<table>
<thead>
<tr>
<th><strong>Project reference no.</strong></th>
<th>H2020 644632</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project working name</strong></td>
<td>MixedEmotions</td>
</tr>
<tr>
<td><strong>Project full name</strong></td>
<td>Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets</td>
</tr>
<tr>
<td><strong>Security (distribution level)</strong></td>
<td>PU</td>
</tr>
<tr>
<td><strong>Contractual delivery date</strong></td>
<td>30 September 2016</td>
</tr>
<tr>
<td><strong>Deliverable number</strong></td>
<td>D4.2</td>
</tr>
<tr>
<td><strong>Deliverable name</strong></td>
<td>Emotion Recognition from Multilingual Content, final version</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>Report</td>
</tr>
<tr>
<td><strong>Version</strong></td>
<td>Final</td>
</tr>
<tr>
<td><strong>WP / Task responsible</strong></td>
<td>WP4</td>
</tr>
<tr>
<td><strong>Contributors</strong></td>
<td>Uni Passau (Hesam Sagha, Björn Schuller), NUIG (Cécile Robin)</td>
</tr>
<tr>
<td><strong>Project officer</strong></td>
<td><a href="mailto:Martina.EYDNER@ec.europa.eu">Martina.EYDNER@ec.europa.eu</a></td>
</tr>
</tbody>
</table>
CONTENTS
Executive Summary ........................................................................................................................ 5
1. Introduction ............................................................................................................................. 5
2. Speech Analysis in Big Data area ............................................................................................ 6
   2.1. Introduction ..................................................................................................................... 6
   2.2. Data: The Availability-Shock .......................................................................................... 8
   2.3. On Efficiency: Learning Cooperatively .......................................................................... 8
      2.3.1. Transfer Learning .................................................................................................... 8
      2.3.2. (Dynamic) Active Learning ..................................................................................... 8
      2.3.3. Semi-Supervised Learning ...................................................................................... 9
      2.3.4. Cooperative Learning .............................................................................................. 9
   2.4. On Decision-Making: Learning Confidence Measures ................................................... 9
      2.4.1. Agreement-based Confidence Measures ............................................................... 10
      2.4.2. Learning Errors ...................................................................................................... 10
   2.5. On Seeing the Larger Picture: Learning Multiple Targets ............................................ 10
   2.6. On Big Data: Distribution ............................................................................................. 11
   2.7. Conclusion ..................................................................................................................... 11
3. Feature Enhancement by Autoencoder with LSTM Neural Networks .................................. 12
   3.1. Introduction ................................................................................................................... 12
   3.2. Feature enhancement ..................................................................................................... 13
      3.2.1. Memory-enhanced recurrent neural networks ....................................................... 13
   3.3. Feature enhancement by an autoencoder ....................................................................... 14
   3.4. Experimental and results ............................................................................................... 15
      3.4.1. RECOLA and noise database ................................................................................ 15
      3.4.2. Experimental setups ............................................................................................... 16
      3.4.3. Results and Discussion .......................................................................................... 17
   3.5. Conclusions ................................................................................................................... 19
4. Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio ........................................................................................................ 19
4.1. Introduction ................................................................................................................... 19
4.2. Related Work ................................................................................................................. 20
4.3. Discriminative objectives for emotion regression ......................................................... 21
4.4. Training algorithm ......................................................................................................... 23
4.5. Experiments and Results ............................................................................................... 24
  4.5.1. Emotions from music: MediaEval ......................................................................... 24
  4.5.2. Emotions from speech: RECOLA ......................................................................... 25
  4.5.3. Results ................................................................................................................... 26
4.6. Conclusions ................................................................................................................... 27
5. End-to-end Speech Emotion Recognition Using a Deep Convolutional Recurrent Network 28
  5.1. Introduction and prior work ........................................................................................... 28
    5.1.1. Related Work ......................................................................................................... 28
    5.1.2. Contribution of this work ...................................................................................... 29
  5.2. Model design ................................................................................................................. 29
    5.2.1. Topology of the network ....................................................................................... 30
    5.2.2. Objective function ................................................................................................. 30
  5.3. Experiments and dataset .............................................................................................. 31
  5.4. Relation to existing acoustic and prosodic features ....................................................... 32
  5.5. Conclusions ................................................................................................................... 33
6. Final Conclusion ................................................................................................................ 33
Executive Summary

This deliverable (D4.2) is an extension to the initial version (D4.1) of emotion recognition from audio contents. In the initial version (D4.1) we investigated the multilingualism aspect of emotion recognition from audio and we found that choosing a trained model of the same language or language family increases the emotion recognition performance. Additionally, we proposed a method based on Canonical Correlation Analysis to reduce the dissimilarity between the corpora of different languages.

In this deliverable, first we describe the necessary components which are needed for audio analysis in Big-Data analysis. Then, we represent the state-of-the-art methods developed since last deliverable (D4.1) on i) deep feature enhancement for improving emotion recognition, ii) modification of objective function of neural network for learning continuous emotion labels, and iii) employment of convolutional neural networks for end-to-end (or featureless) speech emotion recognition.

Most of the analysis in these deliverable are performed on the RECOLA database\(^1\) [1] which has been introduced in the D4.1 Section 3 and as a performance measure we used Concordance Correlation Coefficient (CCC) which is a suitable measure for continuous labels.

1. Introduction

In spoken language analysis tasks, one is often faced with small available corpora of only one up to a few hours of speech material mostly annotated on a single aspect, such as a particular speaker’s state at a time. In stark contrast to this, engines such as for the recognition of speaker's emotions, sentiment, personality, or pathologies, are often expected to run regardless of the speaker, the spoken content, and the acoustic conditions. This lack of large and richly annotated material explains to a large degree the room for improvement left in terms of accuracy in today's engines. Yet, in the Big Data era, and with the increasing availability of crowdsourcing services, as well as recent advances in weakly supervised learning, new opportunities arise to ease this fact. In this light, Section 2 first shows the de-facto standard in terms of data-availability in a broad range of speaker analysis tasks. It then introduces highly efficient ‘cooperative’ learning strategies, based on the combination of active and semi-supervised alongside transfer learning in order to best exploit available data in combination with data synthesis. Further, approaches to estimate meaningful confidence measures in this domain are suggested, as they form (part of) the basis of the weakly supervised learning algorithms. In addition, first successful approaches towards holistic speech analysis are presented in this section using deep recurrent rich multi-target learning with partially missing label information. Finally, steps towards needed distribution of processing for big data handling are demonstrated.

During the last decade, speech emotion recognition technology has matured well enough to be used in some real-life scenarios. However, these scenarios require an almost silent environment not to compromise the performance of the system. Emotion recognition technology from speech thus needs to evolve and face more challenging conditions, such as environmental additive and convolutional noises, in order to broaden its

\(^1\) http://diuf.unifr.ch/diva/recola/
applicability to real-life conditions. Section 3 evaluates the impact of a front-end feature enhancement method based on an autoencoder with long short-term memory neural networks, for robust emotion recognition from speech. Support Vector Regression is then used as a back-end for time- and value-continuous emotion prediction from enhanced features. We perform extensive evaluations on both non-stationary additive noise and convolutional noise, on a database of spontaneous and natural emotions. Results show that the proposed method significantly outperforms a system trained on raw features, for both arousal and valence dimensions, while having almost no degradation when applied to clean speech.

Furthermore, continuous dimensional emotion recognition from audio is a sequential regression problem, where the goal is to maximize correlation between sequences of regression outputs and continuous-valued emotion contours, while minimizing the average deviation. As in other domains, deep neural networks trained on simple acoustic features achieve good performance on this task. Yet, the usual squared error objective functions for neural network training do not fully take into account the above-named goal. Hence, in Section 4 we introduce a technique for the discriminative training of deep neural networks using the concordance correlation coefficient as cost function, which unites both correlation and mean squared error in a single differentiable function. Results on the MediaEval 2013 and AV+EC 2015 Challenge data sets show that the proposed method can significantly improve the evaluation criteria compared to standard mean squared error training, both in the music and speech domains.

Finally, the automatic recognition of spontaneous emotions from speech is a challenging task. On the one hand, acoustic features need to be robust enough to capture the emotional content for various styles of speaking, and on the other hand, machine learning algorithms need to be insensitive to outliers while being able to model the context. Whereas the latter has been tackled by the use of Long Short-Term Memory (LSTM) networks, the former is still under very active investigations, even though more than a decade of research has provided a large set of acoustic descriptors. In Section 5, we propose a solution to the problem of ‘context-aware’ emotional relevant feature extraction, by combining Convolutional Neural Networks (CNNs) with LSTM networks, in order to automatically learn the best representation of the speech signal directly from the raw time representation. In this novel work on the so-called end-to-end speech emotion recognition, we show that the use of the proposed topology significantly outperforms the traditional approaches based on signal processing techniques for the prediction of spontaneous and natural emotions on the RECOLA database (D4.1, Section 3). In Section 6 we draw conclusions and our approach toward incorporating the methods into the MixedEmotions platform.

2. Speech Analysis in Big Data area

2.1. Introduction

Speech recognition has more and more found its way into our everyday lives – be it when searching on small handheld devices, controlling home-entertainment or entering, e.g., an address into a navigation system. However, many other speech analysis tasks are not at that point yet – in particular the ‘paralinguistic’ ones. The most frequently encountered usage in day-to-day life of the latter is the identification of the speaker per
se, such as in some telephone banking settings. Then comes, often underlying, gender and age-group, e.g., in dialogue systems or simply to adapt the speech recogniser. A few applications, e.g., in video games such “Truth or Lies”, promise to recognise deceptive speech or emotion. However, the wide potential of speech analysis tasks such as recognition of a speaker’s personality, physical and mental load, health condition, eating condition, degree of nativeness, intoxication, or sleepiness has not been able yet to found its way into applications noticed by the general public. While certainly of great help if running properly, still they lack of high reliability. However this is subject to variations as it is a matter of diverse factors such as the right pre-processing including de-noising and de-reverberation, optimal feature representation, optimal classification or regression and optimisation of models. The main bottleneck can likely be attributed to the sparseness of learning data for such systems. In comparison, a speech recogniser is partially being trained on more data than a human is exposed to throughout lifetime. For computational paralinguistic tasks however, data often remains at the level of one up to a few hours and a handful to some hundred speakers. This data material is mostly annotated with a single phenomenon such as a particular speaker state at a time. In stark contrast to this, engines such as for the recognition of speaker's emotions, sentiment, personality, or pathologies, are often expected to run regardless of the speaker, the spoken content, and the acoustic conditions. While one may argue that a human might not need as much data to learn certain paralinguistic characteristics as for learning a whole language, clearly, more data are needed than the one available now – also because one may wish to aim at surpassing humans in some tasks. Three factors are mainly responsible for this sparseness of speech data and their labels: the data are often 1) sparse in itself, such as in the case of a sparsely occurring speaker state or trait, 2) considerably more ambiguous and thus challenging to annotate than, e.g., orthographic transcription of speech usually is, and 3) of highly private nature such as highly emotional or speech disorders. Yet, in the Big Data era, the lack of speech data is less and less a reality, as we have now access to ‘big’ amount of data through a large number and a variety of resources, such as the internet, broadcast, voice communication, and increased usage of speech-services including self-monitoring. Instead, what is missing are rather the labels. Luckily, with the increasing availability of crowdsourcing services, and recent advances in weakly supervised, contextual, and reinforced learning, new opportunities arise to ease this fact.

In this light, Section 2.2 first shows the current standard in terms of data-availability in a broader range of speaker analysis tasks. It then presents highly efficient ‘cooperative’ learning strategies based on the combination of active and semi-supervised, as well as transfer learning to best exploit available data (Section 2.3). Further approaches to estimate meaningful confidence measures in this domain are then suggested, as they form (part of) the basis of the weakly supervised learning algorithms (Section 2.4). In addition, first successful approaches towards holistic speech analysis are presented using deep recurrent rich multi-target learning with partially missing label information (Section 2.5). Then, steps towards needed distribution of processing for big data handling are demonstrated (Section 2.6). Finally, some remaining aspects are discussed and conclusions are drawn (Section 2.7). Overall, we discuss here a system architecture and methodology that holds the promise of a major breakthrough in performance and ability of generalization in tomorrow’s speech analysis systems.
2.2. Data: The Availability-Shock

Few speech analysis tasks are lucky enough to have a full day-length of labelled speech material available for training and testing of models. Taking the Interspeech 2009–2015 series of Computational Paralinguistics Challenges as a reference [2], one can see that, in fact, mostly around one or ‘some’ hours is all that is given as a starting point to train a model for a new speech analysis task, such as recognising Alzheimer, Autism, or Parkinson’s Condition of a speaker. Obviously, one can hardly expect to train models that are able to run regardless of the speaker, language, cultural background, and co-influencing factors from such little data. Indeed, some attempts at cross-corpus studies show the very weak generalisation observed for most systems trained in such a way (e.g., [3]).

2.3. On Efficiency: Learning Cooperatively

This reality of little labelled speech data, but availability of large(r) amounts of unlabelled one has led to a number of recent approaches that try to most efficiently exploit both of these with little human labour involved.

2.3.1. Transfer Learning

In the past, work labelling of data for a defined and specific domain or task has been done, such as recognising emotion of adult speakers, but little to no (labelled) data for our current interest – let’s say recognising emotion of children. The solution here would be to train a model that best learns how to ‘transfer’ the knowledge from the previous to the new domain, even if no labelling has been done in corpus of the new target domain [4], [5] (D4.1. Section 3). Some interesting experiments have shown that the knowledge acquired through the training of a model for the recognition of emotion in speech could be transferred to the music domain. In [6], this was reached by the use of a sparse autoencoder that learns a compact representation of one of the domains (either speech or music) and ‘transfers’ features to the respective other one. In [7] a more efficient approach has been tested through the training of several autoencoders and the learning of the differences with an additional neural network. Further, usage of related data for the initialisation of models such as in deep learning has been shown useful in general speech processing, e.g. in [8], but is less exploited in paralinguistics as for now.

2.3.2. (Dynamic) Active Learning

Better models can usually nonetheless be reached if one does label at least some data in the new domain or for the new task. To keep human efforts to a minimum, the computer can first decide which data points are of interest, for example by identifying sparse instances such as emotional data (leaving the human to decide which emotion it is) versus non-emotional data (which usually appears in much higher frequency and is thus less interesting after some such data points have already been seen) ([9], [10]). Accordingly, rather than recognising different emotions where data for each class may be sparse, a coarser model can be first chosen, and simply discriminates neutral versus non-neutral speech. This way, one can initialise an active learning
system by collecting only emotionally neutral speech and then execute a novelty detection or alike for unlabelled data. As neutral emotional speech is available in large amounts or can even be synthesised [11], this ‘one class’ model can easily be trained with such amount of data. Then, when some new speech is deviant in some form, a human can be asked for labelling help. Other aspects can include the likelihood of change of model parameters, i.e., the learning algorithm decides if the data would allow to change its parameters significantly at all before asking for human aid on ‘what it is’. Such approaches were also extended for actively learning regression tasks rather than discrete classes [12].

An interesting more recent option for fast labelling is crowdsourcing [13], as it offers to quickly reach a large amount of annotators (in fact often even in real-time, which may become necessary when dealing with ‘big’ and growing amounts of data). However, laymen most often form the majority of the crowd, rather than experts in phonetics, linguistics, psychology, medicine or other related disciplines that may be of relevance to the speech analysis task. A much higher number of annotators is thus often needed and the resulting corpus has to cope with noisy labels. ‘Learning’ the annotators’ behaviour and dynamically deciding on how many annotators and ‘whom to ask when’ allows to select the crowd more efficiently [14].

### 2.3.3. Semi-Supervised Learning

More efficiently, the computer can label data itself once it was trained on some first training data [15]. Obviously, this brings a risk of erroneously labelling data and then re-training the system on partially noisy labels. Accordingly, one usually needs to make a decision based on some confidence measure (cf. below) on whether to add a computer-labelled data instance to the learning material for (re-)training or not. In addition, one can use multiple ‘modalities’ in ‘co-training’ to decide on the labels of the data [16], [17]. In [18] it was shown for a range of speech analysis tasks that this following this method makes it possible to have a self-improved speech analysis system by feeding it with new (unlabelled) speech data observations.

### 2.3.4. Cooperative Learning

Associating the above two (i.e., active and semi-supervised learning) methods together leads to ‘cooperative learning’ [19]. The principle can be described as follows: For a new data instance, the computer first decide if it can label it by itself – if not, make a decision if it is worth or not to ask for human aid. [19] shows that this can be more efficient than any of the above to forms.

### 2.4. On Decision-Making: Learning Confidence Measures

As both, active, and semi-supervised learning mostly base on some confidence measure, it seems crucial to find ways of reliably estimating such. In stark contrast to speech recognition [20], there does not yet exist much literature on this topic for the field of paralinguistic speech analysis. The current approaches use the inherent confidence of a learning algorithm, such as the distance to the separating hyperplane of the winning class as compared to the next best class. However, as it is, this learning algorithm that makes the decision on a class is often wrong. It seems thus more reliable to explore additional ways of measuring the confidence of
the result of recognition. Two approaches have been tested recently in the field of Computational Paralinguistics, partially exploiting the characteristics of this field.

2.4.1. Agreement-based Confidence Measures

The first approach aims at estimating the agreement humans are likely to have in the review of the paralinguistic targeted features [21]. Thus, for a subjective task requiring several annotation classes such as emotion of a speaker, one does not train the emotion class or degree of likability as target, but rather the percentage of human raters that agreed upon the label. Then, this percentage can automatically be estimated for new speech data, and serves as a measure on how difficult is the task of assessing a unique agreed label/opinion on this new data. Thereby this can be interpreted as an indirect measure of confidence.

2.4.2. Learning Errors

Alternatively, it is possible to train additional recognition engines along with the ‘in charge’ paralinguistic engine, choosing whether or not selecting errors as learning target. If several such engines are trained on different data, their estimates can be used as confidence measure. In fact this measure’s reliability can even be improved by semi-supervised learning [22].

2.5. On Seeing the Larger Picture: Learning Multiple Targets

As all our personal speaker traits as well as our mental and physical state have an impact on the mechanism of production of speech, it seems wise to attempt to see the ‘larger picture’. Up to now, most works in the field of Computational Paralinguistics are dealing with one phenomenon at a time, focusing for example on the recognition of emotion only or concentrating exclusively on the health state feature or else the personality. However, a variation in voice is not shaped by one factor only. For example one could be angry at some point, but could also have a flu in parallel and an introverted personality. All of those characteristics are felt through the voice. Most attempts to learn multiple traits in parallel such that the knowledge of each other aspect positively influences the overall recognition accuracy have so far been focused on learning several emotion primitives all together, cf., e.g., [23]. However, more recent work heads for learning a richer variety of human’s states and traits as multiple targets [24]. This introduces the challenge to find data that are labelled in such manifold states and traits – something hardly available these days. One can easily imagine how this raises the demand in the above described efficient ways of quickly labelling data by the aid of the crowd in intelligent ways. In addition, it requires learning algorithms that are not only able to learn with multiple targets, but also with an unbalanced quantity (even missing) of those distinctive features, given that one will not always be able to obtain a broad range of attributes for any voice sample ‘found’ or newly observed by the computer.
2.6. On Big Data: Distribution

Referring to ‘Big Data’ usually implies that the amount of data is so large that ‘conventional’ approaches of processing cannot be applied [25]. This may require partitioning of the data and distribution of efforts [26]. While a vast body of literature exists in the field of ‘core’ Machine Learning on how to best distribute processing, it still remains to tailor these approaches to the needs of speech analysis. Distributed processing has been targeted considerably to speech recognition, but hardly to paralinguistic tasks where only very first experiences are reported, e.g., on optimal compression of feature vectors [27].

2.7. Conclusion

The next major leap towards the field of Computational Paralinguistics and the broader field of Speech Analysis is likely expected to be achieved by overcoming the ever-present sparseness of learning data by making efficient use of the big amounts of available speech, through adding a big amount of enriched labelled data, likely with help from crowdsourcing. Such resources will have a partially noisy gold standard thus requiring potentially larger amounts of labelled speech than if labelled by experts, but it will be easier to reach large amounts of data. In particular, a wide range of information may all be labelled at once rather than targeting a single phenomenon such as emotion or sleepiness of a speaker per task. The labelling effort will very likely become manageable by the pre-processing of a machine which makes first decisions on the pertinence of the data, and also learns how many and which evaluator to ask in which situation. It is thus not only learning about the pertinent phenomena to target but also about the people who will best help in the relevant task. One can probably best depict this by the metaphor of a child that not only learns about its world, but also whom to best ask about each part of it and thus sometimes better collecting several opinions. When a task does not require to know the details of those aspects of the speech data, then it is still possible to make further use of it, e.g., in a spoken dialogue system without attaching a human-interpretable label, in unsupervised learning, or also clustering may be another suited variant ([28], [29]) for exploiting big amount of speech data.

On the technical side, efficient big data speech analysis will require reliable confidence measure estimation to decide which data can be labelled by the computer and which needs to be reviewed by humans. Furthermore, distributed processing may become necessary if data becomes ‘too’ large. The handling of big data processing comes with new concerns such as how to share the data and trained models, dealing with ethical aspects like privacy, transparency, and responsibility for what has been learnt by the machine once decisions are made ([30], [31]).

Other concerns are brought by the ‘big’ aspect. Some speech analysis tasks will remain sparse in terms of data, e.g., for some pathological speech analysis tasks. In this case, zero resource [32] or sparse resource approaches are an alternative to circumvent the data sparseness. Such approaches are known from speech recognition, keyword spotting and spoken term detection [33]. The usual application scenario is to recognise words in languages where only very sparse speech resources exist. For the recognition of paralinguistic tasks, implementing rules is only possible once some information is known about the phenomenon of interest, for
example to be able to make statements like “IF the speech is faster and the pitch is higher THEN the speaker is more aroused” etc.

As a final statement, one can easily imagine that these conclusions may hold in similar ways to a broader range of audio analysis tasks and, consequently, many other fields – the era of big data and increasingly autonomous machines has room for much more exploration.

3. Feature Enhancement by Autoencoder with LSTM Neural Networks

3.1. Introduction

Technology for automatic emotion recognition from speech (ERS) has gained increasing commercial attention in the last decade. Rapid progress of this technology has indeed enabled application of ERS in various domains, such as, health care [34], education [35], serious games [36], robotics [37], and call-centers [38]. However, while good performance has been reported in research papers under laboratory conditions [39], or with systems tailored towards specific databases [40], real-life applications of ERS still remain an open challenge. Indeed, various factors make this task highly challenging. They can be grouped into three main categories: (i) the contextual dependencies of the meaning and significance of affective expressions across different speakers, languages and cultures [41], (ii) the presence of varying and degraded acoustic conditions caused by reverberation, background noise, and acoustic properties of the recording devices used, and (iii) the necessity to use distributed systems in a client-server architecture, which introduce some latency and distortion in the data [27]. Stationary, non-stationary, and convolutional noise severely degrade performance of systems, and affect consequently the user experience in real-life conditions [42]–[44]. Therefore, many studies have been performed for speech and acoustic feature enhancement (FE), especially for automatic speech recognition (ASR). Recurrent Neural Networks (RNN) are widely used in this field to enhance corrupted features, which is an application of the de-noising autoencoder [45] principle: neural networks are trained to map noisy features to clean features. This method has recently also been exploited for speech enhancement in the time domain [46], [47]. RNN have been also studied for blind non-linear source separation, with the aim to enhance the acoustic features by separating noise and speech sources [48], [49]. In the context of speech enhancement, the authors in [47] use deep neural networks to map noisy to clean Mel features, but the network output is synthesised directly into a time domain signal, instead of constructing a filter based on speech and noise magnitudes. A combination of unsupervised noise estimation and Deep Neural Network (DNN) based speech power spectrum estimation is used in [46] to construct a Wiener filter. Supervised training of deep neural networks was performed to predict the ideal ratio mask in an uncertainty decoding framework for ASR [50]. Studies on noise robustness for ERS are much more sparse, despite being necessary for real-life applications of this technology. To the best of our knowledge, only a few studies have addressed this issue so far. Large acoustic feature sets were investigated in [42]. Adaptive noise cancellation was proposed as a front end in [43]. Speech enhancement based on spectral subtraction and masking properties was studied in [44]. Wavelet decomposition [51] and feature selection techniques [52] have also
been proposed. Additionally, supervised Nonnegative Matrix Factorization (NMF) was investigated for the robustness of emotion recognition engines [53]. One may note that most existing work on noise robustness for ERS has been performed on acted emotions, which are rarely observable in real-life. Furthermore, only a few of those studies have analysed the impact of reverberated noise, which is known to impact severely the performance of ASR systems [53], [54]. In this light, this section studies the impact of non-stationary additive noise and convolutional noise on the automatic recognition of spontaneous emotions from speech. We propose the use of a FE method based on a memory-enhanced recurrent Denoising Autoencoder (rDA) as a front end, and show that this method can significantly improve the performance, while having almost no degradation when applied to clean speech. The following sections are structured as follows: the proposed FE method based on rDA is introduced in Section 3.2, then extensive experiments on spontaneous emotions are described in Section 3.3, and a conclusion with future work is given in Section 3.4.

3.2. Feature enhancement

In the past few years, Long Short-Term Memory (LSTM) models have been widely applied to a variety of pattern recognition tasks [55], [56], and show a powerful capability of learning long-range contextual information. In this section, we give a quick overview of such a memory-enhanced RNN, on which the proposed rDA is built.

3.2.1. Memory-enhanced recurrent neural networks

Compared to conventional RNN, the LSTM-RNN model proposed by Hochreiter and Schmidhuber [57] uses one or multiple LSTM blocks to replace hidden neurons. Every memory block consists of self-connected linear memory cells $c$ and three multiplicative gate units: an input gate $i$, a forget gate $f$, and an output gate $o$, which are responsible for writing, reading, and resetting the memory cell values, respectively. Given an input $x_t$ at the step time $t$, the activations of the input gate $i_t$, the forget gate $f_t$, the memory cell state $c_t$, and the output gate $o_t$ are separately updated by the following formulas:

$$i_t = f_g(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$
$$f_t = f_o(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$
$$c_t = f_g(W_{xc}x_t + W_{hc}h_{t-1} + b_c) + f_t c_{t-1}$$
$$o_t = f_g(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$
$$h_t = o_t f_o(c_t)$$

where $f_g$, $f_i$, and $f_o$ denote the logistic sigmoid, tanh, and tanh activation functions, respectively; $W$ is a weight matrix of the mutual connections; $h_t$ presents the output of the hidden block; $b$ indicates the block bias. From the equations mentioned above, it is observed that the values of all memory cells and block outputs in the previous time step $t - 1$ will certainly affect the activations of all three gates, even the input units in the current time step $t$ in the same layer, except for the case between memory cell and output gate. More details about the memory structure can be found in [58]
The main advantage of using such a memory-enhanced block over a traditional neuron in a RNN is that the cell state in a LSTM block sums activations over time. Since derivatives distribute over sums, the back-propagated error does not blow up or decay over time (the vanishing gradient problem) [57], [58]. The general structure of rDA with memory-enhanced neural networks proposed here is illustrated in Fig. 4.1, which includes an input layer, an output layer, and one or multiple hidden layers that are implemented by the LSTM blocks. In comparison with the conventional DA given in [45] where the DA is modelled with Feedforward Neural Networks (FNN), the presented DA is structured with the above described LSTM-RNN in the hidden layers. Additionally, it also should be noticed that the recurrent autoencoder also differs from the ones described in [56], [59], where an encoder is used to map an input sequence into a fixed length representation, and a decoder is used to decode the target sequence from the representation.

Figure 4.1: Structure of a recurrent denoising autoencoder with LSTM neural networks.

### 3.3. Feature enhancement by an autoencoder

As discussed in Section 3.1, the speech signal \( s(k) \) is easily distorted by the environmental noise and recording devices when facing realistic application scenarios with the Acoustic Impulse Response (AIR) \( r(k) \) of finite length \( T_{60} \) and the background additional noise \( n(k) \). Therefore, the distorted speech signal \( s(k) \) can be expressed as

\[
\hat{s}(k) = s(k) * r(k) + n(k)
\]  

(6)

The signal in the time domain \( \hat{s}(k) \) can be approximately transformed into the spectrum domain as

\[
|\hat{S}(f)|^2 \approx |S(f)|^2 |R(f)|^2 + |N(f)|^2
\]  

(7)

by applying a Short-Time Discrete Fourier Transform (STDFT) with three assumptions: 1) \( T_{60} \) is shorter than the analysis window size \( w \); 2) The power spectrum of the additive noise in each analysis window \( w \) is a slowly varying process, which means that the additive noise is assumed to be stationary in each analysis window; 3) The phase of different analysis windows are non-correlated. To extract the feature vectors in the cepstral domain such as Mel-Frequency Cepstrum Coefficients (MFCC) for emotion recognition from speech,
logarithms and Discrete Cosine Transform (DCT) are performed over the above spectrum. Therefore, Eq. (7) can be further formulated into

$$\mathcal{D}(ln|\hat{S}(f)|^2) \approx \mathcal{D}(ln|S(f)|^2) + \mathcal{D}(ln|R(f)|^2) + \left(ln \left(1 + \frac{|N(f)|^2}{|S(f)|^2 \cdot |R(f)|^2}\right)\right)$$

(8)

From Eq. (8), we can see that the goal of denoising is to eliminate the impact of the last two terms. For the non-stationary noise, however, the cepstrum does not only fluctuate over time, but is also involved with the original speech spectrum which is non-stationary as well. Therefore, the last term in Eq. (8) cannot be simply subtracted due to its non-linear property.

To tackle this non-linear problem, we choose the memory-enhanced rDA as described in Section 3.2.1 with the purpose of exploiting its advantage of accessing long-range contextual information. The goal of the DA is to reconstruct the features $x^c$ in the clean speech feature domain $\mathcal{X}^c$ from the corresponding features $x^n$ in the corrupted speech feature domain $\mathcal{X}^n$, as shown in Fig. 4.1. When providing these corrupted features as the input $x^n$ to the first layer, we want the output $\hat{x}^n$ to be highly similar to the clean features $x^c$. To learn the required mapping between noisy and clean features, an objective function – Mean Squared Error (MSE) – is defined to minimise the reconstruction error during training:

$$\mathcal{J}(\theta) = \frac{1}{T} \sum_{t=1}^{T} (\hat{x}^n_t - x^c_t)^2$$

(9)

where $T$ is the number of frames of the training set.

3.4. Experimental and results

In the following, we firstly describe the selected spontaneous emotion database, then evaluate the performance of the proposed FE method based on the rDA with LSTM neural networks for time- and value-continuous emotion recognition.

3.4.1. RECOLA and noise database

For the experiments, we chose the RECOLA database (D4.7. Section 3) which was used as a database for the 5th Audio/Visual+ Emotion Challenge (AV+EC 2015) [60]. The motivation of the database collection was to study the complex phenomena, especially emotion, portrayed by humans during social interactions in daily-life. To generate additive noisy speech, we added the CHiME15 database [61] into the clean (raw or original) speech with various levels of SNR (i.e., 0–12dB at a step of 3dB). This database was used for the 3rd CHiME Challenge [61], and was collected in five different locations, such as booth, bus, cafe, pedestrian area, and street junction. The goal of this database is to simulate emotional speech in different places with various additive background noises. To generate convolution noise, we applied (via convolution) the Microphone Impulse Response (MIR) of the Google Nexus One smartphone to the recordings from RECOLA using the Audio Degradation Toolbox (ADT) [62]. The goal is to simulate reverberant speech being recorded with a smartphone. Moreover, other noises are further simulated via the MIR, by applying the Room Impulse...
Response (RIR) of classroom or grand hall as the second convolutional noise. This aims to simulate the scenarios that someone speaks on the phone in different environments. Note that, when adding the CHiME noise, each noise recording was firstly normalised to 0dB peak energy and concatenated according to the type of noise. Then, the recording was cut into three partitions of the same length for the training, the validation, and the test sets, respectively. Finally, for each recording of RECOLA, we randomly chose an excerpt of the concatenated noise signal from the relevant partition, to ensure both speaker and noise independent partitions.

### 3.4.2. Experimental setups

At the front-end of the emotion recognition system, 13 Low-Level Descriptors (i.e., MFCCs 0–12) were firstly extracted. In detail, the feature vectors of $x^n_t$ and $x^c_t$ were separately extracted from the distorted speech signals and the original clean speech signals at every 10ms using a window size of 25ms. Before training the rDA, the global means and variances were calculated of the noisy and clean speech. Then, standardisation was performed over the network inputs and targets using the means and variances from the corresponding training sets, respectively. For the rDA, both input and output node numbers are equal to the dimension of the feature vector (13 in our case). Two bidirectional LSTM hidden layers were chosen, and each layer consists of 30 memory blocks. During network training, gradient descent was implemented with a learning rate of $10^{-6}$ and a momentum of 0.9. Zero mean Gaussian noise with standard deviation 0.1 was added to the input activations in the training phase such as to improve generalisation. All weights were randomly initialised in the range from -0.1 to 0.1. Note that, all these parameters were optimised on the validation set. Finally, the early stopping strategy was used, i.e., training was stopped when no improvement of the MSE on the validation set has been observed during 20 epochs or the predefined maximum number of training epochs (200 in our case) has been executed. Further, to accelerate the training process, we updated the network weights after running every mini batch of 8 sequences for computation in parallel. The training was performed with CURRENNT toolkit [63]. After the procedure of FE, functionals – mean and variance – were applied over each of the enhanced MFCCs with a window size of 8s at a step of 0.04s, which leads to 26 attributes for each window. These statistical features were then fed into the back-end of the system used for emotion recognition, where L2-regularised L2-loss Support Vector Regression (SVR) implemented in the LIBLINEAR toolbox [64] was used. The complexity value of SVR was optimised by the best performance of the validation set, i.e., $C=5 \cdot 10^{-5}$ for arousal and $C=5 \cdot 10^{-3}$ for valence in our experiments. For the performance evaluation, we choose the Concordance Correlation Coefficient (CCC) [65] (D4.1. Section 3). Compared to Pearson’s Correlation Coefficient (PCC), CCC can estimate not only the linear correlation, but also the difference of the bias between two variables. Formally, CCC is formulated as

$$
ρ_ε = \frac{2ρσ_xσ_y}{σ_x^2 + σ_y^2 + (μ_x - μ_y)^2}
$$

where $ρ$ is the correlation coefficient between the two variables; $μ_x$ and $μ_y$ are the means of the two variables; and $σ_x^2$ and $σ_y^2$ are the corresponding variances. Moreover, it is worth noting that the gold standard ratings for
all the recordings were shifted back in time with a four seconds delay. This is due to compensate the reaction delay of human for continuous emotion annotation [66].

3.4.3. Results and Discussion

To verify the effectiveness of the rDA with LSTM Neural Networks for FE, we separately performed two experiments: 1) on the non-stationary additive noisy speech only, i.e., by adding CHiME15 noise; 2) on the smartphone related convolutional noisy speech, i.e., the speech is distorted by applying MIR only (smartphone), or additionally by applying RIR of the classroom/hall (+classroom/hall), or additionally by adding various levels of CHiME noise (+CH). Apart from that, we carried out two FE strategies: 1) matched FE: Several FE models are trained separately on the data sets with different noise conditions. For example, when testing on clean speech, the same quality of speech is used, i.e., clean speech is employed to train the rDA; 2) mixed FE: One FE model is trained on a data set with mixed noise conditions. Therefore, the distinction between the two FE strategies is based on the noise condition of the training data that can match or not with the one of the testing data. For example, when testing the clean speech, the mixed conditional speech, i.e., all kinds of CHiME noisy speech or the smartphone related noisy speech together with the clean speech, are utilised to train the rDA. Table 4.1 shows the performance of the non-enhanced and enhanced speech (with the matched or mixed FE strategies) evaluated on the emotion recognition model trained on the clean speech for both, arousal and valence regression. In almost all cases, the proposed FE method significantly outperforms the system trained on the non-enhanced noisy speech (baseline). Taking the additive noisy speech (CHiME15) for example, the average CCC over the recordings at different levels of SNRs on the test set is boosted from 0.563 to 0.596 and 0.594, respectively, by matched and mixed FE for arousal, and from 0.176 to 0.223 and 0.199, respectively, by matched and mixed FE for valence. Furthermore, it is expected that the matched FE outperforms the mixed FE, since the matched FE uses different FE models for denoising corresponding noisy data, whereas the mixed FE trains only one FE model for denoising all kinds of noisy data.

Specifically, the performance obtained on the clean speech condition almost does not degrade when executing the matched FE method, but this conclusion is not supported by performing the mixed FE method. This should be mainly due to huge mismatch between the clean speech and the mixed noisy speech. However, this can be easily solved by inserting a noise detector at the front-end to distinguish whether the signal is noisy or clean [67]. If it was clean, the speech signal can directly be fed into the recognition model without any procedures of FE. Moreover, for the convolutional noise of smartphone in the classroom or in the hall, we can see that our proposed method does not work quite well, however, for the convolutional noise of the smartphone with CHiME noise, the proposed method can surprisingly improve the baseline. This may be because LSTM-RNN does not work efficiently for a linear problem, as the $T_{60}$ of the MIR used for these experiments is short and the convolutional noise can be regarded as a constant value in the spectral domain (see Eq. (8)). To further investigate the efficiency of the proposed FE, we calculated the CCC between the enhanced (by matched FE)/non-enhanced speech and the clean speech over the whole test set with the CHiME noise or the smartphone-in-the-hall noise, as illustrated in Fig. 4.2. It can be seen that the enhanced speech could deliver
higher correlation coefficients with the speech, which possibly contributes to the better emotion recognition performance.

Table 4.1: Performance (Concordance Correlation Coefficient [CCC]) of the validation and test sets for the proposed matched and mixed feature enhancement (FE) model on the CHiME15 noisy speech only or on the smartphone related noisy speech, in the evaluation of arousal and valence emotional tasks. class.: classroom; CH: the average CCC over five different ‘smartphone + CHiME’ noisy speeches with 0–12 dB of SNRs at a step of 3 dB. The mark of “/”: no other noise is added onto the smartphone related noisy speech.

<table>
<thead>
<tr>
<th></th>
<th>clean</th>
<th>12 dB</th>
<th>9 dB</th>
<th>6 dB</th>
<th>3 dB</th>
<th>0 dB</th>
<th>mean</th>
<th>clean</th>
<th>/</th>
<th>class.</th>
<th>hall</th>
<th>CH</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>arousal</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>0.736</td>
<td>0.680</td>
<td>0.657</td>
<td>0.626</td>
<td>0.584</td>
<td>0.526</td>
<td>0.635</td>
<td>0.736</td>
<td>0.726</td>
<td>0.629</td>
<td>0.634</td>
<td>0.436</td>
<td>0.545</td>
</tr>
<tr>
<td>matched FE</td>
<td>0.735</td>
<td>0.715</td>
<td>0.710</td>
<td>0.692</td>
<td>0.666</td>
<td>0.627</td>
<td>0.691</td>
<td>0.735</td>
<td>0.723</td>
<td>0.641</td>
<td>0.662</td>
<td>0.675</td>
<td>0.682</td>
</tr>
<tr>
<td>mixed FE</td>
<td>0.693</td>
<td>0.721</td>
<td>0.711</td>
<td>0.691</td>
<td>0.648</td>
<td>0.594</td>
<td>0.676</td>
<td>0.690</td>
<td>0.686</td>
<td>0.650</td>
<td>0.651</td>
<td>0.599</td>
<td>0.630</td>
</tr>
<tr>
<td><strong>valence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>0.732</td>
<td>0.628</td>
<td>0.590</td>
<td>0.542</td>
<td>0.480</td>
<td>0.404</td>
<td>0.563</td>
<td>0.732</td>
<td>0.719</td>
<td>0.618</td>
<td>0.609</td>
<td>0.356</td>
<td>0.495</td>
</tr>
<tr>
<td>matched FE</td>
<td>0.729</td>
<td>0.658</td>
<td>0.646</td>
<td>0.611</td>
<td>0.510</td>
<td>0.422</td>
<td>0.596</td>
<td>0.729</td>
<td>0.713</td>
<td>0.614</td>
<td>0.694</td>
<td>0.682</td>
<td>0.684</td>
</tr>
<tr>
<td>mixed FE</td>
<td>0.717</td>
<td>0.683</td>
<td>0.651</td>
<td>0.598</td>
<td>0.499</td>
<td>0.418</td>
<td>0.594</td>
<td>0.712</td>
<td>0.690</td>
<td>0.612</td>
<td>0.600</td>
<td>0.532</td>
<td>0.586</td>
</tr>
</tbody>
</table>

Figure 4.2: Concordance Correlation Coefficient (CCC) of 13 Low-Level Descriptors (MFCC 0–12) between the enhanced/non-enhanced speech and the clean speech over the whole test set with the CHiME noise (a) or the smartphone-in- the-hall noise (b).
3.5. Conclusions

We presented a feature enhancement method based on a denoising autoencoder with Long Short-Term Memory neural networks for spontaneous emotion recognition from speech. Extensive experiments were carried out with non-stationary additive noise and convolutional noise. The results show that, the presented feature enhancement method is significantly superior to the baseline without any enhancement methods. With the fast development of deep learning technologies, there are many possibilities that could be used to further improve the robustness performance of emotion recognition systems from speech. For example, Convolutional Neural Networks are good at reducing spectral variation for the clean speech, which could also be effective for noisy speech. Methods combined with Deep Neural Networks in an end-to-end structure [68] is worth evaluating as well in future. Further, some other traditional denoising approaches, e.g., minimum mean square error [69], may be also of interest in this task.

4. Discriminatively Trained Recurrent Neural Networks for Continuous Dimensional Emotion Recognition from Audio

4.1. Introduction

Continuous dimensional emotion recognition from audio is a sequential learning problem that has attracted increasing attention in the past few years [70]–[74]. There, sequences of acoustic features have to be mapped to emotion contours in several dimensions that represent the emotion communicated by means of audio, e.g., speech utterances or excerpts of music. Typical emotion dimensions comprise arousal and valence [75], as explored in this study, although other dimensions such as dominance and expectation can be added [23]. Defining the target labels as real-valued mappings from time instants to targets helps capturing the temporal dynamics of emotion, which cannot be assumed to be constant over time [76]. To learn such mappings, deep recurrent neural networks are a promising model [70], as they take into account temporal dependencies in inputs and outputs and can handle correlated features.

Continuous emotion recognition is typically evaluated in terms of the correlation between the learner’s outputs and the target values (such as by the correlation or determination coefficient), as well as the average deviation of outputs and targets, such as by the mean linear or mean squared error (MLE/MSE) [77], [78]. Since neural networks are usually trained using criteria such as the (root) MSE, this only takes into account the latter while neglecting the former. Further, although it is well known that the minimization of the MSE and the maximization of the (Pearson) correlation coefficient (CC) are equivalent if the outputs and targets are standardized, such standardization cannot be assumed in emotion regression, as the emotional intensity, and hence the mean and variance, is of crucial importance.

Moreover, the CC is insensitive to scaling and shifting, which is problematic for training neural networks with this metric. Imposing a cost function based on CC may lead to an infinite number of local minima with different prediction behaviour. In fact, a neural network trained on CC cannot learn the correct scales and offsets from the target values (‘gold- standard’) because the CC is not sensitive to such variations. We
illustrate those issues on different variants of the same time-series in Figure 5.1. Because CC is insensitive to scaling and shifting, it always provides the same perfect prediction score (CC = 1.00) on different versions (shifted and/or scaled) of the same time-series, although such variations are actually far from a perfect prediction. As a result, the CC is not sufficient as evaluation criterion in practice, and additional measures such as mean squared error need to be taken into account.

From the above considerations, we can conclude that firstly, the usual objective functions for neural network regression do not fully match the evaluation criteria used for continuous dimensional emotion recognition, and secondly, the use of CC as objective function cannot lead to satisfying results. To alleviate these problems, we propose to use the concordance correlation coefficient (CCC) [65] as a differentiable objective function that unites both correlation and mean squared error, and can be thought of as a CC that enforces the correct scale and offset of the outputs. As a result, CCC takes into account the effects of shifting and scaling the prediction when computing the performance (cf. Fig. 5.1), and can thus be used as an objective function to train neural networks for time-continuous prediction tasks.

In the following, we will show that the choice of objective function (sum of CCCs per sequence, overall CCC, or MSE) for network training significantly influences the evaluation outcome on standard corpora for continuous dimensional emotion recognition from music and speech. The remainder of this part is as follows: we describe some related work on task-specific discriminative objective functions for neural network training in Section 4.2, we introduce different discriminative objectives for emotion regression in Section 4.3, and the optimization of CCC by stochastic gradient descent in Section 4.4, we evaluate the performance in dimensional emotion recognition tasks (arousal and valence) for different objective functions on two different corpora (music and speech) in Section 4.5, and provide a conclusion in Section 4.6.

4.2. Related Work

Task-specific discriminative objective functions for neural network training are well known. For example, in training networks for automatic speech recognition, the minimum phoneme error or minimum Bayes risk...
objectives are used in place of the standard cross entropy objective [79], [80]. In [81], it is proposed to optimize the prediction of time-frequency masks for acoustic source separation based on local signal-to-noise ratio rather than MSE. Joint optimization of masking functions and deep recurrent neural networks has been investigated for monaural source separation tasks by [82], with a discriminative criterion designed to enhance the source to interference ratio. Regarding neural network approaches for emotion recognition, most of the existing work has been focused on classification tasks. Deep Neural Network Hidden Markov Models (DNN-HMMs) were investigated with discriminative pretraining and restricted Boltzmann Machine (RBM) based unsupervised pre-training by [83]. Experimental results have shown the superiority of the hybrid DNN-HMMs with discriminative pretraining in comparison with other models, such as GMM-HMMs and MLP-HMMs. Another hybrid architecture combining DNN with an Extreme Learning Machine (ELM) was also successfully utilised for classifying emotion from utterances in [84]. Continuous prediction of dimensional emotion was investigated with a Deep Belief Network in [76]. [23] dealt with multi-task recurrent neural network based speech emotion regression. The winning team of the last edition of the Audio-Visual Emotion recognition Challenge [60] employed bidirectional long short-term memory recurrent neural networks (BLSTM-RNN) to perform unimodal emotion recognition and multimodal fusion [66]. While these works use a similar learning framework as in our paper, none of them uses the discriminative objective based on correlation coefficients as introduced below.

4.3. Discriminative objectives for emotion regression

In the following, we will introduce two different objectives, one based on the CCC for each sequence, and one based on the overall CCC. We will denote by $y^i_f$ the regression outputs for sequence $i$ and target variable $f$ (in case of neural networks, the sequences of activations of unit $f$ of the output layer), while $y^{i*}_{f}$ denotes the corresponding training targets (i.e., gold-standard). The standard sum of squared errors (SSE) training objective for a mini-batch $\mathcal{B}$ is given by

$$\sum_{i \in \mathcal{B}} \sum_{f \in \mathcal{F}} \sum_t (y^i_{f,t} - y^{i*}_{f,t})^2$$

(1)

where $\mathcal{F}$ is the set of target variables (e.g., arousal, valence, etc. in the case of emotion recognition) and $t$ denotes the index of a time step at which the target variable is annotated.

While the above objective is discriminative, it is not discriminative on the sequence level. Let us thus introduce the proposed objective function based on the CCC per training sequence. The total cost function $\mathcal{O}$ is:

$$\mathcal{O} = - \sum_{i \in \mathcal{B}, f \in \mathcal{F}} CCC^2_i$$

(2)

This objective will be denoted by $\Sigma CCC$ below, as it is computed on all training sequences $i$ available in the mini-batch $\mathcal{B}$. The CCC per sequence $i$ and target $f$ is defined in accordance with [65] as:
\[ CCC^i_f = \frac{2 \text{Cov}(y_f^i, y_f^{i*})}{\text{Var}(y_f^i) + \text{Var}(y_f^{i*}) + (E(y_f^i) - E(y_f^{i*}))^2} \] (3)

where \( E, \text{Var}, \) and \( \text{Cov} \) denote sample mean, variance, and covariance, respectively.

Let us consider the mean squared error, \( E(y_f^i - y_f^{i*})^2 \), which is equivalent to

\[ \text{Var}(y_f^i - y_f^{i*}) + E(y_f^i) - y_f^{i*})^2 = \text{Var}(y_f^i) + \text{Var}(y_f^{i*}) + E(y_f^i - y_f^{i*})^2 - 2 \text{Cov}(y_f^i, y_f^{i*}) \] (4)

Based on this observation, we can rewrite the CCC as:

\[ CCC^i_f = \frac{Q_f^i}{S_f^i + Q_f^i} \] (5)

defining \( N^i \) as the length of sequence \( i \), the covariance-related quantity as:

\[ Q_f^i := N^i \text{Cov}\{y_f^i, y_f^{i*}\} = N^i(E\{y_f^i y_f^{i*}\} - E\{y_f^i\} E\{y_f^{i*}\}) = \sum_{t=1}^{N^i} y_f^i t y_f^{i*} - E\{y_f^i\} \sum_{t=1}^{N^i} y_f^i \] (6)

and the sum of squared errors (SSE) related quantity as:

\[ S_f^i := \frac{1}{2} \sum_{t=1}^{N^i} (y_f^i t - y_f^{i*})^2 \] (7)

An alternative objective to maximize (denoted simply by CCC below) is the ‘total’ CCC on the training set. This can be achieved by simply considering the entire training set as a single sequence \( i \) in (3). As shown in Fig. 5.2, the \( \Sigma \text{CCC} \) objective differs from the CCC objective in that it necessarily enforces accurate prediction of the target contour within each sequence, while the CCC objective could assign a good score to over-smoothed regression outputs that only predict the average label right. Conversely, if the target label has low variance within the sequences, the \( \Sigma \text{CCC} \) objective is hard to optimize and might emphasize on noise in the ‘gold-standard’, which is often given in emotion recognition. Thus, which objective is preferable certainly depends on the application.

Note that since the CCC on two partitions of the training set is not equivalent to the sum (or average) of the CCCs on these two partitions, it is not directly possible to optimize the total CCC on the training set, unless the mini-batch size comprises the total training set, which might be impractical. This is in contrast to SSE training and the \( \Sigma \text{CCC} \) maximization, where batch learning can be implemented by summing up the gradients from the mini-batches. In the case of mini-batch learning with \( |B| = 1 \) (one sequence per mini-batch), the optimization of \( \Sigma \text{CCC} \) and CCC is equivalent. However, in case of recurrent neural network training as considered here, \( |B| \gg 1 \) is required for efficiency [63]. Further, we could also define an objective in analogy to (2) yet based on the Pearson’s CC,

\[ - \sum_{i \in E, f \in \mathcal{F}} CCC^i_f = - \sum_{i \in B, f \in \mathcal{F}} \frac{\text{Cov}(y_f^i, y_f^{i*})}{\sigma_f^i \sigma_f^{i*}} \] (8)
where $\sigma$ denotes the standard deviations of outputs and targets in analogy to the above. However, since the Pearson CC is invariant w.r.t. the scale of the network outputs (in contrast to the CCC, cf. Fig. 5.1), this function has infinitely many minima, making its minimization hardly feasible.

Because CCC is not a linear function, its computation between the predictions and the gold-standards provides different results whether we sum it over the two sequences ($\Sigma CCC = 0$), or compute it on the two concatenated sequences ($CCC = 0.37$).

### 4.4. Training algorithm

In this study, optimization of the discriminative objectives is performed by stochastic gradient descent. For the proposed CCC objective, we compute the gradient $\nabla_y O = (\partial O / \partial y^i_{f,t})^t$. Using the quotient rule, the partial derivative for a sequence $i$, target $f$ and time step $t$ is computed as:

$$
\frac{\partial O}{\partial y^i_{f,t}} = -\frac{\partial Q^i}{\partial y^i_{f,t}} \cdot \left( S^i_f + Q^i_f \right) - Q^i_f \left( \frac{\partial S^i}{\partial y^i_{f,t}} + \frac{\partial Q^i}{\partial y^i_{f,t}} \right)
$$

With the partial derivatives:

$$
\frac{\partial Q^i}{\partial y^i_{f,t}} = y^i_{f,t} - E\left\{ y^i_f \right\}, \quad \frac{\partial S^i}{\partial y^i_{f,t}} = y^i_{f,t} - y^i_{f,t}
$$

With this, the desired derivative $\partial O / \partial y^i_{f,t}$ is obtained as:

$$
\frac{y^i_{f,t} - E\{y^i_f\} (S^i_f + Q^i_f) - Q^i_f (y^i_{f,t} - E\{y^i_f\})}{(S^i_f + Q^i_f)^2} = \frac{(y^i_{f,t} - E\{y^i_f\}) S^i_f - (y^i_{f,t} - y^i_{f,t}) Q^i_f}{(S^i_f + Q^i_f)^2}
$$

Having obtained the output gradient as above, the gradient w.r.t. the weights, $\partial O / \partial w$ is determined by backpropagation through time as usual. Discriminative training is implemented on top of the open source,
GPU-enabled neural network training software CURR ENNT [63], which supports deep feedforward and recurrent neural networks. The additional code which minimizes the proposed objectives will be made available upon publication of this manuscript.

During the forward pass, we compute and store the target means $E\{y_f^i\}$, as well as $S_f^i$ and $Q_f^i$ in matrices. The summations are computed using an outer loop over time steps, and using matrix addition for all $i$ and $f$ in parallel. For sequence lengths which are large in comparison to the number of target features and the batch size, this might still get inefficient. However, in our experiments, the speed of network training using the $\Sigma$CC, CCC, or the standard SSE cost function was in the same ballpark. The backward pass is similar to the calculation of the SSE backward pass. The output gradient can be computed for all $i, f$, and $t$ in parallel. In case of optimizing CCC rather than $\Sigma$CC, sequence boundaries need not be taken into account ($i = 1$). Consequently, the computation of the quantities $S_f^i$ and $Q_f^i$ for each $f$ is much simpler and similar to the SSE calculation.

4.5. Experiments and Results

We present in this section the results of time-continuous dimensional emotion (arousal and valence) prediction tasks on two different corpora from different domains (speech and music). The objective of those experiments is to empirically demonstrate the benefits of using CCC as cost function for network training, in comparison to the traditional SSE, for emotion recognition from speech and music.

4.5.1. Emotions from music: MediaEval

Experiments on emotion recognition from music are done on the ‘Emotion in Music Database’ which was used in the MediaEval 2013 evaluation campaign [71]. The task is to recognize time-varying emotion contours in the arousal and valence dimensions at a rate of 1Hz from music signals. The data set includes excerpts of 45 seconds randomly selected (uniform distribution) from 744 songs taken from the online library Free Music Archive1, and split between a development set (619 songs) and an evaluation set (125 songs). Ratings of emotion were performed on a crowdsourcing platform (MTurk) by a pool of 100 selected workers (57 male, 43 female, mean age is 32 years and standard deviation of 10 years) from different countries (72% from the USA, 18% from India and 10% from the rest of the world).

Both features extraction and machine learning steps are based on the setup reported in [85]. The 6373-dimensional ComParE set of generic affective features, and the Long Short-Term Memory (LSTM) [86] architecture for deep recurrent neural networks (DRNNs) are used. LSTM networks have two hidden layers with 192 or 256 hidden units. The ComParE set consists of supra-segmental acoustic features, i.e., summarization of frame-level features over segments of constant length. In the present study, supra-segmental features are extracted from non-overlapping segments of 1 second length, in accordance with the time resolution of the emotion contours.

The training parameters are preserved from [85]. Input noise with $\sigma = 0.6$ is added to help generalization, and an early stopping strategy is used to alleviate overfitting. Stochastic gradient descent with a batch size of 25
sequences is used in all experiments. The learning rate $\eta$ is determined in a preliminary cross-validation experiment for each objective function. Note that the objective functions are on different scales and hence the optimal step size varies between various objectives.

Both the sum of CCC and total CCC objectives are investigated. As baseline, standard SSE training is used. In accordance with the MediaEval challenge, the evaluation metrics comprise the overall Pearson’s correlation coefficient (CC)$^2$ as well as the average Kendall’s rank correlation coefficient per sequence ($\mathbb{E}\{\tau\}$), which is related to our $\Sigma$CCC objective function but not differentiable. Furthermore, we report the average CCC ($\mathbb{E}\{(\text{CCC})\}$) per sequence, which directly corresponds to the $\Sigma$CCC objective.

4.5.2. Emotions from speech: RECOLA

We used the full dataset of RECOLA, partitioned as in Table 5.1. The same procedure as the one used in the latest edition of the Audio-Visual Emotion Recognition Challenge (AV+EC 2015) [60] has been used to extract acoustic features from the speech recordings: the extended Geneva minimalistic acoustic feature set (eGeMAPS – 102 features) [87] has been applied at a rate of 40ms (to match the sampling frequency of the gold-standard) using overlapping windows of 3 seconds length.

For the prediction task, we used LSTM-DRNNs with three hidden layers with 128 units each. Input noise with $\sigma = 0.1$ is added and early stopping is also used to prevent overfitting. The networks were trained with stochastic gradient descent on a batch size of 5 sequences with a fixed momentum of 0.9, at different values of learning rate $\eta = \{10^{-2}, 10^{-3}, \ldots, 10^{-7}\}$. An optimal learning rate $\eta$ was chosen based on the CCC on the development set for each emotional dimension and objective function. The CCC metric was computed on the gold-standard and prediction values concatenated over all recordings, in accordance with the AV+EC challenge. In addition, we also report the average CCC ($\mathbb{E}\{(\text{CCC})\}$) per sequence in analogy to the experiments on music. For all the networks (regardless of the training objective), a chain of post-processing was applied to the predictions obtained on the development set: (i) median filtering (with size of window ranging from 0.4 second to 20 seconds) [60], (ii) centring (by computing the bias between gold-standard and prediction) [88], (iii) scaling (using the ratio of standard-deviation of gold standard and prediction as scaling factor) and (iv) time-shifting (by shifting the prediction forward in time with values ranging from 0.04 second to 10 seconds), to compensate for delays in the ratings [89]. Any of these post-processing steps was kept when an improvement was observed on the CCC of the development partition, and applied then with the same configuration on the test partition.
Table 5.1: Partitioning of the RECOLA database into train, dev(velopment), and test sets for continuous emotion recognition.

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Male</td>
<td>6</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>French</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Italian</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>German</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Portuguese</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Age μ (σ)</td>
<td>22.3 (3.4)</td>
<td>21.6 (2.1)</td>
<td>21.2 (2.0)</td>
</tr>
</tbody>
</table>

4.5.3. Results

Table 5.2 shows the results on the MediaEval 2013 test set (music). We can observe that the evaluation metrics exactly reflect the choice of the objective function: SSE training works best for minimizing the MLE, while CCC based training yields the best CCC on the test set.

The official Challenge evaluation metric, E{τ}, is significantly (according to a z-test, α = .05) improved by using the ΣCCC objective function (.221 → .251) for arousal but only slightly (.189 → .199) for valence. Generally, it is observed that the larger network with 256 hidden units performs worse on the test set, which can be attributed to the relatively small data set which causes over-fitting. The discrepancy between E{CCC} and CCC on this data set is astonishing; we found that for some test sequences, the variance in the annotated emotion contours is very low, which makes it hard to achieve good CC on these. One may further notice that the best performance measured as CC on valence is obtained with the CCC objective. The improvement over the SSE objective is significant (.637 → .653). Regarding the optimization of the network, results show that each objective function requires a specific learning rate to perform best.

Next, in Table 5.3 we report the metrics on the RECOLA database (speech). Here, we observe a significant improvement in the CC, CCC and E{CCC} metrics by using the ΣCCC objective function, particularly on the test set, where SSE training does not deliver useful results in the arousal dimension: CCC = .097 with SSE training and .350 with ΣCCC training. Since this difference is less pronounced on the development set, for which the network is tuned, we have some evidence that the ΣCCC objective function leads to better generalization. In fact, when training using the SSE criterion, we observed a bias of the network towards predicting the mean annotation on the training set, which leads to good RMSE but low correlation; conversely, the RMSE is significantly increased by using the CCC-based criteria. This result can also be observed on the CC evaluation metric, where a significant improvement over the SSE objective function is obtained when using ΣCCC for both arousal and valence. One may also note that the same results hold for the optimization of the network: each objective function requires a specific learning rate in order to provide the best performance.
4.6. Conclusions

We have demonstrated that the SSE objective in neural network regression can be effectively replaced by a criterion derived from the CCC, which has a significant impact on performance in continuous dimensional emotion recognition of arousal and valence from speech and music. Indeed, the CCC is an elegant solution to the issue of scaling and shifting time-continuous predictions, as it is sensitive to both of these variations and thus alleviates the problem of local minima in neural network training. Still, the observed increase in MSE-related criteria indicates that further investigations need to be performed in order to find an appropriate trade-off between MSE- and CC-like criteria.

Furthermore, note that the proposed approach based on CCC optimization can be applied to any sequence regression task where the correlation between the regression outputs and the ground truth should be maximized. There are no assumptions made on the underlying problem, other than that there be one or more continuous-valued target labels and that the regression model can be effectively trained by a first-order method such as stochastic gradient descent. Thus, we will verify its efficiency on other recognition tasks involving time-continuous measurements.
Finally, there are many more areas in emotion recognition and related fields where usage of the CCC can be explored, including the definition of the ‘gold-standard’ by computing the agreement of the raters, estimating an annotation delay, or the selection of features by using CCC instead of CC.

5. End-to-end Speech Emotion Recognition Using a Deep Convolutional Recurrent Network

5.1. Introduction and prior work

With the advent of deep neural networks in the last decade a number of ground-breaking improvements have been observed in several established pattern recognition areas such as object, speech and speaker recognition, as well as in combined problem solving approaches, e.g. in audio-visual recognition, and in the rather recent field of paralinguistics. For this purpose a series of new neural network architectures have been proposed, such as autoencoder networks, convolutional neural networks (CNN), or memory enhanced neural network models like Long Short-Term Memory (LSTM) models [55]. Numerous studies have shown the favourable property of these network variants to model inherent structure contained in the speech signal [90], with more recent research attempting end-to-end optimisation utilising as little human a-priori knowledge as possible [91]. Nevertheless, the majority of these works make use of commonly hand-engineered features have been used as input features, such as Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) coefficients, and supra-segmental features such as those used in the series of ComParE [92] and AVEC challenges [60], which build upon knowledge gained in decades of auditory research and have shown to be robust for many speech domains. Recently, however, a trend in the machine learning community has emerged towards deriving a representation of the input signal directly from raw, unprocessed data. The motivation behind this idea is that, ultimately, the network learns an intermediate representation of the raw input signal automatically that better suits the task at hand and hence leads to improved performance.

5.1.1. Related Work

In one of the first studies that suggested learning better features for automatic speech recognition (ASR) that used directly the speech waveform was Jaitly and Hinton [93]. Although they did not train the system in an end-to-end manner, they proposed learning an intermediate representation by training a Restricted Boltzmann Machine directly on the speech time signal. Experiments on the TIMIT phoneme recognition task demonstrated results that were on-par, or better than, state-of-the-art results at the time. More interestingly, the resulting learnt filters show the bandpass behaviour that auditory research has shown to exist in the human inner ear. Bhargava and Rose [94] used stacked bottleneck deep neural networks (DNNs) trained on windowed speech waveforms and obtained results only slightly worse than corresponding MFCC on the same architecture. Sainath et al. match the performance of a large-vocabulary speech recognition (LVCSR) system based on log-Mel filter-bank energies by using a Convolutional, LSTM-DNN [95], [96]. They observed that a time convolutional layer helps in reducing temporal variation, another frequency convolutional layer aids
in preserving locality and reducing frequency variation, while the LSTM layers serve for contextual modelling of the speech signal. Palaz et al. [97], [98] used CNNs directly trained on the speech signal to estimate phoneme class conditional probabilities and observed that the features learnt between the first two convolutional layers tend to model the phone-specific spectral envelope of sub-segmental speech signal, which leads to a more robust performance in noisy conditions. Deep CNN end-to-end learning was successfully applied on a music information retrieval task [99], and a similar model architecture was recently used for polyphonic music transcription [100]. In the field of paralinguistics, several studies have been carried out using CNNs for feature learning, e.g., recently by Milde and Biemann [101], and Mao et al. [102]. However, these works rely on a low-dimensional Mel filter-bank feature vector and hence did not do a full end-to-end training of their system.

5.1.2. Contribution of this work

In this work we study automatic affect sensing and prediction by training – directly on the underlying audio time signal – an end-to-end model that combines CNN and memory enhanced neural networks. To our knowledge this is the first work in literature that applies such a model to an emotion recognition task and our results show that this can successfully outperform state-of-the-art approaches based on designed features. Furthermore, we suggest using explicit maximisation of the concordance correlation coefficient (ρc) [103] in our model and show that this improves performance in terms of emotion prediction compared to optimising the mean square error objective, which is traditionally used. Finally, by further studying the activations of different cells in the recurrent layers, we find the existence of interpretable cells, which are highly correlated with several prosodic and acoustic features that were always assumed to convey affective information in speech, such as the loudness and the fundamental frequency.

5.2. Model design

One of the first steps in a traditional feature extraction process in audio is to use finite impulse response filters which perform time-frequency decomposition to reduce the influence of background noise [104]. More complicated hand-engineered kernels, such as gamma-tone filters [105], which were formulated by studying the frequency responses of the receptive fields of auditory neurons of grass frogs, can be used as well. A key component of our model are the 1-d convolutions that operate on the discrete-time waveform \( h(k) \).

\[
(f * h)(t) = \sum_{k=-T}^{T} f(t) \cdot h(t-k)
\]  

(1)

where \( f(x) \) is a kernel function whose parameters are learnt from the data of the task in hand. After the spatial-modelling of the signal which removes background noise and enhances specific parts of the signal for the task in hand, we model the temporal structure of speech by using a recurrent network with LSTM cells. We use LSTM for (i) simplicity, and (ii) to fairly compare against existing approaches which concentrated in the combination of hand-engineered features and LSTM networks. Finally, both subparts of our model are then trained jointly by backpropagation using the same objective function, cf. Equation 2.
5.2.1. Topology of the network

In contrast to previous work done in the field of paralinguistics, where acoustic features are first extracted and then passed to a machine learning algorithm, we aim at learning the feature extraction and regression steps in one jointly trained model for predicting the emotion. Our convolutional recurrent model is depicted in Figure 6.1 and summarised below.

**Input.** We segment the raw waveform to 6s long sequences after we preprocess the time-sequences to have zero mean and unit variance to account for variations in different levels of loudness between the speakers. At 16kHz sampling rate, this corresponds to a 96000-dimensional input vector. Temporal Convolution. We use \( F = 40 \) space time finite impulse filters with a 5ms window in order to extract fine-scale spectral information from the high sampling rate signal. Pooling across time. The impulse response of each filter is passed through a half-wave rectifier (analogous to the cochlear transduction step in the human ear) and then down-sampled to 8kHz by pooling each impulse response with a pool size = 2.

**Temporal Convolution.** We use \( M = 40 \) space time finite impulse filters of 500ms window. These are used to extract more long-term characteristics of the speech and the roughness of the speech signal.

**Max pooling across channels.** We perform max-pooling across the channel domain with a pool size of 20. This reduces the dimensionality of the signal while preserving the necessary statistics of the convolved signal.

**Recurrent layers.** We segment the 6s sequences to 150 smaller sub-sequences to match the granularity of the annotation frequency of 40ms. We use two bidirectional LSTM layers with 128 cells each [57], [106], although we get similar performance with the unidirectional approach.

![Diagram of convolutional recurrent network topology](image)

Fig. 1. Illustration of the proposed convolutional recurrent network topology for emotion prediction from the raw waveform signal. The convolutional layers replace the need for hand-engineering features which were used till now in the paralinguistics community.

5.2.2. Objective function

To evaluate the agreement level between the predictions of the network and the gold-standard derived from the annotations, a similar objective function as in Section 4.3 has been used.
5.3. Experiments and dataset

We used the full RECOLA dataset (D4.7. Section 3). The same procedure as the one used in the latest edition of the Audio-Visual Emotion Recognition Challenge (AV+EC 2015) is used to extract acoustic features from the speech recordings: the extended Geneva minimalistic acoustic feature set (eGeMAPS) [87] is applied at a rate of 40ms using overlapping windows of 3s length. Because the complexity of this feature set is quite low, and could thus make the comparison unfair with the CNN approach, we also extracted the low-level descriptors (LLDs) that are used in the series of computational paralinguistic challenges (ComParE) [92]. We then applied functionals (max, min, range, mean, and standard-deviation) [40] with the same rate and window length as used for eGeMAPS, on those LLDs.

As a first baseline machine learning algorithm, we used Support Vector Regression models with a linear kernel – polynomial and RBF kernels provided lower performance, using the libsvm library. The complexity parameter is optimised with a logarithmic grid in the range \([10^{-6} - 10^0]\). As a second baseline algorithm, we utilised a BLSTM-DRNNs with the architecture preserved from [40], [60], i.e., we used three hidden layers with 64 units for each layer. Input noise with \(\sigma = 0.1\) is added and early stopping is also used to prevent overfitting. Stochastic gradient descent with a batch size of 5 sequences is used in all experiments. The learning rate of the network is optimised on the validation set for each emotional dimension (arousal, valence) and objective function (MSE, \(\rho_c\)), using the \(\rho_c\) as evaluation performance, which is computed on the gold-standard and prediction values concatenated over all recordings, in accordance with the approach defined in the AV+EC challenge [60]

For training our proposed model, we utilised stochastic optimisation, with a mini-batch of 50 samples, Adam optimisation method [107], and a fixed learning rate of \(2 \cdot 10^{-3}\) throughout all experiments. Also, for regularisation of the network, we used dropout [108] with \(p = 0.5\) for all layers except the recurrent ones. This step is important as our models have a large amount of parameters (≈ 1.5M) and not regularising the network makes it prone on overfitting on the training data.

Finally, for all investigated methods, a chain of post-processing is applied to the predictions obtained on the validation set: (i) median filtering (with size of window ranging from 0.4s to 20s) [60], (ii) centring (by computing the bias between gold-standard and prediction) [88], (iii) scaling (using the ratio of standard-deviation of gold-standard and prediction as scaling factor) [88] and (iv) time-shifting (by shifting the prediction forward in time with values ranging from 0.04s to 10s), to compensate for delays in the ratings [89]. Any of these post-processing steps is kept when an improvement is observed on the \(\rho_c\) of the validation set, and applied then with the same configuration on the test partition. Results obtained for each method are shown in Table 6.1. In all of the experiments, our model outperforms the designed features in terms of \(\rho_c\).

One may note, however, that the eGEMAPS feature set provides close performance on valence, which is much more difficult to predict from speech compared to arousal. Furthermore, we show that by incorporating \(\rho_c\) directly in the optimisation function of all networks allows us to optimise the models on the metric (\(\rho_c\)) on which we evaluate the models. This provides us with (i) a more elegant way to optimise models, and (ii) gives consistently better results across all test-runs as seen in Table 6.1.
Table 6.1. RECOLA dataset results (in terms of $\rho_c$) for prediction of arousal and valence. In parenthesis are the performance obtained on the validation set. In a) we optimised the models w.r.t. MSE whereas in b) w.r.t. $\rho_c$.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Features</th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVR</td>
<td>eGeMAPS</td>
<td>.318 (.489)</td>
<td>.169 (.210)</td>
</tr>
<tr>
<td>SVR</td>
<td>ComParE</td>
<td>.366 (.491)</td>
<td>.180 (.178)</td>
</tr>
<tr>
<td>BLSTM</td>
<td>eGeMAPS</td>
<td>.300 (.404)</td>
<td>.192 (.187)</td>
</tr>
<tr>
<td>BLSTM</td>
<td>ComParE</td>
<td>.132 (.221)</td>
<td>.117 (.152)</td>
</tr>
<tr>
<td>Proposed</td>
<td>raw signal</td>
<td>.684 (.728)</td>
<td>.249 (.312)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Concordance correlation coefficient objective</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLSTM</td>
<td>eGeMAPS</td>
<td>.316 (.445)</td>
<td>.195 (.190)</td>
</tr>
<tr>
<td>BLSTM</td>
<td>ComParE</td>
<td>.382 (.478)</td>
<td>.187 (.246)</td>
</tr>
<tr>
<td>Proposed</td>
<td>raw signal</td>
<td>.686 (.741)</td>
<td>.261 (.325)</td>
</tr>
</tbody>
</table>

5.4. Relation to existing acoustic and prosodic features

The speech signals convey information about the affective state either explicitly, i.e., by linguistic means, or implicitly, i.e., by acoustic or prosodic cues. It is well accepted amongst the research community that certain acoustic and prosodic features play an important role in recognising the affective state [109]. Some of these features, such as the mean of the fundamental frequency (F0), mean speech intensity, loudness, as well as pitch range [87], should thus be captured by our model.

To gain a better understanding of what our model learns, and how this relates to existing literature, we study the statistics of gate activations in the network applied on an unseen speech recording; a visualisation of the hidden-to-output connections of different cells in the recurrent layers of the network is given in Figure 6.2. This plot shows that certain cells of the model are very sensitive to different features conveyed in the original speech waveform.

Fig. 6.2. A visualisation of three different gate activations vs. different acoustic and prosodic features that are known to affect arousal for an unseen recording to the network. From top to bottom: range of RMS energy ($\rho = 0.81$), loudness ($\rho = 0.73$), mean of fundamental frequency ($\rho = 0.72$)
5.5. Conclusions

In this section, we propose a convolutional recurrent model that operates on the raw signal, to perform an end-to-end spontaneous emotion prediction task from speech data. Further, we propose the direct optimisation of the concordance correlation coefficient, which is used to evaluate the agreement rate between the predictions and the gold-standard. The proposed method achieves significantly better performance in comparison to traditional designed features on the RECOLA database, thus demonstrating the efficacy of learning features that better suit the task-at-hand. As a final contribution, we study the gate activations of the recurrent layers and find cells that are highly correlated with prosodic features that were always assumed to cause arousal.

6. Final Conclusion

One goal of the MixedEmotions’ project is to develop an emotion recognition system from speech signals. The third pilot of MixedEmotions (Call-Centre) is a potential environment to deploy such system, where automatic analysis of the emotions of the customers is a great value for enhancing the interaction between customers and agents. In such scenarios, continuous (in time) evaluation of emotions is highly important in order to locate which durations in a conversation are getting (un-)pleasant. In this deliverable, we presented some developed state-of-the-art methods for improving continuous emotion recognition by enhancing features, defining a time-continuous objective function, as well as an end-to-end emotion recognition. We observed significant improvements with respect to the traditional approaches on emotional speech corpora. We will further investigate the possibilities of employing these approaches into the MixedEmotions’ framework for the Call-Centre and Social-TV pilots.

References:


[70] E. Coutinho and A. Cangelosi, “A neural network model for the prediction of musical emotions,” in *Advances in*...


