33.

Zhao, J., Mao, X. and Chen, L. (2019), "Speech emotion recognition using deep 1D & 2D CNN LSTM networks", *Biomedical Signal Processing and Control*, Vol. 47, pp. 312-323.

Further reading

Barron, A.R. (1993), “Universal approximation bounds for superpositions of a sigmoidal function”, *IEEE Transactions on Information Theory*, Vol. 39 No. 3, pp. 930-945.

Mozer, M.C. (1989), “A focused backpropagation algorithm for temporal pattern recognition”, *Complex Systems*, Vol. 3 No. 4, pp. 349-381.

Ton-That, A.H. and Cao, N.T. (2019), “Speech emotion recognition using a fuzzy approach”, *Journal of Intelligent and Fuzzy Systems*, Vol. 36 No. 2, pp. 1587-1597.

Appendix

The Appendixes files are available online for this article.

Corresponding author

Stefano Bromuri can be contacted at: stefano.bromuri@ou.nl