ADIEU FEATURES? END-TO-END SPEECH EMOTION RECOGNITION USING A DEEP CONVOLUTIONAL RECURRENT NETWORK

George Trigeorgis[1], *Fabien Ringeval*[2,3], *Raymond Brueckner*[3,4], *Erik Marchi*[3]
Mihalis A. Nicolaou[5], *Björn Schuller*[1,2,6],*Stefanos Zafeiriou*[1]
[1]Department of Computing, Imperial College London, London, UK
[2]Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany
[3]Machine Intelligence & Signal Processing Group, MMK,
Technische Universität München, Munich, Germany
[4]Nuance Communications Deutschland GmbH, Ulm, Germany
[5]Department of Computing, Goldsmiths, University of London, UK
[6]audEERING UG, Gilching, Germany
g.trigeorgis@imperial.ac.uk

ABSTRACT

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 while on the other, 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 this paper, 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.

Index Terms— end-to-end learning, raw waveform, emotion recognition, deep learning, CNN, LSTM

1. INTRODUCTION AND PRIOR WORK

With the advent of deep neural networks in the last decade a number of groundbreaking 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 such as Long Short-Term Memory (LSTM) models [1].

Numerous studies have shown the favourable property of these network variants to model inherent structure contained in the speech signal [2], with more recent research attempting *end-to-end* optimisation utilising as little human a-priori knowledge as possible [3]. 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 [4] and AVEC challenges [5], 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.

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 [6]. 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 [7] used stacked bottleneck deep neural networks (DNNs) trained on windowed speech waveforms and obtained results only slight 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 filterbank energies by using a Convolutional, LSTM-DNN [8, 9]. They observed that a time convolution layer helps in reducing temporal variation, another frequency convolution layer aids in preserving locality and reducing frequency variation, while the the LSTM layers serve for contextual modelling of the speech signal.

Palaz et al. [10, 11] used CNNs directly trained on the speech signal to estimate phoneme class conditional probabilities and observed that the features learnt between the first two convolution 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 [12], and a similar model architecture was recently used for polyphonic music transcription [13].

In the field of paralinguistics, several studies have been carried out using CNNs for feature learning, e.g., recently by Milde and Biemann [14], and Mao et al. [15]. However, these works rely on a low-dimensional Mel filterbank feature vector and hence did not do a full *end-to-end* training of their system.

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* (ρ_c) [16] 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.

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 [17]. More complicated hand-engineered kernels, such as gammatone filters [18], which were formulated by studying the frequency responses of the receptive fields of auditory neurons of grassfrogs, can be used as well.

A key component of our model are the 1-d convolutions that operate on the discrete-time waveform $h(k)$.

$$(f \star h)(t) = \sum_{k=-T}^{T} f(t) \cdot h(t-k) \tag{1}$$

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.

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 reccurrent model is depicted in Figure 1 and summarised below.

Input. We segment the raw waveform to 6 s 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 16 kHz 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 8 kHz 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 6 s sequences to 150 smaller sub-sequences to match the granularity of the annotation frequency of 40 ms. We use two bidirectional LSTM layers with 128 cells each [19, 20], although we get similar performance with the uni-directional approach.

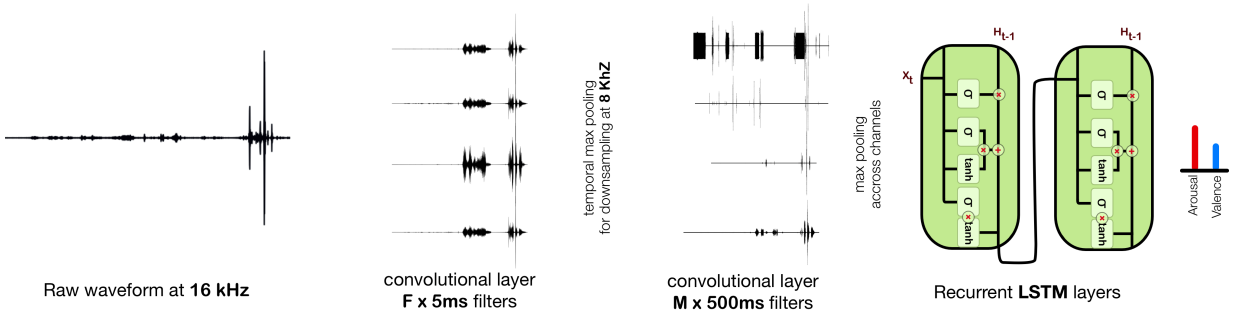

Fig. 1. Illustration of the proposed convolutional recurrent network topology for emotion prediction from the raw waveform signal. The convolutional layers replace the need for hand-engineering features which were used till now in the paralinguistics community.

2.2. Objective function

To evaluate the agreement level between the predictions of the network and the gold-standard derived from the annotations, the concordance correlation coefficient (ρ_c) [16] has recently been proposed [21, 5]. Nonetheless, previous work minimised the MSE during the training of the networks, but evaluated the models with respect to ρ_c [21, 5]. Instead, we propose to include the metric used to evaluate the performance in the objective function ($\mathcal{L}_c$) used to train the networks. Since the objective function is a cost function, we define $\mathcal{L}_c$ as follow:

$$\mathcal{L}_c = 1 - \rho_c = 1 - \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} = 1 - 2\sigma_{xy}^2\psi^{-1} \quad (2)$$

where $\psi = \sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2$ and $\mu_x = \mathbb{E}(\mathbf{x})$, $\mu_y = \mathbb{E}(\mathbf{y}), \sigma_x^2 = \text{var}(\mathbf{x}), \sigma_y^2 = \text{var}(y)$ and $\sigma_{xy}^2 = \text{cov}(\mathbf{x}, \mathbf{y})$ Thus, to minimise $\mathcal{L}_c$ (or maximise ρ_c), we backpropagate the gradient of the last layer weights with respect to $\mathcal{L}_c$,

$$\frac{\partial \mathcal{L}_c}{\partial \mathbf{x}} \propto 2\frac{\sigma_{xy}^2(\mathbf{x} - \mu_y)}{\psi^2} + \frac{\mu_y - \mathbf{y}}{\psi}, \quad (3)$$

where all vector operations are done element-wise.

3. EXPERIMENTS AND DATA SET

Time-continuous prediction of spontaneous and natural emotions (arousal and valence) is investigated on speech data by using the RECOLA database [22]; the full dataset is used for the purpose of this study, which corresponds to speech recordings from 46 French-speaking participants with 5 minutes for each. The dataset is split equally in three partitions – train (16 subjects), validation (15 subjects) and test (15 subjects) – by stratifying (i.e., balancing) the gender and the age of the speakers. The same procedure as the one used in the latest edition of the Audio-Visual Emotion Recognition Challenge (AV$^+$EC 2015) [5] is used to extract acoustic features from the speech recordings: the extended Geneva minimalistic acoustic feature set (eGeMAPS) [23] is applied at a rate of 40 ms using overlapping windows of 3 s 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) [4]. We then applied functionals (max, min, range, mean, and standard-deviation) [21] 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 [5, 21], 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, ρ_c), using the ρ_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 [5].

For training our proposed model, we utilised stochastic optimisation, with a mini-batch of 50 samples, Adam optimisation method [24], and a fixed learning rate of $2 \cdot 10^{-3}$ throughout all experiments. Also, for regularisation of the network, we used dropout [25] with $p = 0.5$ for all layers except the recurrent ones. This step is important as our models have a large amount of parameters ($\approx 1.5M$) 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 devel-

opment set: *(i)* median filtering (with size of window ranging from 0.4 s to 20 s) [5], *(ii)* centring (by computing the bias between gold-standard and prediction) [26], *(iii)* scaling (using the ratio of standard-deviation of gold-standard and prediction as scaling factor) [26] and *(iv)* time-shifting (by shifting the prediction forward in time with values ranging from 0.04 s to 10 s), to compensate for delays in the ratings [27]. Any of these post-processing steps is kept when an improvement is observed on the ρ_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 1. In all of the experiments, our model outperforms the designed features in terms of ρ_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 ρ_c directly in the optimisation function of all networks allows us to optimise the models on the metric (ρ_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 1.

Predictor	Features	Arousal	Valence
a. Mean squared error objective			
SVR	eGeMAPS	.318 (.489)	.169 (.210)
SVR	ComParE	.366 (.491)	.180 (.178)
BLSTM	eGeMAPS	.300 (.404)	.192 (.187)
BLSTM	ComParE	.132 (.221)	.117 (.152)
Proposed	raw signal	.684 (.728)	.249 (.312)
b. Concordance correlation coefficient objective			
BLSTM	eGeMAPS	.316 (.445)	.195 (.190)
BLSTM	ComParE	.382 (.478)	.187 (.246)
Proposed	raw signal	**.686 (.741)**	**.261 (.325)**

Table 1. RECOLA dataset results (in terms of ρ_c) for prediction of arousal and valence. In parenthesis are the performance obtained on the development set. In *a)* we optimised the models wrt. MSE whereas in *b)* wrt. ρ_c.

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 [28]. Some of these features, such as the mean of the fundamental frequency (F0), mean speech intensity, loudness, as well as pitch range [23], should thus be captured by our model.

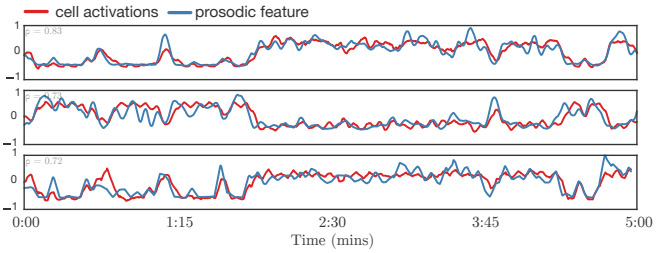

Fig. 2. A visualisation of three different gate activations vs. different acoustic and prosodic features that are known to affect arousal for an unseen recording to the network. From top to bottom: range of RMS energy ($\rho = 0.81$), loudness ($\rho = 0.73$), mean of fundamental frequency ($\rho = 0.72$)

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 2. This plot shows that certain cells of the model are very sensitive to different features conveyed in the original speech wave form.

5. CONCLUSIONS

In this paper, 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. ACKNOWLEDGEMENTS

George Trigeorgis is a recipient of the fellowship of the Department of Computing, Imperial College London, and this work was partially funded by it. This work was partially supported by the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the EU's Horizon 2020 Programme through the Innovative Action No. 644632 (MixedEmotions), No. 645094 (SEWA) and the Research Innovative Action No. 645378 (ARIA-VALUSPA). We would like to thank the NVIDIA Corporation for donating a Tesla K40 GPU used in this work.

7. REFERENCES

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, January 2015.

[2] G. Hinton, L. Deng, Y. Dong, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. ICML*, Beijing, China, 2014, pp. 1764–1772.

[4] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 148–152, ISCA.

[5] F. Ringeval et al., "AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proc. AVEC*, Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic, Eds., Brisbane, Australia, October 2015, pp. 3–8, ACM.

[6] N. Jaitly and G. Hinton, "Learning a better representation of speech sound waves using restricted Boltzmann machines," in *Proc. ICASSP*, Prague, Czech Republic, May 2011, pp. 5884–5887, IEEE.

[7] M. Bhargava and R. Rose, "Architectures for deep neural network based acoustic models defined over windowed speech waveforms," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 6–10, ISCA.

[8] T. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. ICASSP*, Brisbane, Australia, April 2015, pp. 4580–4584, IEEE.

[9] T. Sainath, R. Weiss, A. Senior, K. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 1–5, ISCA.

[10] D. Palaz, R. Collobert, and M. Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proc. INTERSPEECH*, Lyon, France, August 2013, pp. 1766–1770, ISCA.

[11] D. Palaz, M. Magimai-Doss, and R. Collobert, "Analysis of cnn-based speech recognition system using raw speech as input," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 11–15, ISCA.

[12] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. ICASSP*, Florence, Italy, April 2014, pp. 7014–7018.

[13] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic music transcription," *arXiv*, vol. arXiv:1508.01774, 2015.

[14] B. Milde and C. Biemann, "Using representation learning and out-of-domain data for a paralinguistic speech task," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 904–908, ISCA.

[15] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec 2014.

[16] L. I-Kuei Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, March 1989.

[17] H.G. Hirsch, P. Meyer, and H.W. Ruehl, "Improved speech recognition using high-pass filtering of subband envelopes," in *Proc. EUROSPEECH*, Genoa, Italy, September 1991, pp. 413–416.

[18] R. Schlüter, L. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. ICASSP*. IEEE, April 2007, vol. 4, pp. 649–652.

[19] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks, IJCNN Special Issue*, vol. 18, no. 5-6, pp. 602–610, July-August 2005.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] F. Ringeval et al., "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, November 2015.

[22] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions," in *Proc. of EmoSPACE, FG*, Shanghai, China, 2013.

[23] F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, 2015, in press.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, January 2014.

[26] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels, "Ensemble methods for continuous affect recognition: Multimodality, temporality, and challenges," in *Proc. AVEC*, Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic, Eds., Brisbane, Australia, October 2015, pp. 9–16.

[27] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, April-June 2015.

[28] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.