Kathania, Hemant; Singh, Mittul; Grósz, Tamás; Kurimo, Mikko

Data augmentation using prosody and false starts to recognize non-native children's speech

Published in:
Proceedings of Interspeech

DOI:
10.21437/Interspeech.2020-2199

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Data augmentation using prosody and false starts to recognize non-native children's speech

Hemant Kathania, Mittul Singh, Tamás Grósz, Mikko Kurimo

Department of Signal Processing and Acoustics, Aalto University, Finland
firstname.lastname@aalto.fi

Abstract

This paper describes AaltoASR's speech recognition system for the INTERSPEECH 2020 shared task on Automatic Speech Recognition (ASR) for non-native children's speech. The task is to recognize non-native speech from children of various age groups given a limited amount of speech. Moreover, the speech being spontaneous has false starts transcribed as partial words, which in the test transcriptions leads to unseen partial words. To cope with these two challenges, we investigate a data augmentation-based approach. Firstly, we apply the prosody-based data augmentation to supplement the audio data. Secondly, we simulate false starts by introducing partial-word noise in the language modeling corpora creating new words. Acoustic models trained on prosody-based augmented data outperform the models using the baseline recipe or the SpecAugment-based augmentation. The partial-word noise also helps to improve the baseline language model. Our ASR system, a combination of these schemes, is placed third in the evaluation period and achieves the word error rate of 18.71%. Post-evaluation period, we observe that increasing the amounts of prosody-based augmented data leads to better performance. Furthermore, removing low-confidence-score words from hypotheses can lead to further gains. These two improvements lower the ASR error rate to 17.99%.

Index Terms: Children speech recognition, prosody modification, SpecAugment, hesitation noise, confidence score filter

1. Introduction

Many schools around the world teach English as a foreign language to their pupils. As the trend grows, it creates a demand to develop an objective and reliable English language skills assessment for young learners. In turn, the assessment system requires a robust speech recognition system. However, recognizing children's speech is challenging and the non-native nature of the speech adds further complexity to this task.

INTERSPEECH 2020 shared task offers a unique opportunity to recognize non-native children's speech. As part of the shared task, the provided dataset includes various speech phenomena like code-switching between multiple languages, a large number of spontaneous speech phenomena (like hesitations, false starts, fragments of words), presence of non-collaborative speakers (students often joke, laugh, speak softly, etc.) and multiple age-groups (9 - 16 years). Handling these phenomena is exacerbated by the presence of background noise in the dataset. Furthermore, the dataset provides only limited amounts of labeled audio ($\sim$ 50 hours).

In this paper, we aim to handle the data scarcity, acoustic variability and false starts in text for building an Automatic Speech Recognition (ASR) system. To alleviate data scarcity, we augment training audio data. Specifically, we compare SpecAugment- [1] and prosody-based [2] data augmentation (section 4). SpecAugment, recently popularized for building a robust ASR, has not been explored for processing children's speech. This technique ignores the prosodic variability of children's speech from different age-groups, which is relevant to this task. In contrast, prosody-based data augmentation leverages this acoustic variability to increase the amount of data. Our expriments also show that the prosody-based augmentation is more beneficial than SpecAugment for children's speech.

Table 1: *The table reports statistics for the organizer provided datasets used in our paper: two training sets: train-1 and train-2, the development set (dev) and the evaluation set (eval).*

Corpus	train-1	train-2	dev	eval
No. of words	22450	136578	5287	6206
No. of pupils	338	3112	84	84
Duration (hours)	8.59	40.29	2.05	2.20

The shared task represents false starts in speech as partial words. Approximately 10% of the development and evaluation set utterances contain such words. Prior work [3, 4] has noted that handling disfluencies like false starts can lead to better prediction of the next word. To handle the partial words, we randomly add partial words by splitting existing words to the language modeling text (section 5.1). We observe that noising text in this manner improves the ASR performance.

During the evaluation period, our ASR system ranked third, using a combination of prosody- and SpecAugment-based data augmentation while handling false starts (section 7). We also release tools for prosody-based augmentation and handling false starts publicly [1]. Post-evaluation period, we experiment with increasing the augmented data, which improves the performance of prosody-based ASR, whereas the performance of SpecAugment-based ASR drops (section 4). We also implement a filtering scheme to remove words with low confidence scores (section 6). The filtering technique can deal with non-English words, which are discounted by the shared task metric. The filtering scheme helps to improve the performance of the combined system further.

2. ASR for non-native children's speech

2.1. Dataset

In this challenge, the organizers provide us with an English portion of speech dataset called TLT-school corpus [5]. The corpus consists of audio from Italian pupils speaking English collected between the years 2016 and 2018. The pupil's ages range from 9 to 16 years, belonging to four different school grade levels. The pupils are divided into three age-based groups A1 (9-10),

[1] https://github.com/kathania/Interspeech-2020-Non-native-children-ASR.git

http://dx.doi.org/10.21437/Interspeech.2020-2199

Table 2: *The table shows a portion of an utterance from the reference text (Reference) in the evaluation set. The evaluation script removes the words in red in this text to produce the modified text for WER calculations (Modified). We also present the example utterance's hypothesis predicted by our best system before confidence-score-based filtering (Prediction) and the filtered version (Filtered) produced by dropping the words with very low-confidence scores (marked in blue). The filtering process is described in section 6.*

Reference	<unk> in <unk> the people i watch the
Modified	in the people i watch the
Predicted	@uh interesting ping in work people i watch the
Filtered	interesting in people i watch the

A2 (12-13) and B1 (14-16). Each age-group is asked to answer questions according to their language skills. The recorded answers form the speech provided in this challenge.

The dataset is provided in two parts Train-1 and Train-2. Train-1 contains 8.6 hours of manually transcribed data from 2017 recordings and Train-2 contains 40.3 hours of data from 2016 and 2018 recordings. For development (dev) and evaluation (eval) sets, the organizers provide around two hours of data each from 2017 recordings. Data statistics, like the number of speakers and words, are presented in Table 1.

As the speech is from different age-groups, the dataset has variations in the speaker's pitch and speaking rates. These variations are reported in Table 3. Comparing pitch, we notice that A1 and A2 groups are similar to some extent but different from the B1 group. While comparing the speaking rate, we observe that A2 and B1 are similar and faster than A1. These differences form the basis for our experiments, where we leverage the pitch and speaking rate to augment speech data.

2.2. Evaluation procedure

An interesting aspect of this task is the modified Word Error Rate (WER) metric used to compare ASR performance. Before computing the regular WER, the modified procedure filters out non-English tokens like unknown tokens (e.g. `<unk>`, `<unk-it>`, `<unk-de>`), disfluencies like false starts (e.g. `pro-` from `pro- program`) and filler tokens (e.g. `@m`, `@e`) from both the hypotheses and references. Intuitively, removing such words leads to a comparison of only English words to calculate WER. Table 2 shows an example of this process.

3. Baseline systems

The challenge organizers provide a Kaldi toolkit-based recipe to train on the Train-1 portion (9 hours) of their data [6]. This setup utilizes MFCC features as input to train TDNN-based acoustic models [7] on LDA-MLLT+SAT based GMM alignment labels. The recipe also performs speaker adaptation of the acoustic model using i-vectors [8]. The decoding is then performed using a 4-gram maximum entropy language model built using SRILM toolkit [9]. The recipe can be further modified to train acoustic and language models on combined Train-1 and Train-2 portions of the data. Training on the combined data leads to large performance improvements in terms of Word-Error-Rate (WER), as noted in Table 4. We also include two spelling corrections in these results, where the American English spelling *favorite* is replaced by its British variant *favourite* and the word *coca-cola* is split in two words *coca cola*. These

Table 3: *The table reports the average pitch and speaking rate among different age-groups: A1 (9-10 years), A2 (12-13 years) and B1 (14-16 years). We measure the speaking rate in words per second for the same duration across the different age groups. The table also shows the number of utterances with false starts in the development set (dev).*

Data	A1	A2	B1
Pitch scale			
Pitch	218	212	194
Speaking rate			
Word/sec	0.614	1.159	1.036
No. of words	11501	21699	19401
Duration in sec	18720		
Partial words			
Utterances % (dev)	8.4	14.5	27.7

spelling corrections help normalize the spelling used across training and development portions of the dataset.

We also train bidirectional LSTM-based TDNN (TDNN-BLSTM) acoustic models as a comparative baseline system. This model performs worse than the regular TDNN on WER, as shown in Table 4. However, this system shows benefits when combining with other different acoustic models (Section 7) employed in this paper. We also plan to investigate other acoustic model architectures like CNN-TDNN as well for this task. In the rest of the paper, we use the baseline recipe unless specified otherwise.

To handle the speech of different age groups, we also perform Vocal Tract Length Normalization (VTLN) [10] per age-group, which can aid speaker adaptation at the input feature level. Training the TDNN acoustic model with the modified data outperforms the simplistic TDNN baseline (Table 4).

4. Data augmentation

In this paper, we implement a prosody-based data augmentation technique, which we describe in Section 4.1. We also contrast this technique to SpecAugment-based data augmentation [1], which has recently shown benefits for ASR in general.

4.1. Prosody modification based data augmentation

We change the pitch scale and the speaking rate systematically to leverage prosodic variation in the children's speech (Section 2.1). This process introduces more acoustic variability to the original children's speech corpora. We then augment the modified data to the original corpora for further system development. Figure 1 summarizes the augmentation process. Intuitively, increasing acoustic variability adds noise to the input features regularizing the learning process.

To modify pitch and speaking rate, we have explored Time Scale Modification (TSM) based on Real-Time Iterative Spectrogram Inversion with Look-Ahead (RTISI-LA) algorithm [11, 12, 13, 14]. This algorithm constructs a high-quality time-domain signal from its short-time magnitude spectrum.

RTISI-LA algorithm scales down the pitch per frame of spectrogram with a factor s ($0 < s < 1$) and upsamples this frame to maintain the original size. Next, the algorithm computes a short-time Fourier transform magnitude (STFTM) of the obtained frame. The STFTM describes the audio signal, perceived in terms of its frequency components, by combining the imaginary and real parts into a single number. The RTISI-LA reconstructs the audio signal from its STFTM through an iterative process. Similarly, the RTISI-LA [12, 13, 14] can vary

Table 4: *The table reports WER for various ASR systems. The systems vary in acoustic models, amount of data used for training and acoustic modification applied (if any). For details on augmentation techniques refer to section 4. Asterisks (*) denote statistical significance while comparing against TDNN (23.06) using the matched pairs test with* $p < 0.001$.

AM	Data size	Acoustic mods.	WER
Baselines			
TDNN (9 hrs)	0.2x	-	37.75*
TDNN (baseline)	1x	-	23.06
TDNN-BLSTM	1x	-	29.07*
TDNN	1x	VTLN	**22.68**
SpecAugment			
TDNN	2x	SpecAug	**22.21***
TDNN	4x	SpecAug	22.85
Prosodic Modifications			
TDNN	2x	Speaking rate (SR)	22.58
TDNN	2x	Pitch (P)	21.92*
TDNN	3x	SR-P	21.75*
TDNN	5x	SR2-P2	**21.58***

the speaking-rate by changing the length of the speech signal per unit time by varying the speaking rate factor α.

Table 4 reports the ASR performance for different prosody-based augmentations made to the original data. These include:

1. when only speaking rate (SR) modified data with $\alpha = 1.1$ is augmented to the original,
2. when only pitch scale (P) modified data with $s = 0.9$ is augmented to the original,
3. both SR and P modified data are augmented (SR-P) to the original, and
4. the original data further augmented with speaking-rate modification with $\alpha = 1.2$ and pitch scale modification with $s = 0.85$ and added to SR-P (SR2-P2).

In cases 1) and 2), augmentation doubles the amount of data. Here the pitch-scale based modifications result in a better performance between the two. We observe subsequent improvements when increasing the data via pooling of data from pitch-scale and speaking rate modifications, i.e., SR-P and SR2-P2. In all these cases, the increased prosodic variability helps to improve ASR performance.

4.2. SpecAugment

SpecAugment [1] modifies the input spectrogram by removing time and frequency information randomly. It further warps information across the time axis producing variable speaking rates in different segments of the audio. In our experiments, we use *Librispeech double* augmentation policy, which had performed well on the Librispeech dataset [1], and applied it directly to the MFCC features to create additional data. In the future, we explore SpecAugment applied to filter bank features as done in the original recipe [1]. Similar to prosody-augmentation, the modified data is augmented to the original for further use in ASR development, as shown in Figure 1.

Table 4 shows that doubling the data through SpecAugment (2x SpecAug) improves performance while subsequent increase (4x SpecAug) leads to worse results. This effect is in contrast with Prosody-based augmentation, which shows a consistent improvement with subsequent increase in augmented data size.

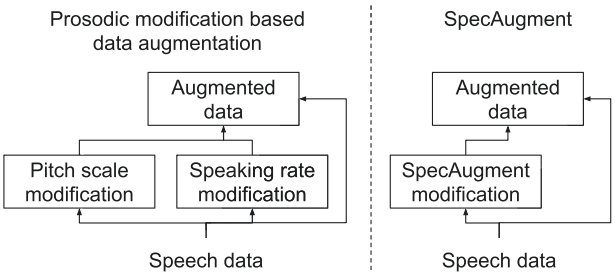

Figure 1: *The figure displays a block diagram depicting the different data augmentation methods used in our work.*

Table 5: *This table shows WERs for the TDNN-based system with language models trained using different language modeling text. Asterisks (*) denote statistical significance while comparing against (1) using the matched pairs test with* $p < 0.001$.

#	LM Text	WER	+Conf. Filter
1	Original text	23.06	22.38
2	+False start noise	23.32*	23.89*
3	Unnormalized text	23.09	22.58
4	+False start noise	23.93	23.61*
Linear combinations			
	1+2	23.00	23.61*
	1+3	22.92*	**22.32**
	1+4	**22.83**	22.51

5. Language modeling

The organizers normalize 2016 and 2018 transcripts for language modeling (Original) by removing some of the filler tokens. Training language models with the normalized text bias them, as contexts with filler tokens become unseen. The 2017 transcripts are made available in an *unnormalized* form with the filler tokens intact. Along with training language models with normalized text, we also create language models with unnormalized text. Like the baseline recipe, we create a 4-gram maximum entropy model for both normalized and unnormalized text. Table 5 shows the performance of these two models. We also linearly interpolate these models trained on different corpora, which improves over the constituent models.

5.1. Handling false starts in text

In TLT-school corpus, children across different age-groups frequently hesitate while speaking. These false starts are marked as partial words (like `pro-` in `pro- program`). Table 3 shows the percentage of utterances per age-group containing partial words. We observe that the partial words affect around 10% portion of the development set. The B1 speakers report the most number of false starts as they usually speak longer sentences having a higher probability of hesitating per sentence.

To handle partial words, we artificially add such words in the language modeling corpus. In this process, we first sample words with at least three characters from the language modeling corpora. The sampled word is split in a random position. Finally, the sampled word is replaced by the partial word and itself (e.g. `program` → `prog- program`). Then the noised text can be used to build language models. The noised-text models do not perform well compared to the source text models, as shown in Table 5. However, they show improvement when augmented with original-text models via linear interpolation. We note that interpolating with original-text models is crucial; otherwise, the noise-text models do not perform well.

Table 6: *The table presents WERs of different ASR system combinations on development (dev) and evaluation (eval) sets. † marks our best-submitted system. Post-evaluation, we improved on these results and our best results are marked in boldface. Asterisks (*) denote statistical significant results compared to the baseline using the matched pairs test with* $p < 0.001$.

AM	LM	dev	eval
Individual systems			
TDNN (baseline)	Original	23.06	20.26
SR2-P2	Org.+Unnorm.	21.45*	19.71
SR2-P2 + Conf. Filter	Org.+Unnorm.	21.20*	19.10*
System combinations			
TDNN+BLSTM+VTLN	Original	21.92*	19.58
3x(2x SpecAug)	Original	21.18*	19.84
SR-P+SR2-P2	Original	20.92*	19.03
System combinations with TDNN+BLSTM+VTLN			
+3x(2x SpecAug)	Original	21.01*	19.13*
+SR-P	Original	20.86*	18.80*
+SR-P	+Unnormalized	20.54*	19.01*
+SR-P	+Noise	20.43*	18.71*†
+SR2-P2	Original	20.60*	18.68*
+SR2-P2	+Unnormalized	20.31*	18.58*
+SR2-P2	+Noise	20.09*	18.42*
+Conf. Filter	Org.+Unnorm.+Noise	19.86*	**17.99***

6. Filtering the decoding output

In this task, the modified WER (Section 2.2) removes non-English words before calculating WER. This removal is to facilitate the comparison of only English words. However, in practice, the ASR system can incorrectly predict non-English tokens as English words and contribute to the error.

In our post-evaluation experiments, we built a filter to remove these words from the decoding output — the filter inputs the word confidence scores, a combination of acoustic and language model scores. The filter only outputs words that have a confidence score above a certain threshold. We show an example of this filter in action on a segment of text from the evaluation set in Table 2. From the predicted statement (Predicted), the filter can remove incorrectly recognized words in the Filtered statement. Also, filtering out low-confidence score words helps improve the ASR performance, as shown in Table 5. We chose the filter thresholds in the range of $[0, 1]$ that produced the best WER on the development set.

7. System combinations

As no external resources are used to create our ASR systems, we submit our systems as part of the *closed track* for the shared task on recognizing non-native children's speech. On this task, we report the best individual and combined systems' WER as chosen on the development (dev) set and evaluated on the test set (eval) in Table 6. For evaluation, we retrained all our systems on the training and development set. Among the individual systems, the ASR trained on the most amount of augmented data (SR2-P2) achieves the best result. This system also applies the word confidence score filtering and language models built using original 2016-2018 (Original) and unnormalized 2017 (Unnormalized) transcripts.

Earlier, increasing the amount of SpecAugment-based data (4x SpecAug) did not result in an improvement. Nevertheless, combining three different 2x SpecAug ASR systems, labeled as 3x(2x SpecAug), leads to improvement over the individual 2x SpecAug system. Combining both the prosody-augmentation based systems (SR-P+SR2-P2) leads to the best data augmentation based system.

During the evaluation period, we submitted a system combination of 3x(2x SpecAug) with SR-P. This ASR system also utilizes an interpolated language model where the constituents are trained on the original normalized text, unnormalized text and noised unnormalized (noise) text. For individual systems, the noised-text-based language models did not improve performance compared to using just original and unnormalized text-based models but turned out to be essential for building combined systems. In our post evaluation, we improved this system by combining it with SR2-P2, which adds more prosody-augmentation and applies word score confidence filtering. These additions helped to achieve our best WER of 17.99.

8. Related work

In the context of children speech, prosodic features and modifications are well studied [2, 11, 13, 15, 16]. Prior work [16] has leveraged similar prosody modifications for data augmentation in children ASR achieving substantial gains in performance. These benefits also inspire our solution. Though, unlike prior work [16], which used a glottal-closure-instants-based modification, we use a simpler TSM based algorithm to modify the prosodic parameters like pitch and speaking rate.

In the context of language modeling, quite a few researchers [3, 4, 17, 18, 19, 20] have studied disfluencies like false starts, filled pauses and repetition in textual data. Some of these work [3, 4] have noted that modeling disfluencies can be beneficial for language modeling. In the same vein, our work focuses on modeling false starts in text. Most similar to our work, [4] introduces disfluencies to clean text and uses the processed text for the language model. They, however, use an existing set of disfluencies to be introduced in the text. In contrast, we split words in the text to introduce new disfluencies.

9. Concluding remarks

In this work, we presented AaltoASR's system for the task of recognizing non-native children's speech. We focused on applying a data augmentation-based approach. We leveraged prosody- and SpecAugment-based data augmentation to augment the limited training data for building acoustic models. Compared to SpecAugment, prosody-based augmentation achieved better results. Additionally, prosody-based augmentation showed improvement when increasing the amount of augmentation data, whereas increasing the amount of augmented data for SpecAugment led to worse results.

We also modeled false starts (partial words) in the text to augment the language modeling corpora for training. Adding the partial-word noise improved the ASR performance of linearly interpolated models compared to the vanilla language models. In our post-evaluation experiments, we developed a filtering mechanism to remove low confidence scores from the decoded output, which helped improve the ASR performance. Finally, we performed a system combination of the techniques developed in this work to achieve the best result.

10. Acknowledgements

This work was supported by the Academy of Finland (grant 329267) and the Kone Foundation. The computational resources were provided by Aalto ScienceIT.

11. References

[1] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2613–2617.

[2] S. Shahnawazuddin, N. Adiga, and H. K. Kathania, "Effect of prosody modification on children's asr," *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1749–1753, 2017.

[3] D. Prylipko, B. Vlasenko, A. Stolcke, and A. Wendemuth, "Language modeling of nonverbal vocalizations in spontaneous speech," in *Text, Speech and Dialogue - 15th International Conference, TSD 2012, Brno, Czech Republic, September 3-7, 2012. Proceedings*, ser. Lecture Notes in Computer Science, P. Sojka, A. Horák, I. Kopecek, and K. Pala, Eds., vol. 7499. Springer, 2012, pp. 488–495.

[4] J. Staš, D. Hládek, and J. Juhár, "Adding filled pauses and disfluent events into language models for speech recognition," in *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, 2016, pp. 000 133–000 136.

[5] R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna, "Tlt-school: a corpus of non native children speech," *CoRR*, vol. abs/2001.08051, 2020. [Online]. Available: https://arxiv.org/abs/2001.08051

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech recognition toolkit," in *Proc. ASRU*, December 2011.

[7] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech 2018*, B. Yegnanarayana, Ed. ISCA, 2018, pp. 3743–3747.

[8] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*. IEEE, 2013, pp. 55–59.

[9] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Interspeech 2002*, J. H. L. Hansen and B. L. Pellom, Eds. ISCA, 2002.

[10] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96, Atlanta, Georgia, USA, May 7-10, 1996*. IEEE Computer Society, 1996, pp. 346–348.

[11] W. Ahmad, S. Shahnawazuddin, H. Kathania, G. Pradhan, and A. Samaddar, "Improving children's speech recognition through explicit pitch scaling based on iterative spectrogram inversion," in *Interspeech 2017*, 2017, pp. 2391–2395. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-302

[12] H. K. Kathania, S. Shahnawazuddin, W. Ahmad, N. Adiga, S. K. Jana, and A. B. Samaddar, "Improving children's speech recognition through time scale modification based speaking rate adaptation," in *2018 International Conference on Signal Processing and Communications (SPCOM)*, July 2018.

[13] H. K. Kathania, W. Ahmad, S. Shahnawazuddin, and A. B. Samaddar, "Explicit pitch mapping for improved children's speech recognition," *Circuits, Systems, and Signal Processing*, vol. 32, p. 2021–2044, 2018.

[14] X. Zhu, G. T. Beauregard, and L. L. Wyse, "Real-time signal estimation from modified short-time fourier transform magnitude spectra," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1645–1653, 2007.

[15] H. K. Kathania, S. Shahnawazuddin, N. Adiga, and W. Ahmad, "Role of prosodic features on children's speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5519–5523.

[16] S. Shahnawazuddin, N. Adiga, H. K. Kathania, and B. T. Sai, "Creating speaker independent asr system through prosody modification based data augmentation," *Pattern Recognition Letters*, vol. 131, pp. 213 – 218, 2020.

[17] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, ICASSP '96, Atlanta, Georgia, USA, May 7-10, 1996*. IEEE Computer Society, 1996, pp. 405–408.

[18] M. Siu and M. Ostendorf, "Modeling disfluencies in conversational speech," in *Interspeech 1996*. ISCA, 1996.

[19] Y. Liu, E. Shriberg, and A. Stolcke, "Automatic disfluency identification in conversational speech using multiple knowledge sources," in *Interspeech 2003*. ISCA, 2003.

[20] H. Moniz, I. Trancoso, and A. I. Mata, "Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts," in *Interspeech*. ISCA, 2009, pp. 1719–1722.