# LEARNING BANDWIDTH EXPANSION USING PERCEPTUALLY-MOTIVATED LOSS

*Berthy Feng,*[1] *Zeyu Jin,*[1,2] *Jiaqi Su,*[1] *Adam Finkelstein*[1]

[1]Princeton University [2]Adobe Research

## ABSTRACT

We introduce a perceptually motivated approach to bandwidth expansion for speech. Our method pairs a new 3-way split variant of the FFTNet neural vocoder structure with a perceptual loss function, combining objectives from both the time and frequency domains. Mean opinion score tests show that it outperforms baseline methods from both domains, even for extreme bandwidth expansion.

***Index Terms***— Bandwidth expansion, bandwidth extension, audio super resolution, deep learning.

## 1. INTRODUCTION

This paper introduces a deep learning-based method for bandwidth expansion of human speech. The goal of the bandwidth expansion (BWE) problem, also called "bandwidth extension" and "audio super-resolution," is to expand the frequency range of an input audio signal. Its traditional applications are in telephony, where the bandwidth of telephone speech may be limited to below 4 kHz, thus aiming to render muffled speech more intelligible [1].

In the context of newer audio synthesis tasks, such as text-to-speech (TTS) and consumer digital media creation, there arises a need for more extreme BWE, such as to 44.1 kHz or 48 kHz. In WaveNet-like applications, for example, speech is synthesized at a low sampling rate for efficiency reasons [2]. BWE may be applied to synthesized audio to improve the listening experience. In another use case, many consumers record speech on low-bandwidth devices, such as a consumer-grade microphone, and would like higher-resolution versions of their recordings for podcasts or other artistic purposes. In these applications, the input bandwidth might not be as low as that of telephone transmission, but rather around 8 kHz.

Our objective is to super-resolve speech to high-definition audio – in our experiments, we convert 8 kHz to 44.1 kHz, although these are just parameters of the method. By expanding beyond 16 kHz, we emphasize not intelligibility as in traditional BWE, but high perceptual quality and sense of presence in the recording, since the extreme upper bands offer information beyond just speech content, including the finer details of the speaker's voice and environment.

Previous methods for BWE have focused on expanding up to 16 kHz, and most operate in the frequency domain. With the introduction of WaveNet [2] and SampleRNN [3], two waveform generation models that operate directly on the input waveform, recent work has explored BWE directly in the time domain.

In this paper, we propose a method for BWE that is both waveform-based and perceptually motivated. We introduce the three-way split summation FFTNet architecture for bandwidth expansion as well as a perceptual loss to encourage realistic-sounding output. In subjective and objective comparisons, our approach outperforms state-of-the-art baseline methods that work in the time and frequency domains.

## 2. RELATED WORK

Recent work in this area has focused on deep learning-based approaches for artificial BWE [4, 5, 6, 7, 8, 9, 10]. In this section, we review the two broad approaches to BWE and highlight recent deep learning-based work.

### 2.1. Frequency domain

Most approaches to speech BWE are based on the source-filter model of human speech [1, 11, 12]. By the source-filter model, human speech is first produced as a periodic signal by the larynx, or voice box. The larynx signal is then shaped by the vocal tract, and this shaping defines the distinctions between human voices. Approaches based on this model estimate (1) the upper-band (UB) residual signal and (2) the UB spectral envelope (more challenging). Classical methods generate the spectral envelope by codebook mapping [10], Gaussian mixture models, and hidden Markov models. More recent approaches use deep neural networks to classify among pre-trained UB envelopes or directly estimate the UB envelope by regression [4, 5, 6, 7, 8, 9].

Schmidt et al. [13] follow the source-filter model in the upsampling process of the narrowband (NB) signal. They transform the NB waveform to the frequency domain by FFT and then zero-pad the signal to their target bandwidth. They generate the higher frequency content via Intelligent Gap Filling [14], and the resulting magnitude spectrum is fed into a deep neural network (DNN), consisting of a series of convolutional layers followed by LSTM layers. The DNN estimates the energies of the wideband (WB) signal.

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

Li et al. [15] encourage even more realistic sounding output by training their model with an adversarial loss. They propose a four-layer DNN that generates UB energies and line spectral frequencies (LSFs), which further shape the periodic signal. Their network is pre-trained with a MSE loss and then trained against an adversarial network.

Other methods directly estimate the UB magnitudes in the frequency domain and recover the phase from the input signal. Li et al. [16] propose a DNN that maps input log spectrum power to output log spectrum power and suggest correct phase recovery as future work. Addressing the redundancy of generating magnitude and phase information directly from the NB spectrogram, Abel et al. [17] propose first estimating a lower-dimensional cepstral representation.

Previous methods were designed for traditional BWE, expanding to 16 kHz at most. Our method was designed for more extreme BWE, and our experiments expand from 8 kHz to 44.1 kHz. Our method offers an alternative to spectrogram-based approaches for traditional and extreme BWE.

### 2.2. Time domain

WaveNet [2] introduced a fully autoregressive way of generating raw audio waveforms, encouraging BWE methods that operate directly on the input waveform. Kuleshov et al. [18] propose a simple encoder-decoder-type convolutional neural network (CNN) that takes as input the NB waveform and outputs the WB waveform prediction. Their method uses a series of downsampling blocks, followed by a series of upsampling blocks, and it produces results that are intelligible but not necessarily of the best perceptual quality.

SampleRNN [3], the recurrent version of WaveNet for audio waveform generation, inspired Ling et al. [19] to apply hierarchical recurrent neural networks (RNNs) to BWE. They implement a hierarchical network, where the highest-level RNN (LSTM/GRU) takes as input eight samples at a time, the mid-level RNN four samples, and the low-level multilayer perceptron (MLP) one sample. By stacking RNNs with various perceptive fields, they allow the network to process short-term and long-term information from the input signal. As with any LSTM-based approach, however, their method encounters the problem of oversmoothing, especially in noised parts of speech (i.e., rather than learning a phoneme-dependent distribution of the UB energies during noised parts, the network predicts a uniform distribution of energy across the UB frequency bins).

Aiming to merge the time and frequency domains, Lim et al. [20] propose the Time-Frequency Network (TFNet). Our work introduces a simple network that operates in the time domain and is supervised with a perceptual loss calculated in the frequency domain. Our network builds on the FFTNet architecture introduced by Jin et al. [21], which offers a simple yet powerful model for audio generation and vocoding.

## 3. METHOD

### 3.1. FFTNet

Offering an alternative to WaveNet for audio synthesis, FFTNet conditions the prediction of sample $x_n$ on samples $(x_0, x_1, ..., x_{n-1})$ in a neural network architecture inspired by the Cooley-Tukey Fast Fourier Transform (FFT). Similar to WaveNet, it predicts one sample at a time based on $N$ previously generated samples. Given $N$ input samples, FFTNet splits them into two halves, each of size $N/2$. Then each half is transformed using a different 1x1 convolution and activation and then added together to form a vector of size $N/2$. The resulting vector is processed using another 1x1 convolution and then goes through the same split, transformation, and summation to produce a vector of size $N/4$. Iteratively, we will end up with an output of size 1, which predicts the value for the next sample following the $N$ input samples. Because of this repeated split and summation, we can call it a two-way split summation network.

### 3.2. Three-way split summation FFTNet

For autoregressive waveform generation (where one sample in the input does not directly correspond to one sample in the output), the originally proposed two-way split summation architecture of FFTNet suffices. In our task, however, each input sample should correspond to one output sample (since we upsample the NB signal to be the same length as the WB signal for training). We propose a three-way split summation FFTNet architecture, such that the splits are symmetric about the input sample. Fig. 1 shows a diagram of the architecture.

For an input time series with size $n$, we split the input sequence into thirds of size $n/3$ each. Just as in two-way split summation FFTNet, we perform a 1x1 convolutional transformation on each split and then sum the results.

$$z = W_L * x_L + W_C * x_C + W_R * x_R$$

where $W_L$, $W_C$, and $W_R$ are the weights of the kernel applied to the left, center, and right splits, respectively. The output $z$ is of size $n/3$ and can be further transformed with 1x1 convolution and activation before being fed into the next layer. Using the same three-way split and summation scheme, the next layer will reduce the output size to $n/3/3$ and so on for the following layers until we have a one-sample prediction. To increase non-linearity, 1x1 convolution may be replaced with Gated Linear Units [22]; one or more 1x1 convolutions may be introduced after the summation; and skip layers may be added. In this work, we use the aforementioned GLU and one additional 1x1 convolution (both have 256 channels) after summation followed by ReLU activation. The middle split $x_C$ is added to the input of the next three-way split summation FFTNet layer to form skip connections.

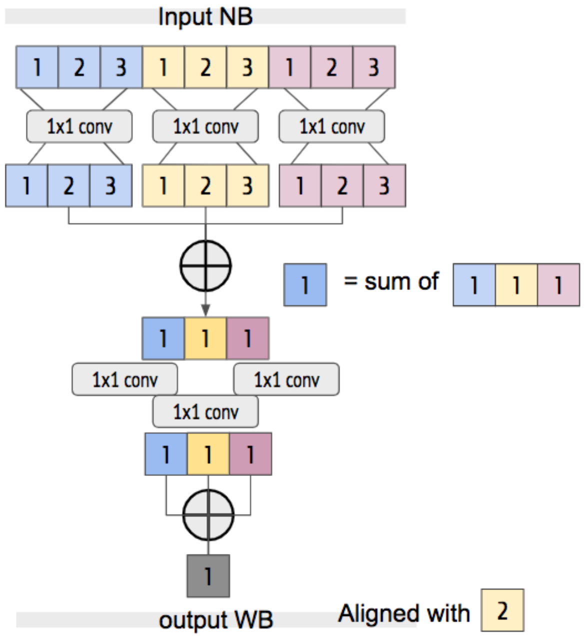


**Fig. 1**: Three-way split summation FFTNet. Starting from an input NB waveform, we iteratively perform a 1x1 convolutional transformation on each split and then sum the results, producing an output WB waveform.

Similar to WaveNet [23], we can stack several FFTNets into deeper networks. To do this, we can apply the same FFTNet of input $N$ for every consecutive N samples in a $(2N-1)$-sized sequence. This will produce $N$ outputs. Then these $N$ outputs can be further processed using another three-way FFTNet to produce one sample. This architecture can be implemented using dilated convolution with 1x3 kernels and dilation of $N/3$ to replace three-way split and summation. Such a network has a smaller number of layers than feed-forward WaveNet, as the receptive field is 3 to the number of layers instead of 2.

### 3.3. Perceptual loss

The second contribution of our paper is the addition of a perceptual loss to encourage perceptually motivated BWE. Our perceptual loss $L_P$ is defined as the L1 loss of the log mel-spectrogram bewteen the predicted waveform and that of the WB waveform, similar to the loss function used in parallel WaveNet [24]:

$$L_P = |\log(\text{MELSPEC}(\hat{y})) - \log(\text{MELSPEC}(y))|$$

The intuition behind this loss is that the spectrogram based on the mel scale is associated with human hearing. Our final loss function consists of two parts: An L1 loss between the predicted waveform and the WB waveform captures the overall shape of the waveform, making sure that the lower frequencies are intact. The WB waveform and the predicted waveform are both passed through STFT to produce a spectrogram that is further transformed into log mel-spectrogram based on triangular filters ranging from 2KHz and 22.05KHz. We choose this frequency range to capture the missing higher band. The L1 distance between these two log mel-spectrograms is the second part of the loss.

## 4. EVALUATION

For our experiments, we implemented a deep FFTNet architecture and trained on the Device and Produced Speech (DAPS) Dataset [25]. The model is composed of two stacks, each of six consecutive FFTNet structures. For the single-speaker case, we trained/tested on speaker F1 (female) and separately on speaker M1 (male). We trained on scripts 1-4 of each speaker, holding out script 5 for testing. We also ran experiments in multi-speaker settings, where we trained a model on all speakers except F1 and M1 and then tested on the held-out voices.

We compare our method to state-of-the-art baselines from both time and frequency domains. For the time domain, we implemented the waveform-based approach proposed by Kuleshov et al. [18]. This baseline is a convolutional neural network that consists of a series of upsampling blocks followed by a series of downsampling blocks. For the frequency domain, we implemented the approach proposed by Li et al. [16]. (described in Section 2). Listening samples and experimental results can be found at our project website.[1]

### 4.1. Subjective evaluation

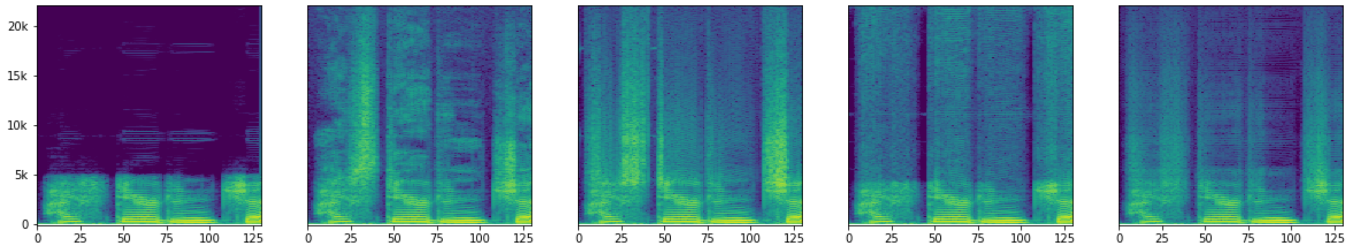
Following an experimental protocol similar to that of Jin et al. [21], we conducted Mean Opinion Score (MOS) tests [26] comparing three methods:

1. OUR: deep three-way split summation FFTNet, trained on L1 and perceptual loss
2. KUL: waveform-based DNN (Kuleshov et al. [18])
3. SPEC: spectrogram-based DNN (Li et al. [16])

| | F | M | cross-F | cross-M |
|---|---|---|---|---|
| Real | 4.54 | 4.68 | 4.47 | 4.36 |
| OUR | **3.39** | **3.74** | **3.04** | **3.70** |
| KUL | 2.95 | 3.29 | 2.86 | 3.06 |
| SPEC | 3.01 | 3.58 | 2.97 | 3.02 |

**Table 1**: MOS test results. OUR method is compared to a waveform-based method (KUL) and a spectrogram-based method (SPEC). Real is the ground-truth WB waveform.

[1] https://pixl.cs.princeton.edu/pubs/Feng_2019_LBE/



(a) NB input (b) WB ground truth (c) OUR (d) KUL (e) SPEC

**Fig. 2**: Comparing log mel-spectrograms. Starting from (a) NB input, (b) WB ground truth can be compared with the generated output of three methods: (c) our method, (d) a waveform-based CNN [18], and (e) a spectrogram-based DNN [16]. Our method avoids oversmoothing the UB frequencies and most closely resembles the appearance of the ground truth spectrogram.

We trained and tested each method in single-speaker and multi-speaker settings. In the single-speaker setting, we trained and tested on F1 (`F`) and separately trained and tested on M1 (`M`). In the multi-speaker setting, we trained on all voices, except F1 and M1, and tested on F1 (`cross-F`) and M1 (`cross-M`).

Our subjects were recruited via Amazon Mechanical Turk, a micro-task platform shown to be as reliable for crowdsourcing experiments as for lab-based studies [27]. Subjects rated the quality of example audio files on a Likert scale of 1-5, where each HIT (human intelligence task) corresponded to one of the four setups and included results from all three compared methods (plus reference and input waveforms, to check if the HIT was valid). The presented MOS results are based on 318 valid HITs for `F`, 186 valid HITs for `M`, 402 valid HITs for `cross-F`, and 384 valid HITs for `cross-M`. Our method outperformed baseline methods in all four scenarios and performed better overall on the male voice.

In subjective evaluations of the generated spectrograms, we found that both baselines are prone to oversmoothing the energy in the upper frequency bins. Kul also results in a clear line between the NB frequency bins of the input and the WB bins of the output, as there is a significant drop in energy across this threshold. We found that our method produces more realistic looking spectrograms.

| | F | | M | |
|---|---|---|---|---|
| | LSD | SNR | LSD | SNR |
| Spec | **4.80** | 11.53 | **3.92** | 13.93 |
| Kul | 9.24 | **22.65** | 7.91 | **19.40** |
| Our | 6.32 | 20.40 | 4.40 | 16.33 |

(a) Results based on networks trained on a single-speaker and tested on the same speaker.

| | cross-F | | cross-M | |
|---|---|---|---|---|
| | LSD | SNR | LSD | SNR |
| Spec | **5.01** | 11.10 | **4.18** | 11.19 |
| Kul | 9.46 | **22.78** | 9.37 | **17.25** |
| Our | 5.95 | 18.36 | 4.83 | 15.14 |

(b) Results based on network trained on multiple speakers and tested on unseen speakers.

**Table 2**: Objective evaluation results.

### 4.2. Objective evaluation

We compare the output of each method to ground truth based on objective metrics of audio quality: signal-to-noise ratio (SNR) and log spectral distance (LSD). The waveform-based baseline [18] should perform well by SNR, while the spectrogram-based baseline [16] should optimize LSD. We expect our method, which uses supervision in both domains, to perform relatively well by both measures, since the L1 loss encourages waveform-level accuracy and the perceptual loss encourages spectral accuracy. We find that Kul indeed performs best by SNR, while Spec performs best by LSD. Our method consistently ranks in between both baselines by both measures, demonstrating that it achieves a balance between waveform-level optimization and spectrogram-level optimization.

## 5. CONCLUSION

We introduce a waveform-based method for extreme bandwidth expansion that uses a deep three-way split summation FFTNet architecture. We train our network using a perceptually motivated loss, which encourages realistic output in the spectral domain as well as time domain. Experiments demonstrate that our method generates more perceptually convincing wideband speech than a state-of-the-art method that operates only in the frequency domain, and it beats the standard waveform-based baseline. Future work may consider alternative approaches to designing a perceptual loss, such as an adversarial loss. It may also consider other applications of our method, such as spectral hole filling and other frequency restoration tasks that arise from audio editing and synthesis.

## 6. REFERENCES

[1] Bernd Iser and Gerhard Schmidt, "Bandwidth extension of telephony speech," in *Speech and Audio Processing in Adverse Environments*, pp. 135–184. Springer, 2008.

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[3] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio, "SampleRNN: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[4] Kehuang Li, Zhen Huang, Yong Xu, and Chin-Hui Lee, "Dnn-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Yingxue Wang, Shenghui Zhao, Wenbo Liu, Ming Li, and Jingming Kuang, "Speech bandwidth expansion based on deep neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] Yu Gu, Zhen-Hua Ling, and Li-Rong Dai, "Speech bandwidth extension using bottleneck features and deep recurrent neural networks," in *Interspeech*, 2016, pp. 297–301.

[7] Bin Liu, Jianhua Tao, Zhengqi Wen, Ya Li, and Danish Bukhari, "A novel method of artificial bandwidth extension using deep architecture," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[8] Johannes Abel, Maximilian Strake, and Tim Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Acoustic Signal Enhancement (IWAENC), 2016 IEEE International Workshop on*. IEEE, 2016, pp. 1–5.

[9] Juho Kontio, Laura Laaksonen, and Paavo Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE transactions on audio, speech, and language processing*, vol. 15, no. 3, pp. 873–881, 2007.

[10] Bernd Iser and Gerhard Schmidt, "Neural networks versus codebooks in an application for bandwidth extension of speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[11] John R Deller, John HL Hansen, and John G Proakis, *Discrete-time processing of speech signals*, IEEE New York, NY, USA:, 2000.

[12] Bernd Iser, Gerhard Schmidt, and Wolfgang Minker, *Bandwidth extension of speech signals*, vol. 13, Springer Science & Business Media, 2008.

[13] Konstantin Schmidt and Bernd Edler, "Blind bandwidth extension based on convolutional and recurrent deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 5444–5448.

[14] Konstantin Schmidt and Christian Neukam, "Low complexity tonality control in the intelligent gap filling tool," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 644–648, 2016.

[15] Sen Li, Stephane Villette, Pravin Ramadas, and Daniel J Sinder, "Speech bandwidth extension using generative adversarial networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE, 2018.

[16] Kehuang Li and Chin-Hui Lee, "A deep neural network approach to speech bandwidth expansion," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4395–4399.

[17] Johannes Abel, Maximilian Strake, and Tim Fingscheidt, "A simple cepstral domain DNN approach to artificial speech bandwidth extension," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 5469–5473.

[18] Volodymyr Kuleshov, S Zayd Enam, and Stefano Ermon, "Audio super resolution using neural networks," *arXiv preprint arXiv:1708.00853*, 2017.

[19] Zhen-Hua Ling, Yang Ai, Yu Gu, and Li-Rong Dai, "Waveform modeling and generation using hierarchical recurrent neural networks for speech bandwidth extension," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 26, no. 5, pp. 883–894, 2018.

[20] Teck Yian Lim, Raymond A Yeh, Yijia Xu, Minh N Do, and Mark Hasegawa-Johnson, "Time-frequency networks for audio super-resolution," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 646–650.

[21] Zeyu Jin, Adam Finkelstein, Gautham J. Mysore, and Jingwan Lu, "FFTNet: a real-time speaker-dependent neural vocoder," in *Proc. ICASSP*, 2018.

[22] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[23] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[24] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, Norman Casagrande, Dominik Grewe, Seb Noury, Sander Dieleman, Erich Elsen, Nal Kalchbrenner, Heiga Zen, Alex Graves, Helen King, Tom Walters, Dan Belov, and Demis Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds., Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 3918–3926, PMLR.

[25] Gautham J Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? A dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[26] Anderson F Machado and Marcelo Queiroz, "Voice conversion: A critical survey," *Proc. Sound and Music Computing (SMC)*, pp. 1–8, 2010.

[27] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.