INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition

Soheil Khorram[1*], *Zakaria Aldeneh*[1*], *Dimitrios Dimitriadis*[2], *Melvin McInnis*[1], *Emily Mower Provost*[1]

[1]University of Michigan at Ann Arbor
[2]IBM T. J. Watson Research Center

{khorrams, aldeneh, mmcinnis, emilykmp}@umich.edu, dbdimitr@us.ibm.com

Abstract

The goal of continuous emotion recognition is to assign an emotion value to every frame in a sequence of acoustic features. We show that incorporating long-term temporal dependencies is critical for continuous emotion recognition tasks. To this end, we first investigate architectures that use dilated convolutions. We show that even though such architectures outperform previously reported systems, the output signals produced from such architectures undergo erratic changes between consecutive time steps. This is inconsistent with the slow moving ground-truth emotion labels that are obtained from human annotators. To deal with this problem, we model a downsampled version of the input signal and then generate the output signal through upsampling. Not only does the resulting downsampling/upsampling network achieve good performance, it also generates smooth output trajectories. Our method yields the best known audio-only performance on the RECOLA dataset.

Index Terms: neural networks, convolutional neural networks, computational paralinguistics, emotion recognition

1. Introduction

Emotion recognition has many potential applications including building more natural human-computer interfaces. Emotion can be quantified using categorical classes (e.g., *neutral, happy, sad, etc.*) or using dimensional values (e.g., *valence-arousal*). In addition, emotional labels can be quantified statically, over units of speech (e.g., utterances), or continuously in time.

In this work, we focus on problems where the goal is to recognize emotions in the valence-arousal space, continuously in time. The valence-arousal space is a psychologically grounded method for describing emotions [1]. Valence ranges from negative to positive, while activation ranges from calm to excited. Research has demonstrated that it is critical to incorporate long-term temporal information for making accurate emotion predictions. For instance, Valstar et al. [2] showed that it was necessary to consider larger windows when making frame-level emotion predictions (four seconds for arousal and six seconds for valence). Le et al. [3] and Cardinal et al. [4] found that increasing the number of contextual frames when training a deep neural network (DNN) for making frame-level emotion predictions is helpful but only to a certain point. Bidirectional long short-term memory networks (BLSTMs) can naturally incorporate long-term temporal dependencies between features; explaining their success in continuous emotion recognition tasks (e.g., [5]).

In this work, we investigate two convolutional network architectures, dilated convolutional networks and downsampling/upsampling networks, that capture long-term temporal dependencies. We interpret the two architectures in the context of continuous emotion recognition and show that these architectures can be used to build accurate continuous emotion recognition systems.

[*]These authors contributed equally to this work

2. Related Work

Even though the problem of emotion recognition has been extensively studied in the literature, we only focus on works that predicted dimensional values, continuously in time. Successful attempts to solving the continuous emotion recognition problem relied on DNNs [4], BLSTMs [5], and more commonly, support vector regression (SVR) classifiers [6]. With the exception of BLSTMs, such approaches do not incorporate long-term dependencies unless coupled with feature engineering. In this work, we show that purely convolutional neural networks can be used to incorporate long-term dependencies and achieve good emotion recognition performance, and are more efficient to train than their recurrent counterparts.

In their winning submission to the AVEC 2016 challenge, Brady et al. [6] extracted a set of audio features (Mel-frequency cepstral coefficients, shifted delta cepstral, prosody) and then learned higher-level representations of the features using sparse coding. The higher-level audio features were used to train linear SVRs. Povolny et al. [7] used eGeMAPS [8] features along with a set of higher-level bottleneck features extracted from a DNN trained for automatic speech recognition (ASR) to train linear regressors. The higher level features were produced from an initial set of 24 Mel filterbank (MFB) features and four different estimates of the fundamental frequency (F0). Povolny et al. used all features to train linear regressors to predict a value for each frame, and considered two methods for incorporating contextual information: simple frame stacking and temporal content summarization by applying statistics to local windows. In contrast, in this work we show that considering temporal dependencies that are longer than those presented in [6, 7] is critical to improve continuous emotion recognition performance.

He et al. [5] extracted a comprehensive set of $4,684$ features, which included energy, spectral, and voicing-related features, and used them to train BLSTMs. The authors introduced delay to the input to compensate for human evaluation lag and then applied feature selection. The authors ran the predicted time series through a Gaussian smoothing filter to produce the final output. In this work, we show that it is sufficient to use 40 MFBs to achieve state-of-the-art performance, without the need for special handling of human evaluation lag.

Trigeorgis et al. [9] trained a convolutional recurrent network for continuous emotion recognition using the time domain signal directly. The authors split the utterances into five-second segments for batch training. Given an output from a the trained model, the authors applied a chain of post-processing steps (median filtering, centering, scaling, time shifting) to get the final

http://dx.doi.org/10.21437/Interspeech.2017-548

output. In contrast, we show that convolutional networks make it possible to efficiently process full utterances without the need for segmenting. Further, since our models work on full-length utterances, we show that it is not necessary to apply any post-processing steps as described in [9].

On the ASR end, Sercu et al. [10] proposed viewing ASR problems as dense prediction tasks, where the goal is to assign a label to every frame in a given sequence, and showed that this view provides a set of tools (e.g., dilated convolutions, batch normalization, efficient processing) that can improve ASR performance. The authors argued that ASR approaches required practitioners to splice their input sequences into independent windows, making the training and evaluation procedures cumbersome and computationally inefficient. In contrast, the authors' proposed approach allows practitioners to efficiently process full sequences without requiring splicing or processing frames independently. The authors showed that their approach obtained the best published single model results on the switchboard-2000 benchmark dataset.

In this work, we treat the problem of continuous emotion recognition as a dense prediction task and show that, given this view of the problem, we can utilize convolutional architectures that can efficiently incorporate long-term temporal dependencies and provide accurate emotion predictions.

3. Problem Setup

We focus on the RECOLA database [11] following the AVEC 2016 guidelines [2]. The RECOLA database consists of spontaneous interactions in French and provides continuous, dimensional (valence and arousal) ground-truth descriptions of emotions. Even though the AVEC 2016 challenge is multi-modal in nature, we only focus on the speech modality in this work. The RECOLA database contains a total of 27 five-minute utterances, each from a distinct speaker (9 train; 9 validation; 9 test). Ground-truth continuous annotations were computed, using audio-visual cues, on a temporal granularity of 40ms from six annotators (three females).

Features. We use the Kaldi toolkit [12] to extract 40-dimensional log MFB features, using a window length of 25ms with a hop size of 10ms. Previous work showed that MFB features are better than conventional MFCCs for predicting emotions [13]. We perform speaker-specific z-normalization on all extracted features. RECOLA provides continuous labels at a granularity of 40ms. Thus, we stack four subsequent MFB frames to ensure correspondence between hop sizes in the input and output sequences.

Problem Setup. Given a sequence of stacked acoustic features $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$, where $\boldsymbol{x}_t \in \mathbb{R}^d$, the goal is to produce a sequence of continuous emotion labels $\boldsymbol{y} = [y_1, y_2, \ldots, y_T]$, where $y_t \in \mathbb{R}$.

Evaluation Metrics. Given a sequence of ground-truth labels $\boldsymbol{y} = [y_1, y_2, \ldots, y_T]$ and a sequence of predicted labels $\hat{\boldsymbol{y}} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_T]$, we evaluate the performance using the root mean squared error (RMSE) and the Concordance Correlation Coefficient (CCC) to be consistent with previous work. The CCC is computed as follows:
$\text{CCC} = 2\sigma^2_{y\hat{y}}/(\sigma^2_y + \sigma^2_{\hat{y}} + (\mu_y - \mu_{\hat{y}})^2)$, where $\mu_y = \mathbb{E}(\boldsymbol{y})$, $\mu_{\hat{y}} = \mathbb{E}(\hat{\boldsymbol{y}})$, $\sigma^2_y = \text{var}(\boldsymbol{y})$, $\sigma^2_{\hat{y}} = \text{var}(\hat{\boldsymbol{y}})$, and $\sigma^2_{y\hat{y}} = \text{cov}(\boldsymbol{y}, \hat{\boldsymbol{y}})$.

4. Preliminary Experiment

We first study the effect of incorporating temporal dependencies of different lengths. The network that we use in the preliminary experiments consists of a convolutional layer with one filter of variable length from 2 to 2048 frames, followed by a $tanh$ nonlinearity, followed by a linear regression layer. We vary the length of the filter and validate the performance using CCC. We train our model on the training partition and evaluate on the development partition. We report the results of our preliminary experiment in Figure 1. The results show that incorporating long-term temporal dependencies improves the performance on the validation set up to a point.

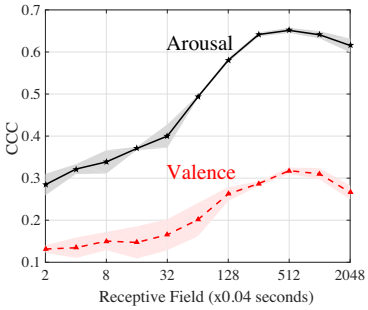

Figure 1: *Increasing the size of the receptive field improves performance for both arousal and valence. Solid lines represent mean CCC from* 10 *runs and shaded area represents standard deviation from the runs.*

The observed diminishing gains in performance past 512 (20.48 seconds) frames may occur either due to the increased number of parameters or because contextual information becomes irrelevant after 512 frames. Covering contexts as large as 512 frames still provided improvements in performance compared to results obtained from covering smaller contexts. The utility of contexts spanning 512 frames (20.48 seconds) is contrary to previous work that considered much smaller time scales. For instance, Valstar et al. [2] only covered six seconds worth of features and Povolny et al. [7] considered a maximum of eight seconds worth of features. Results from the preliminary experiment suggest that continuous emotion prediction systems could benefit from incorporating long-term temporal dependencies. This acts as a motivation for using architectures that are specifically designed for considering long-term dependencies.

5. Methods

In this section, we describe the two architectures that we propose to use to capture long-term temporal dependencies in continuous emotion prediction tasks.

5.1. Dilated Convolutions

Dilated convolutions provide an efficient way to increase the receptive field without causing the number of learnable parameters to vastly increase. Networks that use dilated convolutions have shown success in a number of tasks, including image segmentation [14], speech synthesis [15] and ASR [10].

van den Oord et al. [15] recently showed that it is possible to use convolutions with various dilation factors to allow the receptive field of a generative model to grow exponentially in order to cover thousands of time steps and synthesize high-quality speech. Sercu et al. [10] showed that ASR could benefit from dilated convolutions since they allow larger regions to be covered without disrupting the length of the input signals. Continuous emotion recognition could benefit from such properties.

When compared to filters of regular convolutions, those of dilated convolutions touch the input signal every k time steps,

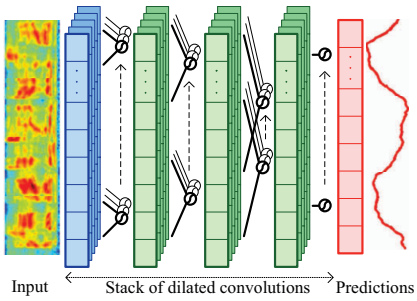

Figure 2: *A visualization of our dilated convolution network. We use convolutions with a different dilation factor for different layers. We use a* 1×1 *convolution for the last layer to produce the final output.*

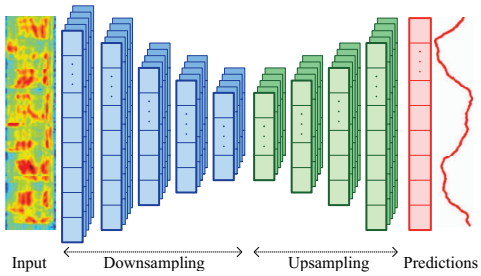

Figure 3: *A visualization of our downsampling/upsampling network. Downsampling compresses the input signal into shorter signal which is then used to reconstruct a signal of the same length by the upsampling sub-network. We use the transpose convolution operation to perform upsampling.*

where k is the *dilation factor*. If $[w_1, w_2, w_3]$ is a filter with a dilation factor of zero, then $[w_1, 0, w_2, 0, w_3]$ is the filter with a dilation factor of one and $[w_1, 0, 0, w_2, 0, 0, w_3]$ is the filter with a dilation factor of two, and so on. We build a network that consists of stacked convolution layers, where the convolution functions in each layer use a dilation factor of 2^n, where n is the layer number. This causes the dilation factors to grow exponentially with depth while the number of parameters grows linearly with depth. Figure 2 shows a diagram of our dilated convolution network.

5.2. Downsampling/Upsampling

The emotion targets in the RECOLA database are sampled at a frequency of 25 Hz. Using Fourier analysis, we find that more than 95 percent of the power of these trajectories lies in frequency bands that are lower than 1 Hz. In other words, the output signals are smooth and they have considerable time dependencies. This finding is not surprising because we do not expect rapid reactions from human annotators. Networks that use dilated convolutions do not take this fact into account while making predictions, causing them to generate output signals whose variance is not consistent with the continuous ground truth contours (Section 6.2). To deal with this problem, we propose the use of a network architecture that compresses the input signal into a low-resolution signal through downsampling and then reconstructs the output signal through upsampling. Not only does the downsampling/upsampling architecture capture long-term temporal dependencies, it also generates a smooth output trajectory.

We conduct an experiment to investigate the effect of downsampling/upsampling on continuous emotion labels. First, we convert the ground truth signals to low-resolution signals using standard uniform downsampling. Given the downsampled signals, we then generate the original signals using spline interpolation. We vary the downsampling factor exponentially from 2 to 128 and compute the CCC between the original signals and the reconstructed ones. The results that we show in Figure 4 demonstrate that distortions caused by downsampling with factors up to 64 are minor ($< 5\%$ loss in CCC relative to original).

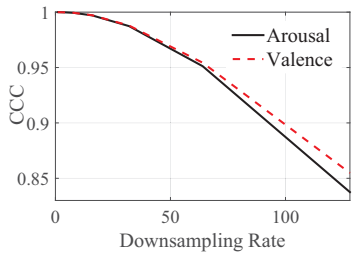

Figure 4: *Effect of downsampling/upsampling on CCC.*

The network that we use contains two subnetworks: (1) a downsampling network; (2) an upsampling network. The downsampling network consists of a series of convolutions and max-pooling operations. The max-pooling layers reduce the resolution of the signal and increase the effective receptive field of the convolution layers. Initial experiments showed that max-pooling was more effective than other pooling techniques.

The upsampling function can be implemented in a number of ways [16]. In this work we use the transposed convolution[1] [17, 18] operation to perform upsampling. Transposed convolutions provide a learnable map that can upsample a low-resolution signal to a high-resolution one. In contrast to standard convolution filters that connect multiple input samples to a single output sample, transposed convolution filters generate multiple outputs samples from just one input sample. Since it generates multiple outputs simultaneously, the transposed convolution can be thought of as a learnable interpolation function.

Downsampling/upsampling architectures have been used in many computer vision tasks (e.g., [16, 19, 20]). For instance, Noh et al. [16] showed that transposed convolution operations can be effectively applied to image segmentation tasks. In addition to vision applications, downsampling/upsampling architectures have been successfully applied to speech enhancement problems [21], where the goal is to learn a mapping between noisy speech spectra and their clean counterparts. Park et al. [21] demonstrated that downsampling/upsampling convolutional networks can be $12\times$ smaller (in terms of the number of learnable parameters) than their recurrent counterparts and yet yield better performance on speech enhancement tasks.

The main goal of a transposed convolution is to take an n_x-dimensional low-resolution vector $\boldsymbol{x}$ and generate an n_y-dimensional high-resolution vector $\boldsymbol{y}$ using an n_w-dimensional filter $\boldsymbol{w}$ (where $n_y > n_x$). Similar to other linear transforms, $\boldsymbol{y}$ can be expressed as: $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{x}$, where $\boldsymbol{T}$ is the linear n_y-by-n_x transform matrix that is given by $\boldsymbol{T} = [\boldsymbol{T}_1, \boldsymbol{T}_2, ..., \boldsymbol{T}_{n_x}]$. $\boldsymbol{T}_i$ is the i-th column of $\boldsymbol{T}$ and can be written as:

$$\boldsymbol{T}_i = [\underbrace{0, ..., 0}_{s(i-1)}, \underbrace{\boldsymbol{w}^T}_{n_w}, \underbrace{0, ..., 0}_{s(n_x - i)}]^T$$

where s is the upsampling factor. This linear interpolator is

[1]Other names in literature include deconvolution, upconvolution, backward strided convolution and fractionally strided convolution.

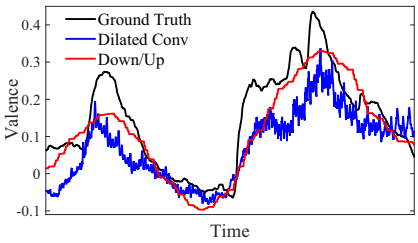

Figure 5: *A visualization of the predictions produced by the two models plotted against ground-truth for a* 40*-second segment.*

able to expand the input vector $\boldsymbol{x}$ to the output vector $\boldsymbol{y}$ with the length of $n_y = s(n_x - 1) + n_w$. Note that the matrix $\boldsymbol{T}$ is nothing but the transposed version of the standard strided convolution transform matrix. Our experiments confirm that the proposed downsampling/upsampling network generates smooth trajectories.

6. Results and Discussion

6.1. Experimental Setup

We build our models using Keras [22] with a Theano backend [23]. We train our models on the training partition of the dataset and use the development partition for early stopping and hyper-parameter selection (e.g., learning rate, number of layers layer size, filter width, $l2$ regularization, dilation factors, downsampling factors). We optimize CCC directly in all setups. We repeat each experiment five times to account for the effect of initialization. The final test evaluation is done by the AVEC 2016 organizers (i.e., we do not have access to test labels). Our test submissions were created by averaging the predictions produced from the five runs.

We report published results from the literature as baselines. Almost all previous works only report their final test results based on multi-modal features. We only show results that are reported on the audio modality in the results tables. We also compare our performance to that of an optimized BLSTM regression model, described in [24]. Our final dilated convolution structure has a depth of 10 layers, each having a width of 32. Our final downsampling/upsampling network contains four downsampling layers, one intermediate layer, and four transposed convolution layers, each having width of 32 for arousal and 128 for valence. We use a downsampling factor of three. We do not splice the input utterances into segments. Instead, we train on full length utterances and use a batch size of one.

6.2. Results

Tables 1 and 2 show the development and test results for arousal and valence, respectively. Each row shows the results for one setup. We only include results from the literature that are based on the speech modality and use "–" to show unreported results.

Both proposed systems show improvements over baseline results by Valstar et al. [2]. Our dilated convolution based system provides improvements of 5.6% and 19.5% over baseline systems for arousal and valence, respectively. Our downsampling/upsampling system provides improvements of 5.1% and 33.9% over baseline systems for arousal and valence, respectively. We report the results we obtain from our BLSTM system to provide a reference point. Our BLSTM system performs well when compared to the baseline results.

The proposed methods outperform BLSTMs and are more efficient to train on long utterances. For instance, given a convolutional network and a BLSTM network with approximately equal number of learnable parameters, one epoch of training on the AVEC dataset takes about 13 seconds on the convolutional network while one epoch of training takes about 10 minutes on the BLSTM network. This suggests that convolutional architectures can act as replacement for recurrent ones for continuous emotion recognition problems.

We show an example 40-second segment of the predictions made by our two networks along with the ground-truth predictions in Figure 5. The figure shows that the predictions produced by the downsampling/upsampling network are much smoother than those produced by the dilated convolution networks. We believe that the structure of the downsampling/upsampling network forces the output to be smooth by generating the output from a compressed signal. The compressed signal only stores essential information that is necessary for generating trajectories, removing any noise components.

Table 1: *Arousal results.*

Method	Dev.		Test	
	RMSE	CCC	RMSE	CCC
Valstar et al. [2]	–	.796	–	.648
Brady et al. [6]	.107	.846	–	–
Povolny et al. [7]*	.114	.832	.141	.682
BLSTM [24]	.103	.853	.143	.664
Dilated	.102	.857	**.137**	**.684**
Down/Up	**.100**	**.867**	**.137**	.681

Table 2: *Valence results.*

Method	Dev.		Test	
	RMSE	CCC	RMSE	CCC
Valstar et al. [2]	–	.455	–	.375
Brady et al. [6]	.132	.450	–	–
Povolny et al. [7]*	.142	.489	.355	.349
BLSTM [24]	.113	.518	**.116**	.499
Dilated	.117	.538	.121	.486
Down/Up	**.107**	**.592**	.117	**.502**

7. Conclusion

We investigated two architectures that provide different means for capturing long-term temporal dependencies in a given sequence of acoustic features. Dilated convolutions provides a method for incorporating long-term temporal information without disrupting the length of the input signal by using filters with varying dilation factors. Downsampling/upsampling networks incorporate long-term dependencies by applying a series of convolutions and max-poolings to downsample the signal and get a global view of the features. The downsampled signal is then used to reconstruct an output with a length that is equal to the uncompressed input. Our methods achieve the best known audio-only performance on the AVEC 2016 challenge.

8. Acknowledgement

This work was partially supported by IBM under the Sapphire project. We would like to thank Dr. David Nahamoo and Dr. Lazaros Polymenakos, IBM Research, Yorktown Heights, for their support.

*Unpublished test results, courtesy of the authors.

9. References

[1] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, *Handbook of emotions*. Guilford Press, 2010.

[2] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[3] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *workshop on automatic speech recognition and understanding (ASRU)*. IEEE, pp. 216–221.

[4] P. Cardinal, N. Dehak, A. L. Koerich, J. Alam, and P. Boucher, "ETS system for AVEC 2015 challenge," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 17–23.

[5] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80.

[6] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 97–104.

[7] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for AVEC 2016 challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 75–82.

[8] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[9] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[10] T. Sercu and V. Goel, "Dense prediction on sequences with time-dilated convolutions for speech recognition," *arXiv preprint arXiv:1611.09288*, 2016.

[11] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2011.

[13] C. Busso, S. Lee, and S. S. Narayanan, "Using neutral speech models for emotional speech analysis." in *Interspeech*, 2007, pp. 2225–2228.

[14] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[15] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[16] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 1520–1528.

[17] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 2528–2535.

[18] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[19] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[20] W. Chen, Z. Fu, D. Yang, and J. Deng, "Single-image depth perception in the wild," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 730–738.

[21] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[22] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[23] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

[24] D. Le, Z. Aldeneh, and E. Mower Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Interspeech, 2017 (to apear)*, 2017.