AUDITORY-BASED DATA AUGMENTATION FOR END-TO-END AUTOMATIC SPEECH RECOGNITION

Zehai Tu, Jack Deadman, Ning Ma, Jon Barker

University of Sheffield, Department of Computer Science, Sheffield, UK
{*ztu3, jdeadman1, n.ma, j.p.barker*}*@sheffield.ac.uk*

ABSTRACT

End-to-end models have achieved significant improvement on automatic speech recognition. One common method to improve performance of these models is expanding the data-space through data augmentation. Meanwhile, human auditory inspired front-ends have also demonstrated improvement for automatic speech recognisers. In this work, a well-verified auditory-based model, which can simulate various hearing abilities, is investigated for the purpose of data augmentation for end-to-end speech recognition. By introducing the auditory model into the data augmentation process, end-to-end systems are encouraged to ignore variation from the signal that cannot be heard and thereby focus on robust features for speech recognition. Two mechanisms in the auditory model, spectral smearing and loudness recruitment, are studied on the LibriSpeech dataset with a transformer-based end-to-end model. The results show that the proposed augmentation methods can bring statistically significant improvement on the performance of the state-of-the-art SpecAugment.

Index Terms— speech recognition, data augmentation, deep neural network, auditory model

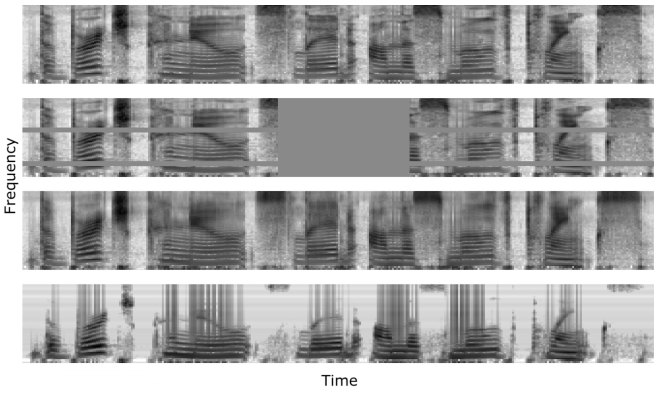

Fig. 1. Augmentation methods applied to the time-frequency representation of a speech signal. From top to bottom, the figures show normalised features of the speech with no augmentation, SpecAugment (SA), spectral smearing (SS), loudness recruitment (LR).

1. INTRODUCTION

In recent years, significant progress has been made in automatic speech recognition (ASR) due to the success of deep learning. Deep learning-based end-to-end models, including Connectionist Temporal Classification (CTC) models [1], attention-based sequence-to-sequence (seq2seq) models [2], and RNN-Transducers [3], outperform conventional hybrid models [4] in various speech recognition tasks. However, the problem of overfitting when training large deep learning models remains a significant issue [5]. Regularization techniques are required. In addition to regularized optimisation approaches, *data augmentation* can also be regarded as an effective data-space solution.

In contrast to most ASR systems, the human auditory system is capable of understanding speech even under poor conditions, including speech in noise and low-quality speech. One of the reasons is that listeners make use of many redundant cues in speech perception; if one set of cues becomes unavailable other strategies can be used. For example, there are studies that show that listeners with high-frequencies hearing loss start adopting different strategies for phoneme classification [6].

Leveraging auditory-based models has always been of interest to ASR researchers. In this work, we employ auditory-based models that simulate various degrees of hearing abilities as part of the data augmentation process. The underlying motivation is that such auditory models could introduce domain knowledge into end-to-end systems, by removing variation from the signal that cannot be heard and thereby encouraging the classifier to pay attention to other (more robust) consistent differences. Such transformations could reduce the chances of spurious correlations existing, which can be useful for wider domain generalisation, especially in noisy conditions where cues are lost due to noise masking.

This study exploits the Cambridge Auditory Group MSBG auditory model [7–10] using a recent open-source implementation [11]. Major auditory phenomena caused by defective hearing abilities, including reduced frequency selectivity and compressed loudness perceptual range, are modelled by spectral smearing (SS) and loudness recruitment (LR), respectively. The behaviours of these two mechanisms are dependent on individual differences in hearing ability. The training speech dataset is augmented with SS and LR by randomly sampling hearing ability related parameters. The augmented training set is then used to fine-tune a transformer-based ASR system that already incorporates the state-of-the-art SpecAugment (SA) data augmentation method. An example time-frequency representation processed by SA, SS, and LR is shown in Fig. 1.

This paper is organised as follows. Section 2 reviews related work on auditory-based ASR and data augmentation for end-to-end systems. Section 3 describes the proposed data augmentation methods and the end-to-end ASR models used in this study. Section 4 presents the database and the experimental setup. The results comparing the performances of the SA baseline with the proposed SS or LR augmentation are presented and discussed in Section 5. Section 6 concludes the paper and presents ideas for future work.

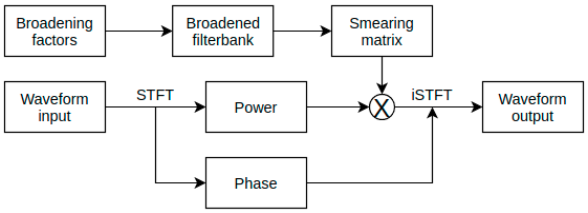

Fig. 2. Procedure of the spectral smearing (SS) process.

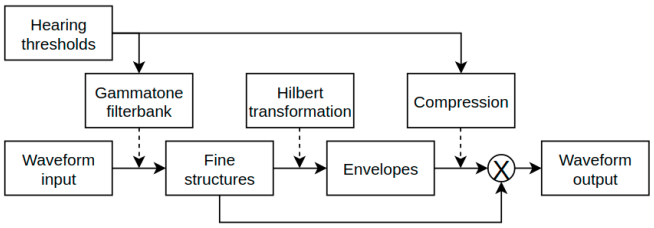

Fig. 3. Procedure of the loudness recruitment (LR) process.

2. RELATED WORK

Exploiting auditory inspired methods for ASR has been long studied. Two early examples are that the auditory perceptually based mel scale filterbank can improve recognition robustness especially in noise [12], and the use of logarithm of speech energy features which approximates the nonlinear dynamic-range compression found in the auditory system. Additionally, a simulated auditory model based on findings from psycho-acoustical and physiological experiments was proposed as an ASR front-end and was found to improve speech-in-noise recognition [13]. Auditory inspired gammatone filterbanks [14, 15] and Gabor filterbanks based on physiologically-motivated modulation frequencies [16, 17] have also been used to improve ASR performance. Recently, a number of works have presented auditory inspired methods for deep neural network based ASR [18–20].

Various data augmentation methods have been proposed for ASR. Vocal tract length perturbation [21] was proposed to augment training utterances by randomising warp factors. Speech perturbation [22, 23] and pitch adjustment [24] were proposed to adjust speed and pitch of the audio. Room impulse response simulation and adding point-source noises were proposed for far-field ASR [25]. Inspired by input dropout, [26] proposed to improve the noise robustness of CNN acoustic models by discarding input features, and [27] proposed to mask time-frequency bins with energy lower than randomised thresholds. Recently, SpecAugment [28] (SA) was proposed to augment speech data by warping spectrograms along the time axis, and masking time and/or frequency bands in the spectral domain. Despite its simplicity, SA showed significant and consistent improvements for end-to-end speech recognition, and has become the standard approach for training state-of-the-art end-to-end ASR models. Methods similar to SA were also used for speech representation learning [29, 30], and proved effective for downstream ASR tasks. Essentially, data augmentation encodes domain knowledge by disturbing the signal in ways 'known' to not change its meaning.

3. METHOD

In this section, the proposed data augmentation mechanisms SS and LR are described. To run the augmentation on the fly in the end-to-end framework for fast computation, the differentiable approximations [31] are used. Both SS and LR take the waveform speech signals as input, and generate waveform variants by randomly sampling hearing ability related parameters. The section concludes with a description of the end-to-end ASR model used for evaluation.

3.1. Spectral smearing

SS smooths the speech spectrum and suppresses details along the frequency axis by broadening the bandwidths of the auditory filters. This is similar to blurring for image data augmentation [32]. In the MSBG model, SS is used to simulate the reduced frequency selectivity in the auditory systems of hearing impaired listeners [7]. Subjective experiments showed that the effect on the intelligibility of spectrally smeared speech was minimal in quiet environments, while the affect would increase in noisy environments as signal-to-noise ratio drops [7, 8]. Speech processed by SS can still preserve key recognition information if not heavily corrupted by noise.

The procedure for SS is shown in Fig 2. An input waveform is first converted into the time-frequency domain with an STFT. The power spectrogram is multiplied with the smearing matrix A, and the phase spectrogram is kept unchanged for later combination with the smeared power spectrogram.

The smearing matrix A_S is the product of the inverse of the normal auditory matrix A_N and broadened auditory matrix A_W, i.e., $A_S = A_N^{-1} A_W$. An auditory matrix A consists of a group of auditory filters, each of which has the the form of roex(p):

$$W(g) = (1 + pg)\exp(-pg), \tag{1}$$

where $W(g)$ is the intensity weighting function describing the filter shape, g is deviation from the centre frequency f_c divided by f_c, and p is the sharpness of the factor, which consists of lower side p_l and upper side p_u and where p is computed as

$$p = \frac{4 f_c}{24.7 \times (0.00437 f_c + 1) r}. \tag{2}$$

p_l and p_u are computed with lower broadening factor r_l and upper broadening factor r_u. Finally, each auditory filter is calibrated by being divided by $(0.00437 f_c + 1)(r_l + r_u)/2$. For normal hearing auditory filters, r_l and r_r are both 1. For the purpose of data augmentation, random smearing matrices are generated by sampling (r_l, r_u) for broadened auditory filters—see Section 4.3 for details.

3.2. Loudness recruitment

LR compresses amplitudes of different frequency bands according to the corresponding hearing losses. In the MSBG model, it is used to simulate the observation that the response of a damaged cochlea to low-level sounds is smaller than that of a healthy one, while the response to high-level sound is roughly the same [9]. Hearing impaired listeners can usually understand clean speech at adequate sound levels. By generating different hearing loss parameters, LR can apply random suppression to different frequency bands. Therefore, LR could introduce randomness to the relative levels of different frequency bands while conserving key recognition information.

The procedure for LR is shown in Fig. 3. Given the input waveform signal, a group of mel scale gammatone filters are first used to extract fine structure $x(n)^{(i)}$ in the time domain. Each filter $h^{(i)}$ can

be expressed as

$$h^{(i)}(t) = A^{(i)} t^{\left(N^{(i)}-1\right)} e^{-2\pi b^{(i)} t} \cos\left(2\pi f^{(i)} t\right), \quad (3)$$

where f^i is the centre frequency, $b^{(i)} = 1.019 \times 24.7 \times (0.00437 f_c + 1)$ is the bandwidth, N is the order of gammatone filters and is set to 4 in this work. $A^{(i)}$ is the normalisation coefficient. The fine structure signals are then aligned and used to estimate the envelope $E^{(i)}$ via a Hilbert transformation. The envelopes are then smoothed with a low pass filter and used for compression. The waveform output y is recruited as

$$y(n) = \sum_{i=1}^{I} \left(\frac{E^{(i)}(n)}{E_\theta} \right)^{\left(\frac{\theta}{\theta - \mathrm{HL}^{(i)}} - 1\right)} x^{(i)}(n), \quad (4)$$

where $\mathrm{HL}^{(i)}$ is the hearing threshold at the centre frequency, θ is the maximal loudness threshold set as 105 dB, and E^θ is the corresponding envelope magnitude. A number of audiograms, i.e., the tables recording hearing thresholds at various frequencies, are sampled for the purpose of data augmentation, and details are described in Section 4.3.

3.3. End-to-end ASR model

Transformer architectures [33] have been widely used for end-to-end ASR and have achieved impressive results. The model used in this work consists of a transformer encoder, decoder, and a convolutional neural network (CNN) based front-end for better utilisation of global context [34]. Both the encoder and decoder consist of multiple transformer blocks, whose core component is the multi-head attention mechanism. For each time step, given the query $Q \in d_k$, key $K \in d_k$, and value $V \in d_v$ projected from the input features, the attention mechanism is expressed as

$$\textit{Attention}(Q, K, V) = \textit{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V. \quad (5)$$

The multi-head attention implements the attention mechanism h times in parallel. For each attention computation, the projection matrices from the input features to the queries, keys, and values are different. The concatenation of h attention heads is at last multiplied by a linear projection matrix. The encoder transformer blocks are built on a multi-head attention, with other operations including residual connections, layer normalisation, and a feedforward network. Based on the encoder transformer block structure, the decoder transformer blocks insert another multi-head attention over the output of the encoder stack, and apply masking so that the predictions only depend on previous decoded outputs.

The joint CTC-attention mechanism [35], which combines the idea of CTC [1] and attention-based seq2seq [2] within the multi-task learning framework, is used for the optimisation. The core idea of CTC is to leverage intermediate label representation by allowing the repetition of labels and adding a special blank label. The CTC loss can be computed as

$$\mathcal{L}_{CTC} = -\log\left(\sum_{\pi \in \beta^{-1}(l)} \prod_{m=1}^{M} P(z^m_{\pi_m}) \right), \quad (6)$$

where β is the many-to-one mapping function that removes the repeated intermediate labels and blank labels, l is the target label sequence, π_m is the intermediate label sequence including also blank labels, and $P(z^m_{\pi_m})$ is the probability of the observing network output intermediate label π_m at time m. The seq2seq training scheme is used with the Kullback-Leibler divergence loss,

$$\mathcal{L}_{KL} = \sum_u P(z_u)(\log P(z_u) - \log P(\hat{z}_u)), \quad (7)$$

where z_u and $\hat{z}_u$ are the u-th ground truth label and the predicted tokens, respectively. Label smoothing is applied for the seq2seq training. The CTC loss is computed on the encoder output, and the seq2seq uses the decoder output for loss computing. During training, the overall multi-task learning loss is then computed as

$$\mathcal{L}_{MTL} = \lambda \mathcal{L}_{CTC} + (1 - \lambda)\mathcal{L}_{KL}, \quad (8)$$

where λ is a predefined weighting coefficient.

Table 1. Maximal r_l and r_r for SS of different hearing impairment degrees.

	$r_l^{\max}$	$r_u^{\max}$
Mild	1.1	1.6
Moderate	1.6	2.4
Severe	2.0	4.0

Table 2. Highest hearing thresholds at different frequencies for LR of different hearing impairment degrees.

	250 Hz	500 Hz	1 kHz	2 kHz	4 kHz	6 kHz
Mild	10 dB	10 dB	10 dB	15 dB	30 dB	40 dB
Moderate	20 dB	20 dB	25 dB	35 dB	45 dB	50 dB
Severe	55 dB	55 dB	55 dB	65 dB	75 dB	80 dB

4. EXPERIMENTS

4.1. Database

We evaluate our proposed augmentation method on the LibriSpeech database [36]. All 960 hours of labeled utterances are used for training the ASR models. The word error rates (WER) on *test-clean* and *test-other* are reported. The difference between the *test-clean* and *test-other* sets is the quality of the utterances, with the quality of *test-clean* being higher.

4.2. End-to-end ASR model

The SpeechBrain [37] LibriSpeech ASR transformer recipe is used to build the ASR model. 80-channel filterbank features are extracted as the input with a 25 ms window with a stride of 10 ms. The front-end context CNN consists of three 2D convolutional layers with kernel size of (3, 3, 1), stride size of (2, 2, 1), and out channel size of (128, 256, 512). The encoder and the decoder consist of 12 and 6 transformer blocks, respectively. For each transformer block, the number of attention heads is 8, and the size of the feedforward layer is 3072. The GELU activation function is used within the transformer blocks. The loss weighting coefficient λ is set to 0.4.

To conserve computational resources, training of all models starts from the LibriSpeech model released by SpeechBrain[1], i.e., we finetune models for a further 15 epochs, rather than train from

[1] huggingface.co/speechbrain/asr-transformer-transformerlm-librispeech

Table 3. The WERs of the ASR models optimised with SA only, SA + SS, and SA + LR, with different degrees of hearing impairment (SA: SpecAugment, SS: spectral smearing, LR: loudness recruitment). Experiments of optimisation with SA only, and the best system (SA + LR with moderate hearing impairment) are run five times independently, and the standard errors of the WERs are shown.

		test-clean	*test-other*
SA		3.36% ± 0.01%	8.12% ± 0.03%
SA + SS	mild	3.32%	8.07%
	moderate	3.38%	8.05%
	severe	3.32%	7.96%
SA + LR	mild	3.27%	7.96%
	moderate	3.28%	7.74%
	severe	3.29%	7.83%
SA + LR	moderate	3.28% ± 0.01 %	7.77% ± 0.03 %

scratch. For the proposed methods, a half of the input signals within a batch are augmented during finetuning. For fair comparison, baseline performances are obtained with further finetuning using the same number of epochs but using only the baseline SpecAugment.

4.3. Experimental setup

SA is applied as the data augmentation method for the baseline models. The released model from SpeechBrain is trained with SA, and SA is also applied when finetuning. The filterbank features are masked along both time and frequency axis with maximal two bins for each axis. The bin widths are sampled up to 30 and 40 for frequency axis and time axis, respectively. Time warping is applied with two-dimensional bicubic interpolation. To gain more confident results, the baseline experiments are repeated five times independently with different random seeds, i.e., five models are finetuned for 15 epochs from the SpeechBrain released model with SA only.

To evaluate the proposed data augmentation methods, we finetuned the ASR models with SS and LR separately, both together with SA. Three parameter settings are used for both SS and LR simulating three hearing impairment degrees: mild, moderate, and severe.

For SS, pairs of $r_l \in [1.001, r_l^{\max})$, and $r_r \in [r_l, r_u^{\max})$ are sampled. The values of $r_l^{\max}$ and $r_u^{\max}$ for different hearing impairment degree settings are shown in Table 1. For LR, the randomly sampled audiograms are represented by the hearing thresholds HL at [250, 500, 1000, 2000, 4000, 6000] Hz, and each threshold HL^f at frequency f is sampled in the range of $[\max(\text{HL}^{f'} | f' < f), \text{HL}^f_{max})$. The highest hearing thresholds HL^f_{max} at these frequencies for different hearing impairment degrees are shown in Table 2. The remaining parameters within SS and LR all follow the MSBG hearing impairment simulator. In the preliminary experiment, ASR models are optimised with the six proposed augmentation candidates, i.e., three for both SS and LR. And for the final results, the method with the overall best performance in the preliminary experiment is run five times and compared with the baselines.

5. RESULTS AND DISCUSSIONS

The results on LibriSpeech *test-clean* and *test-other* are shown in Table 3 and visually presented in Figure 5. As SA is arguably the most successful data augmentation method for end-to-end ASR, it is used as the baseline. Results show that the proposed augmentation methods can further improve on the performance of SA.

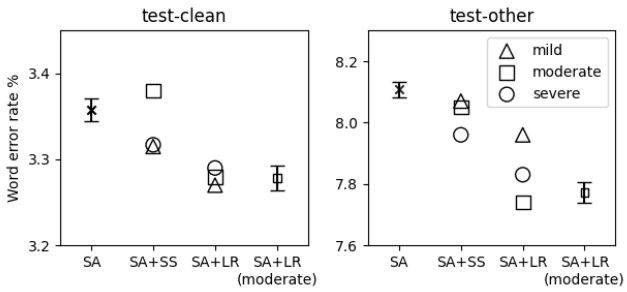

Fig. 4. A visual representation of the results presented in Table 3. Different settings of hearing impairment degrees are shown with different shapes. The results of optimisation with SA only, and SA + LR with moderate hearing impairment, are presented with mean WERs and error bars.

The preliminary results show the optimisation with SA + SS using the mild and severe settings can achieve marginally lower WERs on *test-clean* and *test-other* compared to the average WERs of the baseline models, and the moderate setting performs worse than the baseline on *test-clean*. We further probed the combination of SA + LR + SS, and found SS could not achieve lower WERs compared to SA + LR. In conclusion, SS could not bring improvement to the ASR models. On the other hand, the preliminary results show significant improvement can be achieved with LR augmentation. LR with the moderate hearing impairment setting performs best on *test-other*, and very close to the best performance of *test-clean* achieved by the mild setting. This could be related to the phenomenon that people with moderate hearing loss are able to understand clean speech, and severe hearing impairment setting could damage speech too much so that key information for recognition could be lost.

Finally, repeated independent experiments of the optimisation of SA and LR with moderate hearing impairment with different random seeds are reported. SA can be further improved with the proposed method with relative 2.4% and 4.3% improvements on *test-clean* and *test-other*, respectively, and the improvements are statistically significant [t-test, $p < 0.05$].

We further validate the performances with a transformer-based language model. The baseline models optimised with SA only can reach the WERs of 2.38% ±0.01% and 5.50% ±0.02% on the *test-clean* and *test-other*. Models optimised with SA + LR with moderate setting can achieve the WERs of 2.33% ±0.01% and 5.32% ±0.03% , with statistically significant relative improvements of 2.1% and 3.3% compared to the baseline [t-test, $p < 0.05$].

6. CONCLUSIONS

In this work, two auditory mechanisms SS and LR from the well-verified MSBG hearing impairment model are studied for end-to-end ASR data augmentation. The mechanisms can augment speech datasets on the fly when optimising ASR models. The results show that LR with moderate hearing impairment setting can achieve statistically significant improvement on the performance of the state-of-the-art SA method on both *test-clean* and *test-other* of LibriSpeech when both using and not using an external language model. In the future work, the proposed data augmentation technique will be verified with more challenging speech recognition tasks, such as CHiME 5 [38].

7. REFERENCES

[1] A. Graves et al., "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[2] J. Chorowski et al., "End-to-end continuous speech recognition using attention-based recurrent nn: First results," in *NIPS 2014 Workshop on Deep Learning*, 2014.

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

[4] E. Trentin and M. Gori, "A survey of hybrid ann/hmm models for automatic speech recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91–126, 2001.

[5] Rich Caruana et al., "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," *NeurIPS*, pp. 402–408, 2001.

[6] L. Varnet et al., "High-frequency sensorineural hearing loss alters cue-weighting strategies for discriminating stop consonants in noise," *Trends in Hearing*, vol. 23, 2019.

[7] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in noise," *JASA*, vol. 94, no. 3, pp. 1229–1241, 1993.

[8] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *JASA*, vol. 95, no. 4, pp. 2277–2280, 1994.

[9] B. C. J. Moore and B. R. Glasberg, "Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech," *JASA*, vol. 94, no. 4, pp. 2050–2062, 1993.

[10] Michael A Stone and Brian CJ Moore, "Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear and Hearing*, vol. 20, no. 3, pp. 182–192, 1999.

[11] Simone Graetzer et al., "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Interspeech*, 2021, pp. 686–690.

[12] C. R. Jankowski et al., "A comparison of signal processing front ends for automatic word recognition," *IEEE Transactions on Speech and Audio processing*, pp. 286–293, 1995.

[13] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *JASA*, vol. 106, no. 4, pp. 2040–2050, 1999.

[14] Yang Shao et al., "An auditory-based feature for robust speech recognition," in *ICASSP*. IEEE, 2009, pp. 4625–4628.

[15] Jun Qi et al., "Auditory features based on gammatone filters for robust speech recognition," in *ISCAS*. IEEE, 2013, pp. 305–308.

[16] Marc René Schädler et al., "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," *JASA*, vol. 131, no. 5, pp. 4134–4151, 2012.

[17] Bernd T Meyer et al., "Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition," in *Interspeech*, 2012.

[18] Castro Martinez et al., "Should deep neural nets have ears? the role of auditory features in deep learning approaches," in *Interspeech*, 2014.

[19] Deepak Baby et al., "Investigating modulation spectrogram features for deep neural network-based automatic speech recognition," *Interspeech*, vol. 2015, pp. 2479–2483, 2015.

[20] Castro Martinez et al., "On the relevance of auditory-based gabor features for deep learning in robust speech recognition," *Computer Speech & Language*, vol. 45, pp. 21–38, 2017.

[21] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *ICML WDLASL*, 2013, vol. 117, p. 21.

[22] Naoyuki Kanda et al., "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *ASRU*. IEEE, 2013, pp. 309–314.

[23] Tom Ko et al., "Audio augmentation for speech recognition," in *Interspeech*, 2015.

[24] Syed Shahnawazuddin et al., "Pitch-adaptive front-end features for robust children's asr.," in *Interspeech*, 2016.

[25] Tom Ko et al., "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*. IEEE, 2017, pp. 5220–5224.

[26] László Tóth et al., "A perceptually inspired data augmentation method for noise robust cnn acoustic models," in *SPECOM*. Springer, 2018, pp. 697–706.

[27] Chanwoo Kim et al., "Small energy masking for improved neural network training for end-to-end speech recognition," in *ICASSP*. IEEE, 2020, pp. 7684–7688.

[28] D. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.

[29] M. Ravanelli et al., "Multi-task self-supervised learning for robust speech recognition," in *ICASSP*. IEEE, 2020, pp. 6989–6993.

[30] Eugene Kharitonov et al., "Data augmenting contrastive learning of speech representations in the time domain," in *SLT*. IEEE, 2021, pp. 215–222.

[31] Zehai Tu et al., "Optimising Hearing Aid Fittings for Speech in Noise with a Differentiable Hearing Loss Model," in *Interspeech*, 2021, pp. 691–695.

[32] Connor Shorten and Taghi M Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[33] Ashish Vaswani et al., "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[34] Wei Han et al., "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," 2020.

[35] Suyoun Kim et al., "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *ICASSP*. IEEE, 2017, pp. 4835–4839.

[36] Vassil Panayotov et al., "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*. IEEE, 2015, pp. 5206–5210.

[37] Mirco Ravanelli et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[38] Jon Barker et al., "The fifth'CHiME'speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, 2018.