AUDIO-VISUAL SPEECH INPAINTING WITH DEEP LEARNING

Giovanni Morrone[1] *Daniel Michelsanti*[2] *Zheng-Hua Tan*[2] *Jesper Jensen*[2,3]

[1] University of Modena and Reggio Emilia, Department of Engineering ”Enzo Ferrari”, Italy
[2]Aalborg University, Department of Electronic Systems, Denmark
[3]Oticon A/S, Denmark

ABSTRACT

In this paper, we present a deep-learning-based framework for audio-visual speech inpainting, i.e., the task of restoring the missing parts of an acoustic speech signal from reliable audio context and uncorrupted visual information. Recent work focuses solely on audio-only methods and generally aims at inpainting music signals, which show highly different structure than speech. Instead, we inpaint speech signals with gaps ranging from 100 ms to 1600 ms to investigate the contribution that vision can provide for gaps of different duration. We also experiment with a multi-task learning approach where a phone recognition task is learned together with speech inpainting. Results show that the performance of audio-only speech inpainting approaches degrades rapidly when gaps get large, while the proposed audio-visual approach is able to plausibly restore missing information. In addition, we show that multi-task learning is effective, although the largest contribution to performance comes from vision.

Index Terms— speech inpainting, audio-visual, deep learning, face-landmarks, multi-task learning

1. INTRODUCTION

In real life applications, audio signals are often corrupted by accidental distortions. Impulsive noises, clicks and even transmission errors might wipe out audio intervals. The process of restoring the lost information from the audio context is known as *audio inpainting* [1], and, when applied to speech signals, we refer to it as *Speech Inpainting* (SI). Since human speech perception is multimodal, the use of visual information might be useful in restoring the missing parts of an acoustic speech signal. Visual information was successfully used in many speech-related tasks, such as speech recognition, speech enhancement, speech separation, etc. (cf. [2, 3] and references therein), but it has not been adopted for SI yet. In this paper, we address the problem of *Audio-Visual Speech Inpainting* (AV-SI), i.e. the task of restoring the missing parts of an acoustic speech signal using audio context and visual information.

The first audio inpainting works aimed at restoring short missing gaps in audio signals [1, 4, 5, 6]. For inpainting long gaps, i.e., hundreds of milliseconds, several solutions have been proposed. Bahat et al. [7] tried to fill missing gaps using pre-recorded speech examples from the same speaker and Perraudin et al. [8] exploited self-similarity graphs within audio signals. However, the first approach required a different model for each speaker and the second one was less suitable for speech, since it could only inpaint stationary signals. Prablanc et al. [9] proposed a text-informed solution to inpaint missing speech combining speech synthesis and voice conversion models.

Several researchers attempted to solve audio inpainting using deep learning. In [10], a Convolutional Neural Network (CNN) model was used to inpaint missing audio from adjacent context. Other works exploited Generative Adversarial Networks (GANs) to generate sharper Time-Frequency (TF) representations [11, 12]. Recently, Zhou et al. [13] demonstrated that exploiting visual cues improved inpainting performance. However, these approaches only restored music signals, which usually have long term dependencies, unlike speech. Chang et al. [14] and Kegler et al. [15] both tried to generate speech from masked signals with convolutional encoder-decoder architectures. They evaluated their systems on long gaps (about 500 ms), while in our work we aim at inpainting also extremely long segments (until 1600 ms), where additional information, like video, is essential to correctly restore speech signals. A very recent work proposed a two-stage enhancement network where binary masking of a noisy speech spectrogram was followed by inpainting of time-frequency bins affected by severe noise [16].

In this paper, we propose a deep learning-based approach for speaker-independent SI where visual information is used together with the audio context to improve restoration of missing speech. Our neural network models are able to generate new information and they are designed to fill arbitrarily long missing gaps with coherent and plausible signals. In addition, we present a Multi-Task Learning (MTL) [17] strategy where a phone recognition task is learned together with SI. The motivation of the MTL approach lies in previous work, which showed that speech recognition can improve not only speech enhancement [18] (and vice versa [19, 20]), but also speech reconstruction from silent videos [21].

Additional material, which includes samples of the inpainted spectrograms together with the respective audio clips, can be found at the following link: `https://dr-pato.github.io/audio-visual-speech-inpainting/`.

2. METHODS AND MODEL ARCHITECTURES

In this section we provide a formulation of the problem and describe the architectures of the models that we propose. As done in previous work [1], we assume to know *a priori* the location of reliable and lost data and we use this information in the signal reconstruction stage. In general, the models exploit reliable audio context and visual cues to restore missing gaps in speech signals. As audio and video features, we used log magnitude spectrograms and face landmarks motion vectors, respectively. In a recent work, the specific visual features used here have proven to be effective for audio-visual speech enhancement [22].

2.1. Audio-Visual Speech Inpainting Model

Let $x[n]$ denote an observed acoustic speech signal, i.e., speech signal with missing parts, with n indicating a discrete-time index. We refer to the log magnitude spectrogram of $x[n]$ as $X(k, l)$, where k and l are a frequency bin index and a time frame index, respectively. The information about the location of missing portions of the signal is encoded in a binary mask $M(k, l)$, which indicates whether a time-frequency tile of the spectrogram of the observed signal is lost, $M(k, l) = 1$, or reliable, $M(k, l) = 0$. We assume that $X(k, l) = 0$ if $M(k, l) = 1$. In addition, we denote with $V(l)$ a sequence of visual feature vectors, obtained from the resampled visual frame sequence, since acoustic and visual signals are generally sampled at different rates. We define the problem of AV-SI as the task of estimating the log magnitude spectrogram of the ground-truth speech signal, $Y(k, l)$, given $X(k, l)$, $M(k, l)$, and $V(l)$.

In this paper, $Y(k, l)$ is estimated with a deep neural network, indicated as a function, $\mathscr{F}_{av}(\cdot, \cdot, \cdot)$, whose overall architecture is shown in Fig. 1. The audio and video features are concatenated frame-by-frame and used as input of stacked Bi-directional Long-Short Term Memory (BLSTM) units that model the sequential nature of the data [23]. Then, a Fully Connected (FC) layer is fed with the output of the stacked BLSTM units and outputs the inpainted spectrogram $O(k, l)$. To extract the inpainted spectrogram within the time gaps, $O(k, l)$ is element-wise multiplied with the input mask $M(k, l)$. Finally, the fully restored spectrogram, $\hat{Y}(k, l)$, is obtained by an element-wise sum between the input audio context spectrogram, $X(k, l)$, and the inpainted spectrogram. More formally:

$$\begin{aligned} \hat{Y}(k, l) &\triangleq \mathscr{F}_{av}(X(k, l), M(k, l), V(l)) \\ &= O(k, l) \odot M(k, l) + X(k, l) \end{aligned} \tag{1}$$

where $\odot$ is the element-wise product. The model is trained to minimize the Mean Squared Error (MSE) loss, $J_{MSE}(\cdot, \cdot)$, between the inpainted spectrogram, $\hat{Y}(k, l)$, and the ground-truth spectrogram, $Y(k, l)$.

2.2. Multi-Task Learning with CTC

In addition to the plain AV-SI model, we devised a MTL approach, which attempts to perform SI and phone recognition simultaneously. Our MTL training makes use of a Connectionist Temporal Classification (CTC) loss [24] which is very similar to the one presented in [21] for the task of speech synthesis from silent videos. The *phone recognition subtask* block in Fig. 1 shows the phone recognition module. It is fed with the stacked BLSTM units' output and has a linear FC layer followed by a softmax layer which outputs a CTC probability mass function $\hat{\mathbf{l}} = [\mathbf{p}_1(l), \ldots, \mathbf{p}_P(l)]$, with $l \in [1, L]$, where L is the size of the phone dictionary and P is the number of phone labels in the utterance.

The MTL loss function is a weighted sum between the inpainting loss, $J_{MSE}(\cdot, \cdot)$, and the CTC loss, $J_{CTC}(\cdot, \cdot)$:

$$J_{MTL}(Y, \hat{Y}, \mathbf{l}, \hat{\mathbf{l}}) = J_{MSE}(Y, \hat{Y}) + \lambda \cdot J_{CTC}(\mathbf{l}, \hat{\mathbf{l}}), \tag{2}$$

with $\lambda \in \mathbb{R}$, where $\mathbf{l}$ is the sequence of ground truth phone labels. The phone distribution is used to estimate the best phone sequence. We find the phone transcription applying *beam search* decoding [25] with a beam width of 20.

2.3. Audio-only Inpainting Baseline Model

An audio-only baseline model is obtained by simply removing the video modality from the audio-visual model, leaving the rest unchanged. We consider audio-only models both with and without the MTL approach described in section 2.2.

3. EXPERIMENTAL SETUP

3.1. Audio-Visual Dataset

We carried out our experiments on the GRID corpus [26], which consists of audio-visual recordings from 33 speakers, each of them uttering 1000 sentences with a fixed syntax. Each recording is 3 s long with an audio sample rate of 50 kHz and a video frame rate of 25 fps. The provided text transcriptions were converted to phone sequences using the standard TIMIT [27] phone dictionary, which consists of 61 phones. However, only 33 phones are present in the GRID corpus because of its limited vocabulary.

We generated a corrupted version of the dataset where random missing time gaps were introduced in the audio speech signals. Our models are designed to recover multiple variable-length missing gaps. Indeed, for each signal we drew the amount of total lost information from a normal distribution with a mean of 900 ms and a standard deviation of 300 ms. The total lost information was uniformly distributed between 1 to 8 time gaps and each time gap was randomly placed within the signal. We assured that there were no gaps

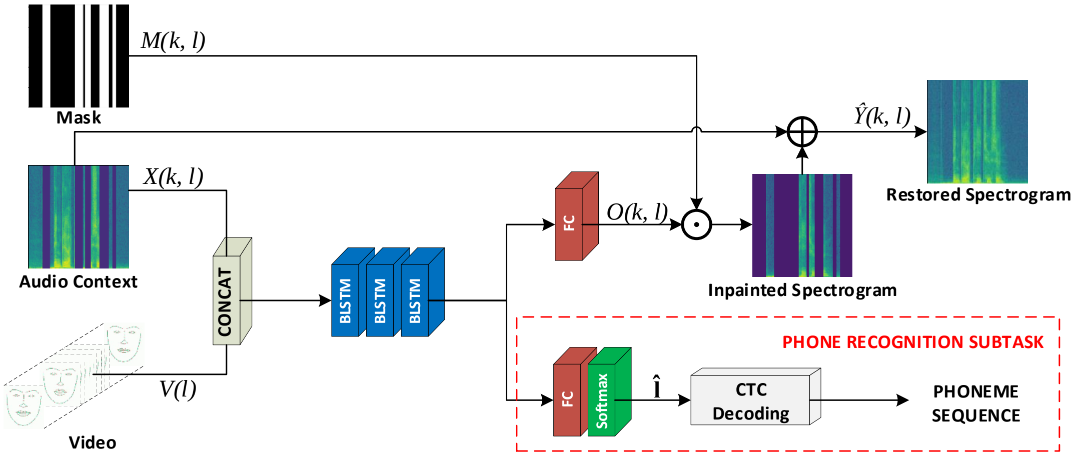

Fig. 1. Overall architecture of the audio-visual speech inpainting system. *CONCAT: frame-by-frame concatenation; BLSTM: Bi-directional Long-Short Term Memory unit; FC: Fully Connected layer; CTC: Connectionist Temporal Classification.*

shorter than 36 ms and the total duration of the missing gaps was shorter than 2400 ms. Similarly to [15], the information loss was simulated by applying binary TF masking to the original spectrograms. The generation process was the same for training, validation and test sets.

Our systems were evaluated in a speaker-independent setting, with 25 speakers (s1-20, s22-25, s28) used for training, 4 speakers (s26-27, s29, s31) for validation and 4 speakers (s30, s32-34) for testing. Each set consists of the same number of male and female speakers, except for the training set which contains 13 males and 12 females. Furthermore, to evaluate the effect of the gap size, we generated additional versions of the test set, each of them containing a single gap of fixed size (100/200/400/800/1600 ms).

3.2. Audio and Video Processing

The original waveforms were downsampled to 16 kHz. A Short-Time Fourier Transform (STFT) was computed using a Fast Fourier Transform (FFT) size of 512 with Hann window of 384 samples (24 ms) and hop length of 192 samples (12 ms). Then, we computed the logarithm of the STFT magnitude and applied normalization with respect to global mean and standard deviation to get the acoustic input features.

The missing phase was obtained by applying the Local Weighted Sum (LWS) algorithm [28] to the restored spectrogram. Finally, we computed the inverse STFT to reconstruct the inpainted speech waveform.

We followed the pipeline described in [22] to extract the video features, i.e., 68 facial landmarks motion vectors. We upsampled the video features from 25 to 83.33 fps to match the frame rate of the audio features.

3.3. Model and Training Setup

The models in Section 2 consist of 3 BLSTM layers, each of them with 250 units. The Adam optimizer [29] was used to train the systems, setting the initial learning rate to 0.001. We fed the models with mini-batches of size 8 and applied early stopping, when the validation loss did not decrease over 5 epochs. The λ weight of the MTL loss, $J_{MTL}(\cdot,\cdot,\cdot,\cdot)$, was set to 0.001. All the hyperparameters were tuned by using a random search and the best configuration in terms of the MSE loss, $J_{MSE}(\cdot,\cdot)$, on the validation set was used for testing.

A	V	MTL	L1 ▼	PER ▼	STOI ▲	PESQ ▲
Unprocessed			0.838	0.508	0.480	1.634
✗			0.482	0.228	0.794	2.458
✗	✗		0.452	0.151	0.811	2.506
✗		✗	0.476	0.214	0.799	2.466
✗	✗	✗	**0.445**	**0.137**	**0.817**	**2.525**

Table 1. Results on the test set. The PER score of uncorrupted speech is 0.069. *A: audio; V: video; MTL: multi-task learning with CTC.*

4. RESULTS

4.1. Evaluation Metrics

We evaluated the system using L1 loss, Phone Error Rate[1] (PER), and two perceptual metrics, STOI [30] and PESQ [31], which provide an estimation of speech intelligibility and speech quality, respectively. While the L1 loss was computed only on the masked parts of the signals, the other three metrics were applied to the entire signals, as it is not possible to perform the evaluation on very short segments. Obviously, PER, STOI, and PESQ show lower sensitivity, when the

[1] PER was obtained with a phone recognizer trained on uncorrupted data. The phone recognizer consists of 2 BLSTM layers (250 units) followed by a FC and a softmax layers.

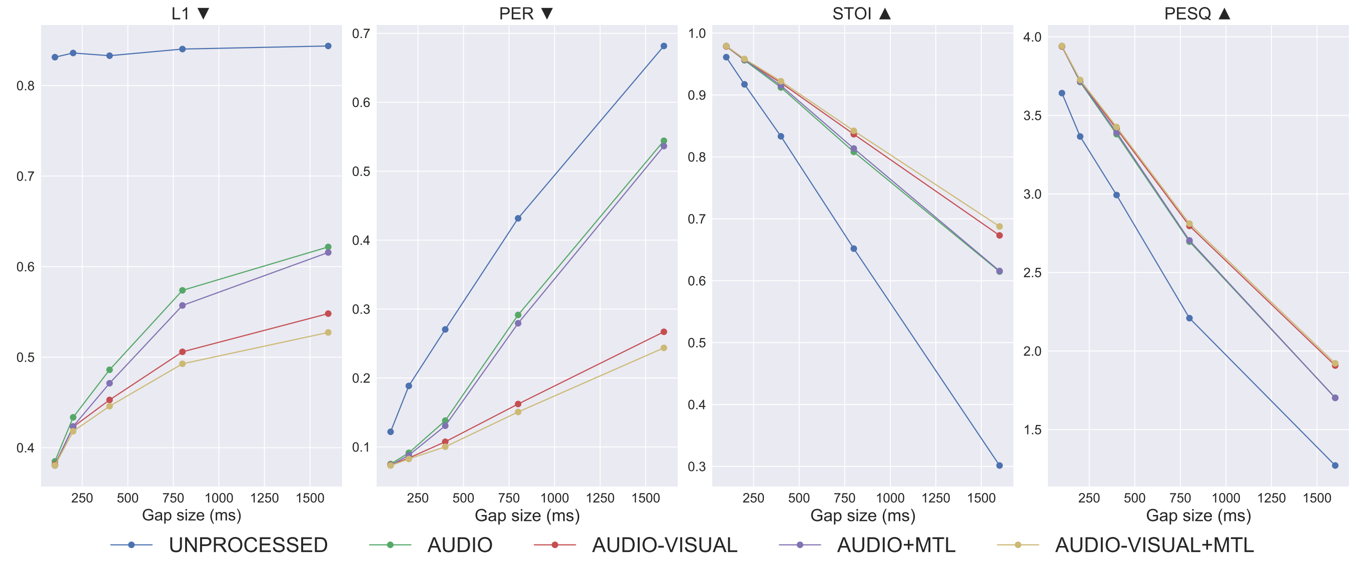

Fig. 2. Effect of gap size on speech inpainting performance.

masked part is small (< 400 ms), since a large fraction of the original signal is unchanged in that case. For L1 and PER, lower values are better, while for STOI and PESQ higher values correspond to better performance.

4.2. Discussion

The evaluation results of the proposed models on the test set are reported in the Table 1. On average, the masking process discarded about half of the original speech information, as confirmed by the PER score of unprocessed data.

Audio-visual models outperform the audio-only counterparts on all metrics, demonstrating that visual features provide complementary information for SI. In particular, the PERs of audio-visual models are lower by a considerable margin, meaning that generated signals are much more intelligible. The improvements in terms of STOI and PESQ are not as large as PER, mainly because the two perceptual metrics are less sensitive to silence than PER. Nonetheless, they are significantly better than the unprocessed data scores confirming the inpainting capability of our models.

The MTL strategy is also beneficial, and results suggest that exploiting phonetic data during the training process is useful to improve the accuracy of SI. However, we observe just a small improvement of the audio-visual MTL model over the plain audio-visual one. This might be explained by the fact that, unlike for the audio-visual system, MTL strategy does not add any additional information at the inference stage.

4.3. Gap size analysis

Table 1 reports the average results using multiple variable-length time gaps, not providing information about how the gap size affects the SI capability of our models. For this reason, we generated other test sets, each of them containing samples with a single time gap of fixed length (100/200/400/800/1600 ms). Fig. 2 shows the inpainting results for each metric on these test sets. As expected, while for short gaps all models reach similar performance, the difference between audio-only and audio-visual models rapidly increases when missing time gaps get larger. The performance of audio-only models drops significantly with very long gaps ($\geq$ 800 ms). Therefore, the audio context does not contain enough information to correctly reconstruct missing audio signals without exploiting vision. In general, audio-only models inpaint long gaps with stationary signals whose energy is concentrated in the low frequencies. On the other hand, audio-visual models are able to generate well-structured spectrograms, demonstrating the benefit that visual features bring to inpaint long gaps. The reader is encouraged to check the difference between audio-only and audio-visual models in the spectrograms provided as additional material (cf. Section 1).

Regarding the models trained with the MTL approach, we can notice a good improvement in terms of L1 loss and PER, even if the contribution is not as high as the one provided by the visual modality.

5. CONCLUSION

This work proposed the use of visual information, i.e., face-landmark motion, for speech inpainting. We tested our models on a speaker-independent setting using the GRID dataset and demonstrated that audio-visual models strongly outperformed their audio-only counterparts. In particular, the improvement due to visual modality increased with duration of time gaps. Finally, we showed that learning a phone recognition task together with the inpainting task led to better results.

6. REFERENCES

[1] Amir Adler, Valentin Emiya, Maria G. Jafari, Michael Elad, Rémi Gribonval, and Mark D. Plumbley, "Audio inpainting," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 922–932, 2012.

[2] Hao Zhu, Mandi Luo, Rui Wang, Aihua Zheng, and Ran He, "Deep audio-visual learning: A survey," *arXiv preprint arXiv:2001.04758*, 2020.

[3] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *arXiv preprint arXiv:2008.09586*, 2020.

[4] Simon Godsill, Peter Rayner, and Olivier Cappé, "Digital audio restoration," in *Applications of Digital Signal Processing to Audio and Acoustics*, pp. 133–194. Springer, 2002.

[5] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka, "Missing data imputation for spectral audio signals," in *Proc. of MLSP, 2009*.

[6] Patrick J. Wolfe and Simon J. Godsill, "Interpolation of missing data values for audio signal restoration using a Gabor regression model," in *Proc. of ICASSP, 2005*.

[7] Yuval Bahat, Yoav Y. Schechner, and Michael Elad, "Self-content-based audio inpainting," *Signal Processing*, vol. 111, pp. 61–72, 2015.

[8] Nathanael Perraudin, Nicki Holighaus, Piotr Majdak, and Peter Balazs, "Inpainting of long audio segments with similarity graphs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 6, pp. 1083–1094, 2018.

[9] Pierre Prablanc, Alexey Ozerov, Ngoc QK Duong, and Patrick Pérez, "Text-informed speech inpainting via voice conversion," in *Proc. of EUSIPCO*, 2016.

[10] Andres Marafioti, Nathanael Perraudin, Nicki Holighaus, and Piotr Majdak, "A context encoder for audio inpainting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2362–2372, 2019.

[11] Pirmin P. Ebner and Amr Eltelt, "Audio inpainting with generative adversarial network," *arXiv preprint arXiv:2003.07704*, 2020.

[12] Andres Marafioti, Piotr Majdak, Nicki Holighaus, and Nathanael Perraudin, "Gacela-a generative adversarial context encoder for long audio inpainting of music," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[13] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang, "Vision-infused deep audio inpainting," in *Proc. of ICCV, 2019*.

[14] Ya-Liang Chang, Kuan-Ying Lee, Po-Yu Wu, Hung-Yi Lee, and Winston Hsu, "Deep long audio inpainting," *arXiv preprint arXiv:1911.06476*, 2019.

[15] Mikolaj Kegler, Pierre Beckmann, and Milos Cernak, "Deep Speech Inpainting of Time-Frequency Masks," in *Proc. of Interspeech*, 2020.

[16] Xiang Hao, Xiangdong Su, Shixue Wen, Zhiyu Wang, Yiqian Pan, Feilong Bao, and Wei Chen, "Masking and inpainting: A two-stage speech enhancement approach for low SNR and non-stationary noise," in *Proc. of ICASSP, 2020*.

[17] Rich Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[18] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. of ICASSP, 2015*.

[19] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. of Interspeech, 2015*.

[20] Luca Pasa, Giovanni Morrone, and Leonardo Badino, "An analysis of speech enhancement and recognition losses in limited resources multi-talker single channel audio-visual asr," in *Proc. of ICASSP, 2020*.

[21] Daniel Michelsanti, Olga Slizovskaia, Gloria Haro, Emilia Gómez, Zheng-Hua Tan, and Jesper Jensen, "Vocoder-based speech synthesis from silent videos," in *Proc. of Interspeech, 2020*.

[22] Giovanni Morrone, Luca Pasa, Vadim Tikhanoff, Sonia Bergamaschi, Luciano Fadiga, and Leonardo Badino, "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments," in *Proc. of ICASSP, 2019*.

[23] Alex Graves, Abdel-Rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of ICASSP, 2013*.

[24] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. of ICML, 2006*.

[25] Alex Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[26] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[27] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, and David S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *STIN*, vol. 93, pp. 27403, 1993.

[28] Jonathan Le Roux, Hirokazu Kameoka, Nobutaka Ono, and Shigeki Sagayama, "Fast signal reconstruction from magnitude STFT spectrogram based on spectrogram consistency," in *Proc. of DAFx, 2010*.

[29] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[31] Anthony W. Rix, John. G. Beerends, Michael P. Hollier, and Andries P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. of ICASSP, 2001*.