



Investigating the Impact of Spectral and Temporal Degradation on End-to-End Automatic Speech Recognition Performance

Takanori Ashihara, Takafumi Moriya, Makio Kashino

NTT Corporation, Japan

takanori.ashihara.vk@hco.ntt.co.jp

Abstract

Humans have a sophisticated capability to robustly handle incomplete sensory input, as often happens in real environments. In earlier studies, the robustness of human speech perception was observed qualitatively by spectrally and temporally degraded stimuli. The current study investigates how machine speech recognition, especially end-to-end automatic speech recognition (E2E-ASR), can yield similar robustness against distorted acoustic cues. To evaluate the performance of E2E-ASR, we employ four types of distorted speech based on previous studies: locally time-reversed speech, noise-vocoded speech, phonemic restoration, and modulation-filtered speech. Those stimuli are synthesized by spectral and/or temporal manipulation from original speech samples whose human speech intelligibility scores have been well-reported. An experiment was conducted on the TED-LIUM2 for English and the Corpus of Spontaneous Japanese (CSJ) for Japanese. We found that while there is a tendency to exhibit similar robustness in some experiments, full recovery from the harmful effect of the severe spectral degradation is not achieved.

Index Terms: locally time-reversed speech, noise-vocoded speech, phonemic restoration, modulation-filtered speech, automatic speech recognition

1. Introduction

Human speech contains rich acoustic redundancy in terms of spectral and temporal information. This redundancy helps humans to recognize the degraded speech robustly, even when parts of it are missing. Various auditory stimuli have been reported for assessing recognition robustness.

Previous studies explored robustness against temporal degradation such as locally time-reversed speech (LTRS) [1] and phonemic restoration (PR) [2, 3, 4]. LTRS is obtained by dividing speech into time segments, temporally reversing each segment, and reconnecting the inverted segments without any spectral distortion. In PR, which is a special case of the continuity illusion, even if a part of speech is removed from the original signal, when noise sufficient for triggering masking is superimposed on the deleted portion, people are capable of perceptually restoring the original continuous signal. Spectrally-degraded approaches also have been used to investigate the importance of spectral cues as in recognition of noise-vocoded speech (NVSS) [5]. NVSS is artificially generated by replacing bandpass-filtered signals by bandpass-filtered noises, while maintaining the amplitude envelope. For directly limiting the information of temporal and spectral modulation, [6] has proposed the modulation-filtering method. This approach is based on extracting the modulation power spectrum (MPS) from two-dimensional Fourier transformation (2D-FFT) of the signal's spectrogram. To restrict the specific modulation frequency, the target modulation area is filtered out in the MPS space and then the filtered MPS is converted back into a waveform to yield the

modulation-filtered speech (MFS). How is certain robustness of human against the distortion achieved? In this paper, we utilize machine speech recognition system to address the question. While the characteristics of perception against the stimuli have been reported in the previous studies, little has been reported on the performance for automatic speech recognition (ASR) comprehensively [7, 8]. Therefore, we began to investigate how the performance of an ASR system is degraded by spectral and/or temporal distortion in a comparison with human intelligibility. Furthermore, we conducted the experiment by using the same model architecture but with different types of training data to explore whether the robustness is acquired through training or innate.

Automatic speech recognition (ASR) system have advanced remarkably by the introduction of deep neural networks (DNNs). Unlike Gaussian Mixture Model (GMM), DNNs trained by using log-mel filterbank outputs (raw waveform in some cases [9]) can perform remarkably well even without any spectral transformation such as Mel-frequency cepstral coefficients (MFCC) additionally. This ability is achieved by the feature learning, in which DNNs are able to automatically acquire internal representations for classification from a large amount of data. Recent end-to-end (E2E) ASR models use a simpler architecture that does not require combinations of different models such as an external pronunciation model and an external language model in comparison with hidden Markov model (HMM)-based ASR system, and hence the direct output of a word or subword sequence from the input speech is possible. In other words, E2E models directly optimize the network parameters from sensory inputs and their transcripts without any additional models/processings. Many studies have proposed structures for E2E models, such as attention-based models [10, 11], connectionist temporal classification (CTC) based models [12, 13], and multi-task learning of the former models (hybrid CTC/Attention) [14, 15]. Recently, Transformer-based models [16, 17, 18] have demonstrated state-of-the-art performance in several benchmarks and now outperform HMM-based ASR systems.

We trained E2E models in the hybrid CTC/Attention with transformer architecture using the TED-LIUM2 [19] for English and Corpus of Spontaneous Japanese (CSJ) [20] for Japanese to make a cross-language comparison. To assess robustness from the aspect of training data, we employ two data augmentation techniques during the training: multi-conditioned training [21, 22] and SpecAugment [23]. The results indicate that while the ASR system offers some robustness to temporal degradation, it fails to match human robustness to severe disruption of spectral cues for both of languages.

2. Method

Here, we introduce 4 types of degraded speech stimuli: LTRS (Section 2.1), NVSS (Section 2.2), PR (Section 2.3), and MFS

(Section 2.4).

2.1. Locally-time Reversed Speech (LTRS)

As shown in Figure 1, LTRS is made by cutting the input speech into segments of fixed length w_{ltrs} , time-reversing each segment, and reconnecting the segments. Note that there are no overlaps or gaps between the reconnected segments. This stimulus maintains the spectral information whereas the temporal components are altered as a function of w_{ltrs} .

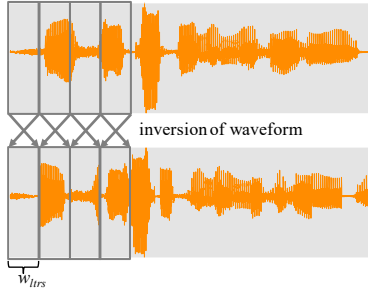


Figure 1: Schematic diagram for the generation of LTRS. The upper panel indicates the original waveform and the lower panel indicates the temporal flipping with width w_{ltrs} segments.

When the illusion is perceived, the listener can integrate the locally time-reversed segments and recover the original speech despite the waveform flipping. Psychophysical experiments indicate that when w_{ltrs} is relatively short (e.g. 25ms), LTRS generally offers sufficiently high intelligibility. As w_{ltrs} becomes longer, intelligibility declines gradually following a sigmoid function, and recovery of the original speech is difficult when w_{ltrs} reaches 100ms or so [24].

2.2. Noise-vocoded speech (NVSS)

In this section, we describe the process used to synthesize NVSS depicted in Figure 2.

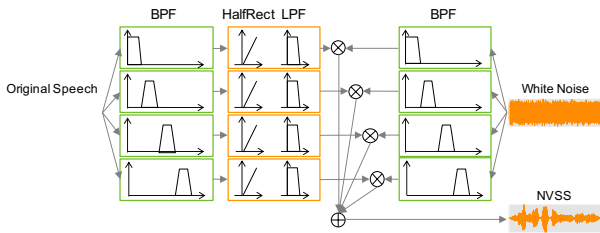


Figure 2: Schematic diagram for the generation of NVSS with the number of bandpass-filter $b_{nvss} = 4$.

First, the original speech waveform is bandpass-filtered (BPF) and divided into a number of frequency bands b_{nvss} . Then, the amplitude envelopes of each frequency band are extracted from the filtered outputs via a half wave rectifier (HalfRect) and lowpass-filter (LPF). The amplitude envelopes are multiplied by a white noise that is BPF with the same frequency band used above. Finally, the amplitude modulated noises are added to yield NVSS. Examples of NVSS are shown in Figure 3. This manipulation roughly maintains the temporal information as the envelope is retained; the spectral information is degraded severely by bandpass noises. The intelligibility increases as a function of b_{nvss} . While the NVSS with $b_{nvss} = 1$ is hard to be intelligible, the intelligibility becomes nearly perfect when $b_{nvss} = 4$ [5].

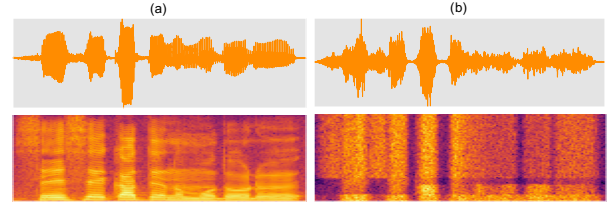


Figure 3: Example of NVSS visualizing waveform (top panel) and spectrogram (bottom panel). (a) Original speech. (b) NVSS with $b_{nvss} = 4$.

2.3. Phonemic Restoration (PR)

Even when the part of a pure tone, speech or music is physically omitted, the brain is able to synthesize the deleted area under certain conditions and humans perceive the result as being continuous. The illusory phenomenon is called continuous illusion, and PR is a special case of it (Figure 4). A typical stimulus for PR is generated by alternately replacing the input speech with louder noise as shown in Figure 4(c). The signal is perceived more natural than the signal replaced by gaps of silence in the same position (Figure 4(b)). In this paper, interruption rate (IR) is introduced to define the gap duration and gap interval [25]. For instance, an IR of 5Hz provide a period of 200ms containing the original and degraded portion with a same duration of 100ms when the duty cycle is 50%. Note that the duty cycle always set 50 % in this paper. From [25], the stimulus becomes intelligible as the IR increased.

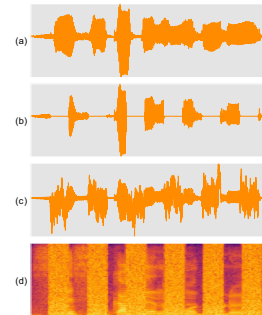


Figure 4: Example of PR signal with IR=5Hz. (a) Original speech. (b) Replacement by silence. (c) Replacement by pink noise with a signal-to-noise ratio of -10dB (d) Spectrogram computed from (c).

2.4. Modulation-filtered Speech (MFS)

In this paper, the method for filtering modulation of spectral and temporal axis is based on [6]. The steps that involved in creating the degraded speech are as follows. (1) The log-spectrogram is computed from the original speech via short-time Fourier transform with Gaussian windows. (2) The modulation amplitude and phase is yielded by two-dimensional discrete Fourier transform (2D-FFT). (3) The specific spectral and temporal modulation filtering is applied to the 2D-FFT result. This filtering operation is realized by multiplying zero at target modulation frequency. (4) The modulation-filtered spectrogram is obtained by inverse 2D-FFT and exponentiation. (5) The Griffin-Lim algorithm [26] is applied to the modulation-filtered spectrogram to yield MFS.

The spectrograms of MFS which is synthesized by LPF of temporal and spectral modulation frequency are exemplified in

Figure 5. In general, [6] reported that the intelligibility gradually declined as the cutoff frequency for spectral or temporal modulation filtering decreased.

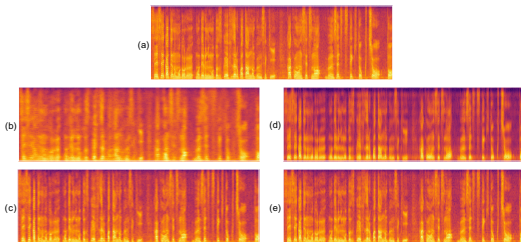


Figure 5: Example of spectrogram of MFS. (a) Original speech. (b, c) MFS with LPF of 6, 12 Hz in temporal modulation respectively. (d, e) MFS with LPF of 1, 4 Hz in temporal modulation respectively.

3. Experiment and Result

3.1. Experimental Setup

3.1.1. ASR configuration and Dataset

The experiments were conducted on the hybrid CTC/Attention model implemented in ESPnet [27], as in the case of the E2E ASR model. The model had the following structure: two convolutional layers with 3×3 filters having the stride of 2 followed by 12 self-attention layers (256 dimensions for attention and 4 attention heads) as an encoder, and 6 self-attention layers as a decoder. During training, the model was optimized by Adadelta with warmup steps of 25000 [16]. The number of epochs was 150. While decoding, the external language model was not utilized, and the beam width was set to 5. The 80-dimensional log-mel filterbank features were extracted from the raw audio with a 25ms window width and 10ms shift. To normalize the features, we subjected the data set to pre-estimated global cepstral mean and variance normalization (CMVN). Validation using the development set was conducted at the end of each epoch, and early-stopping was applied with up to 3 times patience. The evaluation metrics used the word error rate (WER) for TED-LIUM2 task and the character error rate (CER) for CSJ task.

To assess the robustness from the aspect of training data, we also prepared enhanced models based on data augmentation approaches. The technique generates additional training data by transforming the original acoustic data; the new data covers the acoustic variations not present in the original data while maintaining the class labels. In this work, we evaluate the multi-condition trained model (MC), SpecAugment [23] applied model (SA) and their combination model (SA+MC) in addition to the baseline model trained by the original data only.

We trained the MC for the purpose of simulating the noisy situations often encountered in dairy life. MC consists of two main operations, reverberation and noise addition, and is based on the Switchboard recipe of Kaldi toolkit [28]. The former applies artificial reverberation by convolving impulse responses of simulated room [22] and we utilized the simulated small and medium room impulse responses with the same probability of occurrence for all original data. The latter augments the training data by superimposing 3 type of noises from MUSAN [29]: noise, music and babble. The noise, music and babble data were applied with randomly selected signal-to-noise ratios (SNR) from the SNR set of $\{15, 10, 5, 0\}$, $\{15, 10, 8, 5\}$ and $\{20, 17, 15, 13\}$ [dB], respectively. For babble, the number of additional utterances was randomly selected from 3 to 7. As

a result, the amount of MC dataset was five times larger than the original dataset (i.e. original, reverberation, noise, music, and babble), therefore the number of epochs was set to 30 for fairness.

SpecAugment manipulates a feature sequence in three different ways: by time warping, frequency masking, and time masking. Time warping shifts the input feature sequence at a random point on the temporal axis by using a distance chosen from $\text{unif}(0, W)$, where W is the time warp parameter. On the other hand, frequency and time masking apply zero values to consecutive feature sequences over some length of the frequency and time axes, respectively. As the parameters of the last two components, the masking width is chosen from $\text{unif}(0, F)$ with the number of applications m_F of frequency masking and $\text{unif}(0, T)$ with the number of applications m_T of time masking. The method is applied to an input feature under the online policy during training, therefore the epoch size was the same as the baseline (i.e. 150) unlike MC. The hyperparameters of SpecAugment were $W = 5$, $F = 30$, $T = 40$, and $w_F = w_T = 2$. This technique contributes to the improvement for over-fitting/regularization [23]. Furthermore, it is assumed that time masking is equivalent to the manipulation for PR replaced by silence except for the processing space (i.e. feature and waveform space).

We extracted the training and evaluation set from the TED-LIUM2 corpus [19] for English and the CSJ corpus [20] for Japanese based on the recipes of Kaldi toolkit [28]. The both of corpora are consist of spontaneous speech recorded at 16000Hz/16bit. A dataset summary is shown in Table 1.

Table 1: The amount of TED-LIUM2 and CSJ dataset [hours]. Each column shows the data for training, development and test set. CSJ contain 3 test sets (eval1 / eval2 / eval3).

Dataset	train	dev	test
TED-LIUM2	211.1	1.3	2.6
CSJ	515.6	6.5	1.8 / 1.9 / 1.3

For more details of the system and data preparation refer to the recipes for TED-LIUM2 and CSJ of ESPnet¹.

3.1.2. Stimulus configuration

For LTRS, we set the segment length w_{ltrs} to 4 values $\{25, 50, 75, 100\}$ [ms].

In the NVSS experiment, the degraded speech samples were generated using $b_{nvss} = \{1, 2, 3, 4, 5\}$ [bands]. The frequency band of bandpass-filter was set to 0-8000Hz for $b_{nvss} = 1$, 0-600 & 600-8000Hz for $b_{nvss} = 2$, 0-600 & 600-1500 & 1500-8000Hz for $b_{nvss} = 3$, 0-600 & 600-1500 & 1500-2100 & 2100-8000Hz for $b_{nvss} = 4$, 0-600 & 600-1500 & 1500-2100 & 2100-4000 & 4000-8000Hz for $b_{nvss} = 5$. To extract the envelope, the FIR LPF (≤ 16 Hz) of the order of 512 with Hanning window was applied.

As for the conditions of PR, we employed 3 types of IR (2.5, 5, 10 Hz) and SNR of -10dB to replace pink noise against their original signal. A raised cosine ramp of 5 ms was applied to the onset and offset of every target and noises and the ramps of adjacent vibrations overlapped completely.

MFS was synthesized by spectral and temporal modulation filtering described above. In the case of spectral modulation filtering, we adopted 5 types of LPF (1, 2, 4, 8, 16 cycle/kHz);

¹<https://github.com/espnet/espnet>

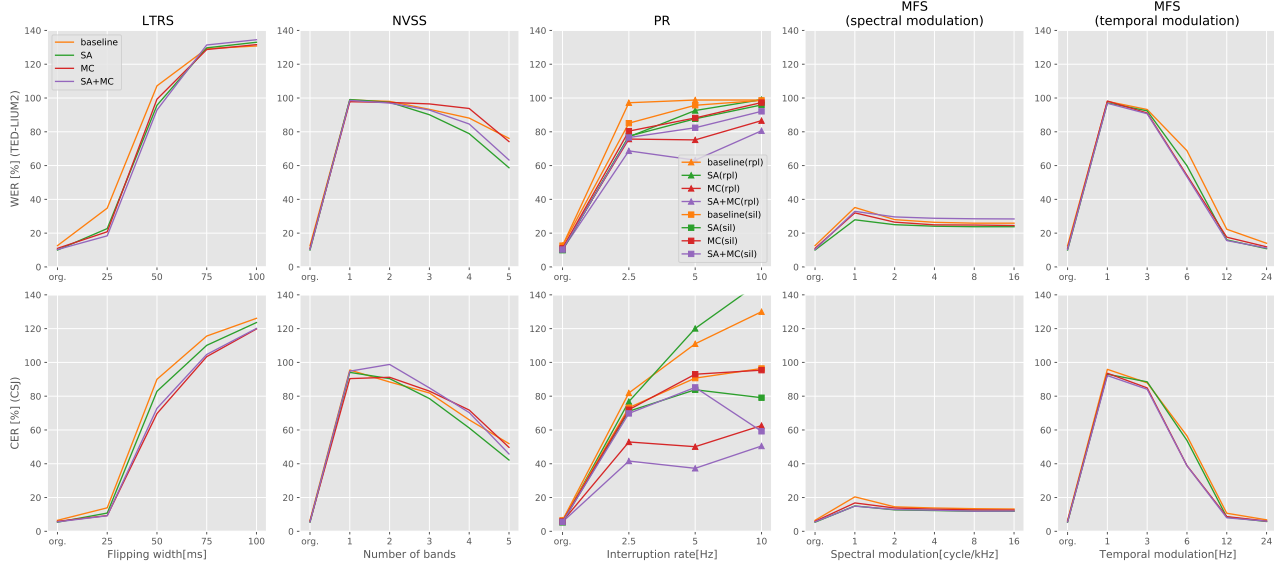


Figure 6: Evaluation results for each stimulus and each models. Top and bottom panels indicate the performance in TED-LIUM2 (WER[%]) and CSJ (CER[%]) dataset, respectively. The columns show the results of LTRS, NVSS, PR and MPS (spectral and temporal modulation LPF). In the PR panels, rpl and sil indicate the test sets based on the replacement by pink noise and silence, respectively.

in the case of temporal modulation filtering, we also utilized 5 types of LPF (1, 3, 6, 12, 24 Hz).

3.2. Experimental Result

Table 2 shows the results achieved with the original speech data. Note that while CSJ corpus is formed from three test sets, we show only the eval1 result because of the space limitation; eval2 and eval3 yielded similar results. We confirmed that the results do not deviate significantly from existing studies [30].

Figure 6 shows the evaluation results for each stimuli and each model. The left end values in each panel show the performance of original data (the same value appears in Table 2). Note that the WER/CER greater than 100% is led by the insertion of words/characters that do not appear in the reference.

4. Discussion

From the results of NVSS (second column in Figure 6), the performance did not recover fully even when frequency band was 5. This result was better than those of existing studies [7] that experimented with GMM-HMM using MFCC as input feature, but still not good enough. On the other hand, the results of MFS with spectral modulation LPF experiments demonstrated less performance degradation than NVSS. Therefore, these results suggest that while the system is somewhat robust to low-resolution (stretching) in the frequency axis of spectrogram, its robustness is greatly degraded when it is completely destroyed such as a noise.

From the results of LTRS (first column in Figure 6) and MFS with temporal modulation LPF, the curves suggest that E2E ASR well matched the human characteristic curve [24, 6].

Table 2: Evaluation results achieved with the original speech data (WER[%] for TED-LIUM2 and CER[%] for CSJ)

Model	TED-LIUM2	CSJ
baseline	12.6	6.5
SA	10.0	5.4
MC	11.0	5.9
SA+MC	10.3	5.6

However, based on the experimental results of PR, all models failed to match the performance recovery of humans reported in [25]. This may be due to the fact that spectral cues are completely missing in PR or replaced by noise, and the robustness against the severe spectral degradation is weak as described above, resulting in insufficient performance recovery. In comparison with human speech recognition, these results indicate that the ASR system has a certain level of robustness to temporal degradation, but its robustness to severe destruction of spectral information still needs to be improved. For example, we will examine a model structure that relies on temporal information attentively when spectral information is not sufficient, or multi-resolution features/architectures that deal with comparatively long timescales [31].

According to the PR results with all IR conditions, the MC and the SA+MC led to a certain gain in performance by noise replacement compared with the baseline and SA model. Therefore, in terms of the acoustic characteristics of training data, it is assumed that the exposure in noisy condition is important to restore the missing segments.

5. Conclusions

The aim of this work is to measure how the performance of an E2E-ASR system is degraded by spectral and/or temporal distortion of speech for which the robustness of human speech perception has been well examined. Our experiments showed that some stimuli led to similar degradation characteristics such as LTRS and MFS with temporal modulation LPF. However, while severe spectral degradation caused worse performance than human performance, as long as low-resolution spectrograms on the spectral axis are provided, a certain level of performance is assured. Furthermore, it is likely that the exposure in daily noise environment is important for perceptual restoration in PR. In the future, we plan to assess human intelligibility using the same data set (i.e. CSJ and TED-LIUM2) to compare the performance directly. In addition, the internal representation of ASR models will be visualized to understand the recognition strategy between the models.

6. References

- [1] K. Saberi and D. R. Perrott, "Cognitive restoration of reversed speech," *Nature*, vol. 398, no. 6730, pp. 760–760, 1999.
- [2] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392–393, 02 1970.
- [3] R. M. Warren, C. J. Obusek, and J. M. Ackroff, "Auditory induction: Perceptual synthesis of absent sounds," *Science*, vol. 176, no. 4039, pp. 1149–1151, 1972.
- [4] M. Kashino, "Phonemic restoration: The brain creates missing speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 318–321, 2006.
- [5] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.
- [6] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLOS Computational Biology*, vol. 5, no. 3, pp. 1–14, 03 2009.
- [7] P. Lin, F. Chen, S. S. Wang, Y.-H. Lai, and Y. Tsao, "Automatic speech recognition with primarily temporal envelope information," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.
- [8] G. Hu, S. C. Determan, Y. Dong, A. T. Beeve, J. E. Collins, and Y. Gai, "Spectral and temporal envelope cues for human and automatic speech recognition in noise," *Journal of the Association for Research in Otolaryngology*, vol. 21, no. 1, pp. 73–87, 2020.
- [9] D. Palaz, R. Collobert, and M. Magimai-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1766–1770, 2013.
- [10] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-End continuous speech recognition using attention-based recurrent NN: First results," *In Proc. Advances in Neural Information Processing System(NIPS): Deep Learning and Representation Learning Workshop*, 2014.
- [11] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *In Proc. Advances in Neural Information Processing System(NIPS)*, pp. 577–585, 2015.
- [12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," *In Proc. International Conference on Machine Learning (ICML)*, pp. 369–376, 2006.
- [13] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, and et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," *In Proc. International Conference on Machine Learning (ICML)*, pp. 173–182, 2016.
- [14] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, Dec 2017.
- [15] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4835–4839, March 2017.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," *In Proc. Advances in Neural Information Processing System(NIPS)*, pp. 5998–6008, 2017.
- [17] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A non-recurrence sequence-to-sequence model for speech recognition," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5884–5888, April 2018.
- [18] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on Transformer vs RNN in speech applications," *In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019.
- [19] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," *In Proc. International Conference on Language Resources and Evaluation(LREC)*, pp. 3935–3939, 2014.
- [20] S. Furui, K. Maekawa, and H. Isahara, "A Japanese national project on spontaneous speech corpus and processing technology," *In Proc. ASR2000 - Automatic Speech Recognition: Challenges for the new Millennium*, pp. 244–248, 2000.
- [21] R. Hsiao, J. Ma, W. Hartmann, M. Karafiát, F. Grézl, L. Burget, I. Szöke, J. H. Černocký, S. Watanabe, Z. Chen, S. H. Mallidi, H. Hermanský, S. Tsakalidis, and R. Schwartz, "Robust speech recognition in unknown reverberant and noisy conditions," *In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 533–538, Dec 2015.
- [22] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, March 2017.
- [23] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2613–2617, 2019.
- [24] M. Ishida, T. Arai, and M. Kashino, "Perceptual restoration of temporally distorted speech in L1 vs. L2: Local time reversal and modulation filtering," *Frontiers in Psychology*, vol. 9, p. 1749, 2018.
- [25] J. D. Saija, E. G. Akyürek, T. C. Andringa, and D. Başkent, "Perceptual restoration of degraded speech is preserved with advancing age," *Journal of the Association for Research in Otolaryngology*, vol. 15, no. 1, pp. 139–148, 02 2014.
- [26] D. Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 8, pp. 804–807, 1983.
- [27] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End speech processing toolkit," *In Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2207–2211, 2018.
- [28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesel, "The Kaldi speech recognition toolkit," *In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.
- [29] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," *arXiv:1510.08484v1*, 2015.
- [30] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs rnn in speech applications," *In Proc. Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 449–456, 2019.
- [31] H. Hermansky and S. Sharma, "Temporal patterns (traps) in asr of noisy speech," *In Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 289–292, 1999.