Improving Whispered Speech Recognition Performance Using Pseudo-Whispered Based Data Augmentation

Lin, Zhaofeng ; Patel, Tanvina; Scharenborg, Odette

DOI
10.1109/ASRU57964.2023.10389801

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389801

IMPROVING WHISPERED SPEECH RECOGNITION PERFORMANCE USING PSEUDO-WHISPERED BASED DATA AUGMENTATION

Zhaofeng Lin, Tanvina Patel, Odette Scharenborg

Multimedia Computing Group, Delft University of Technology, the Netherlands

ABSTRACT

Whispering is a distinct form of speech known for its soft, breathy, and hushed characteristics, often used for private communication. The acoustic characteristics of whispered speech differ substantially from normally phonated speech and the scarcity of adequate training data leads to low automatic speech recognition (ASR) performance. To address the data scarcity issue, we use a signal processing-based technique that transforms the spectral characteristics of normal speech to those of pseudo-whispered speech. We augment an End-to-End ASR with pseudo-whispered speech and achieve an 18.2% relative reduction in word error rate for whispered speech compared to the baseline. Results for the individual speaker groups in the wTIMIT database show the best results for US English. Further investigation showed that the lack of glottal information in whispered speech has the largest impact on whispered speech ASR performance.

***Index Terms*—** Whispered speech, pseudo-whisper, end-to-end speech recognition, wTIMIT, signal processing

1. INTRODUCTION

Whispered speech differs substantially from normally phonated speech. In whispered speech, vocal fold vibrations are absent, resulting in voicelessness [1], which leads to reduced energy and intensity [2]. Additionally, the airflow in whispered speech is often greater than in normal speech, which leads to breathy speech [3]. Whispering is often used in private conversations or in libraries or meetings. Whispered speech also occurs in pathological speech contexts: Speech from individuals who face vocal system challenges such as diseases affecting the vocal folds or post-larynx surgery [4, 5] is often whisper-like. Moreover, whispering has been found to be beneficial in reducing or avoiding stuttering [6].

While human listeners can largely comprehend the linguistic information in whispered speech [7], automatic speech recognition (ASR) systems face significant challenges in accurately transcribing whispered speech compared to normal speech [2]. The poorer performance of ASR systems on whispered speech is due to the large acoustic differences between whispered and normal speech, e.g., pitch differences due to reduced vocal fold vibrations [8], up-shifted formant frequencies [9, 10], wider formant bandwidth [9], reduced phonetic cues, and a lower signal-to-noise ratio [2], and most importantly the limited amount of training data, which makes it challenging to develop robust ASR models tailored to whispered speech.

Several approaches have been proposed in the literature to improve whispered speech ASR. For hybrid ASR systems, these include using Teager Energy Cepstral Coefficients instead of traditional Mel-frequency cepstral coefficients (MFCC) [11], showing large improvements for Serbian whispered speech. In [9] the authors proposed a method to generate pseudo-whispered speech segments using denoising autoencoders showing considerable performance improvements on their own dataset. For End-to-End (E2E) systems, Chang *et al.* [12] showed that a system trained with a frequency-weighted SpecAugment, a frequency-divided Convolutional Neural Network extractor, a layer-wise transfer learning approach, and pre-training outperformed their baseline with about 44% relative improvement in character error rate on whispered speech from wTIMIT [8]. The work in [13] generated whispered speech from normal speech using Generative Adversarial Networks-based voice conversion (VC) techniques for training data augmentation obtaining the current best results on wTIMIT: 29.4% word error rate (WER). Other methods used multimodal data including articulatory cues from motion data [14, 15] and visual information [16]. Although these techniques show that it is possible to improve the recognition of whispered speech, there is still a performance gap with respect to normal speech.

In this paper, we aim to 1) understand the (detrimental) effects of the specific acoustic characteristics of whispered speech on whispered speech ASR performance 2) and deal with the data scarcity problem by generating artificial whispered speech to augment the training data for improved E2E whispered speech ASR. To that end, we propose a handcrafted signal-processing method to convert normal speech to pseudo-whispered speech in two independent steps. These steps allow us to create "intermediate forms" of normal-to-whispered speech, which in turn allow us to investigate the effect of the specific acoustic characteristics of whispered speech on whispered speech ASR.

For our experiments, we use the whispered TIMIT speech dataset [8], which has two speaker groups with different ac-

979-8-3503-0689-7/23/$31.00 ©2023 IEEE

cents, North American English and Singaporean English. [8] showed that the whispered speech recognition performance was dependent on the accent group, with worse results for the Singaporean English speech. Here we, to the best of our knowledge, for the first time, investigate the effect of different accents (speaker groups) in E2E whispered speech ASR.

2. METHODOLOGY

Section 2.1 describes the three datasets that were used for testing, training, and for generating the pseudo-whispered speech in the experiments. Section 2.2 provides a brief comparison of normal versus whispered speech. Section 2.3 describes our approach to the conversion from normal to pseudo-whispered speech. Section 2.4 explains the experimental setup.

2.1. Datasets

2.1.1. wTIMIT

The Whispered TIMIT (wTIMIT) corpus [8] consists of 450 phonetically balanced sentences in both normal (wTIMIT-n) and whispered (wTIMIT-w) speech from speakers from two accent groups: US and Singaporean English with 28 (12 male and 16 female) and 20 (12 male and 8 female) speakers, respectively. wTIMIT thus consists of four speaker groups: $\mathbf{N_{US}}$: Normal speech with US accent; $\mathbf{N_{SG}}$: Normal speech with SG accent; $\mathbf{W_{US}}$: Whispered speech with US accent; $\mathbf{W_{SG}}$: Whispered speech with SG accent.

wTIMIT was originally partitioned into a training and test set [8]. To prevent overfitting our E2E models, a re-partitioning of wTIMIT into training, development, and test sets is needed. Preliminary experiments performed in [12] showed that a partitioning of the training and test data where there was no speaker overlap in the training and test set, degraded performance by approximately 10% relatively compared to a partitioning of the training and test data where the same speaker could occur in both. This relatively small difference in performance was attributed to the pitch being mostly absent in whispered speech. Also given the fact that partitioning by speakers can lead to less data that can be used as training data, [12] suggested that prohibiting speaker overlap between the training and test sets is unnecessary. Following [12], we re-partitioned wTIMIT into a training, development, and test set allowing speaker overlap. Each data set consisted of 400/25/25 sentences, respectively, split from the 450 sentences.

2.1.2. TIMIT

The 450 prompts of wTIMIT were obtained from the phonetically balanced section (SX) of the TIMIT corpus [17], which makes TIMIT a good option as additional training data for normal speech and to generate pseudo-whispered speech. TIMIT contains US English read speech. To validate the training process, we also evaluate our models on TIMIT test set.

2.1.3. LibriSpeech

The LibriSpeech corpus [18] consists of English read speech from audiobooks. LibriSpeech has recently been used as an additional (pre)training dataset in the recent E2E whispered speech studies [12, 13]. We therefore used its 100 hours subset to generate pseudo-whispered speech and augment the training data.

2.2. Differences between normal and whispered speech

Figure 1 shows the spectrograms of a female speaker from the wTIMIT corpus speaking the same utterance in a normal voice (top panel) and while whispering (bottom panel). Comparison of the spectrograms shows that whispered speech appears to have no formant information, less energy in particularly the lower frequency bands, and altogether a different acoustic profile compared to normal speech (see also [12, 19, 20]). These differences are primarily due to the altered vocal production mechanism and reduced vocal fold activity during whispering [8]. Since in whispered speech voicing is significantly reduced or even absent, whispered speech clearly has less distinct harmonic patterns, which makes it challenging to identify individual formants in the spectrogram.

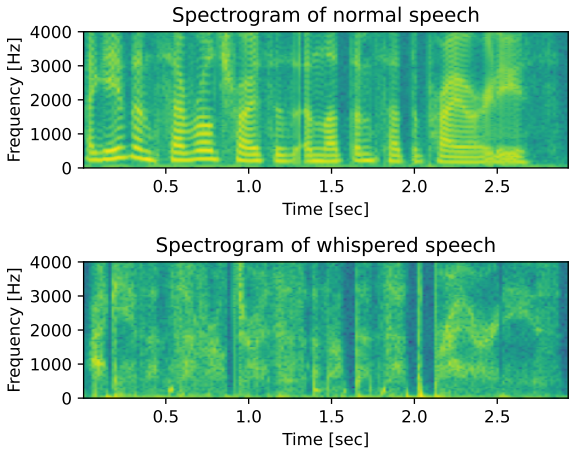

Fig. 1. Spectrograms of the sentence "I gave them several choices and let them set the priorities." produced by the same speaker in a normal and whispered voice. Example is taken from the wTIMIT corpus.

2.3. Proposed pseudo-whispered speech conversion

Our proposed method to convert normal to pseudo-whispered speech is based on that of Cotescu *et al.* [21] who proposed a handcrafted digital signal processing (DSP) recipe that

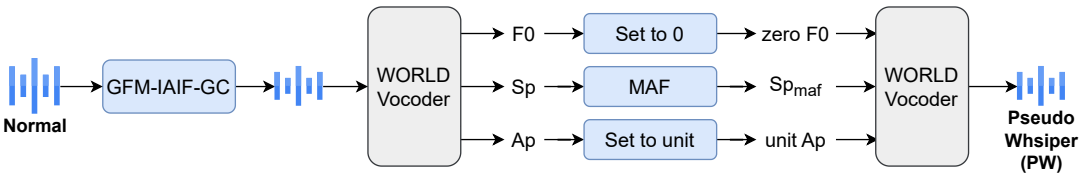

Fig. 2. The proposed pipeline for pseudo-whispered speech conversion, where GFM-IAIF-GC is GFM-IAIF-based glottal cancellation and MAF is moving average filtering. Input is normal speech and the output is pseudo-whispered speech (PW).

converts normal speech into whispered speech in three steps by making acoustic modifications to the normal speech: 1) remove the glottal contribution using spectral subtraction; 2) shift the first formant using frequency warping; 3) increase the formant bandwidth using moving average filtering. The WORLD vocoder [22] is used to extract features for re-synthesizing high-quality speech.

In step 1, instead of using spectral subtraction as in [21], we implemented a glottal cancellation method, which does not require parameters for modelling glottal flow but removes the glottal information directly from a given normal speech signal. Moreover, preliminary experiments using the method from [21] showed that moving average filtering not only widens the formant bandwidth but also up-shifts the formant frequencies. Hence, our proposed method for pseudo-whispered speech conversion is implemented in 2 steps, which is also shown in Figure 2:

1. Removing the glottal source using GFM-IAIF-based Glottal Inverse Filtering (see section 2.3.1),
2. Increasing the formant bandwidth and up-shifting the formant frequencies using moving average filtering (see section 2.3.2).

2.3.1. Step 1: Removing glottal information

Under the assumption of the source-filter model [23], a speech signal is composed of an excitation E, vocal tract filter V, lip radiation filter L, and glottis component G, which can be written in the frequency domain as $S(f) = E(f) \cdot G(f) \cdot V(f) \cdot L(f)$. Glottal Inverse Filtering (GIF) estimates the source of voiced speech, specifically the glottal volume velocity waveform. Iterative Adaptive Inverse Filtering (IAIF) [24] is one of the most widely used algorithms for GIF. IAIF successively models the vocal tract filter $V(f)$, lip radiation $L(f)$, and glottis $G(f)$ using linear prediction (LP) analysis, then removes their effect by inverse filtering. After two iterations, it ultimately removes $V(f)$ and $L(f)$ to leave an estimate of the glottal flow $g(n)$, where n is the discrete-time index.

The Iterative Adaptive Inverse Filtering method based on a Glottal Flow Model (GFM-IAIF) [25] is an improved version of IAIF. It constrains glottal flow by a 3^{rd} order spectral model $G(z) = \left\{\left(1 - az^{-1}\right)\left(1 - a^{*}z^{-1}\right)\left(1 - bz^{-1}\right)\right\}^{-1}$. GFM-IAIF performs competitively for normal phonations [26], which makes it suitable for our case: cancelling glottal contribution of normally phonated speech. In our method, we employ GFM-IAIF to extract glottal flow, after which the effect of the glottis is cancelled by inverse filtering, and the output is a speech signal without glottal contribution.

Subsequently, $F0$, the spectral envelope (Sp), and the aperiodic spectral envelope (Ap) are extracted from the speech signal without glottal information using the WORLD vocoder. To ensure that the pitch is removed entirely, we set the $F0$ to zero and Ap values to all units.

2.3.2. Step 2: Changing formant information

To increase the formant bandwidth and up-shift the formant frequencies, we employ moving average filtering on the spectral envelop Sp extracted by the WORLD vocoder with a 400 Hz-wide triangular window across all frequency axes and get a new spectrogram Sp_{maf}.

The three adapted features, $zero\ F0$, Sp_{maf} and $unit\ Ap$ are passed to the WORLD vocoder for re-synthesising the pseudo-whispered speech PW. Figure 3 shows an example of the conversion results, it shows a zoomed-in spectrogram of normal speech (left panel), whispered speech (middle panel) and pseudo-whispered speech (right panel).

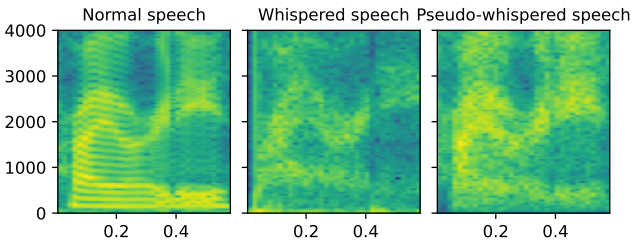

Fig. 3. Spectrogram of normal (left panel), whispered (middle panel), and pseudo-whispered speech (right panel) of the word "priorities" from the same utterance as in Figure 1.

2.4. Experimental setup

All E2E models were trained with the ESPNet toolkit [27]. As speech files in TIMIT and LibriSpeech are recorded with

a sampling rate of 16 kHz, all speech files in wTIMIT were downsampled from 44.1 kHz to 16 kHz. The front-end features are 80 dimensional log-mel filterbank features with 3-dimensional pitch features used for network training.

2.4.1. *Exp. 1: Baseline models*

First, we trained a strong baseline by investigating the

1. Training data: TIMIT plus the normal speech from wTIMIT (TM+wTM-n) vs. TIMIT plus both normal and whispered speech from wTIMIT (TM+wTM-wn);
2. Data augmentation: none vs. speed perturbation (SP) [28] at 90% and 110% of the original rate of the training data and SpecAugment [29] which was used with a maximum width of each time and frequency mask of $T = 20$, $F = 10$, respectively;
3. Dictionary types and sizes: six models used character-level dictionaries, and the remaining two used a Byte Pair Encoding (BPE) token dictionary [30];
4. Model architectures: we compared the Hybrid-CTC [31] and Conformer [32] architectures (from the LibriSpeech recipe from the ESPNet framework).

In total, eight models were trained. The models were evaluated on the TIMIT and wTIMIT-n and wTIMIT-w datasets. Performance was measured in terms of Word Error Rate (WER) for both accent groups separately.

2.4.2. *Exp. 2: Pseudo-whisper data augmentation*

The second experiment investigated the effect of adding pseudo-whispered (PW) data to the training data on whispered speech ASR performance. To that end, the pseudo-whispered speech was created from TIMIT, wTIMIT-n, and LibriSpeech-100h and each was successively added to the training data: first only PW speech from TIMIT (PW(TM)), then the PW speech from wTIMIT-n (PW(wTM-n)) was added, and finally also the PW speech from LibriSpeech-100h (PW(Libri100)). Each set of training data was used to train both the Hybrid-CTC and Conformer architectures, yielding six models. The effect of the pseudo-whisper data augmentation on the two accent groups was also analysed.

2.4.3. *Exp. 3: Acoustic characteristics of whispered speech*

In the final experiment, the effect of the specific acoustic characteristics of whispered speech on whispered speech ASR performance was investigated by comparing the recognition results on speech in which either the glottal information was removed or in which the formant bandwidth had been widened and the formant frequencies shifted.

To that end, we individually applied each of the two steps of our proposed pseudo-whispered speech conversion method on the normal test set in wTIMIT and synthesized the modified speech. Figure 4 shows the pipelines of generating speech without glottal contribution (referred to as NG) and speech with widened formant bandwidth and shifted formant frequencies (referred to as WB). The pipelines are subsets of the full pipeline in Figure 2.

The modified speech was subsequently tested using three models: the Hybrid-CTC architecture trained on only normal speech (row TM+wTM-n in Table 1); trained on normal and whispered speech (row TM+wTM-wn in Table 1); and trained on normal, whispered, and pseudo-whispered speech (TM+wTM-wn+PW(TM+wTM-n)). SP and SpecAug were not applied to all three models in this final experiment.

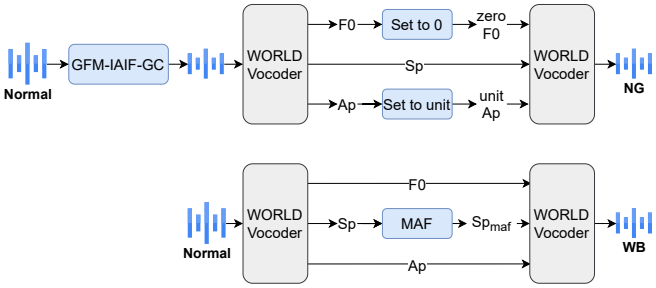

Fig. 4. The pipeline for generating speech with only glottal cancellation (top panel) and with only a widened formant bandwidth and shifted formant frequencies (bottom panel).

3. RESULTS

3.1. Exp. 1: Baseline models

Table 1 presents the results of the baseline models on the TIMIT and wTIMIT test sets for the two accent groups separately and averaged over both accent groups. Only training on normal speech (TM + wTM-n) gave a WER of 40% (Hybrid-CTC) and 52% (Conformer) on the TM test set, while the performance on wTIMIT-n averaged over both accent groups ($\mathbf{N_{Avg}}$) showed a fairly large WER drop of 10-20%. The performance on whispered speech is much lower than that on normal speech, with WERs of over 100%. Including whispered speech in the training data (TM+wTM-wn) improved recognition performance for whispered speech substantially, and reduced the gap with performance on normal speech to less than 10% for both architectures, but at the cost of a slight increase in WER for normal speech. This clearly indicates that using matched training and test sets improves performance.

Applying SP and SpecAug, surprisingly, did not yield improvements for the Hybrid-CTC model; however, it led to a substantial (more than 25%) improvement for both normal and whispered wTIMIT speech for the Conformer model. This suggests that as the training data size increases, the Conformer models outperform Hybrid-CTC models.

When using BPE, the Conformer outperformed the Hybrid-CTC on all three test sets overall and for the individual accent

Table 1. WER (%) of the eight baseline ASR systems on the TIMIT and wTIMIT test sets for the two accent groups separately. $\mathbf{N_{Avg}}$ is normal speech averaged; $\mathbf{W_{Avg}}$ is whispered speech averaged.

Details						TM	wTIMIT					
Training	**Augmentation**	**Hours**	**Architecture**	**Token**	**#Token**	**Test**	$\mathbf{N_{US}}$	$\mathbf{N_{SG}}$	$\mathbf{W_{US}}$	$\mathbf{W_{SG}}$	$\mathbf{N_{Avg}}$	$\mathbf{W_{Avg}}$
TM+wTM-n	None	28.9	Hybrid-CTC	Char	29	40.6	45.4	59.1	99.4	105.2	51.2	101.9
			Conformer	Char	29	52.9	73.4	82.7	102.5	109.6	77.4	105.5
TM+wTM-wn	None	55.1	Hybrid-CTC	Char	29	41.2	51.5	62.8	55.3	74.4	56.3	63.5
			Conformer	Char	29	44.7	78.0	86.4	81.4	92.0	81.6	85.9
TM+wTM-wn	SP + SpecAug	166.3	Hybrid-CTC	Char	29	42.3	52.0	67.3	57.0	77.7	58.5	65.9
			Conformer	Char	29	38.3	49.6	58.3	53.2	68.3	53.3	59.7
TM+wTM-wn	SP + SpecAug	166.3	Hybrid-CTC	BPE	100	44.6	41.2	55.8	44.1	65.9	47.4	53.4
			Conformer	BPE	100	**34.1**	**34.9**	**41.8**	**37.7**	**53.5**	**37.9**	**44.4**

groups, achieving WERs of 37.9% and 44.4% for normal and whispered wTIMIT speech, respectively. These models were selected as our baseline models.

3.2. Exp. 2: Pseudo-whisper data augmentation

Table 2 presents the results of the pseudo-whispered speech experiments. For ease of comparison, the results of the baseline Hybrid-CTC and Conformer models are added (identical to those reported in Table 1. Adding only 3 hours of pseudo-whispered data from TIMIT (with speed perturbation yielding 9 hours; see rows **PW(TM)**) improved the average WER of whispered speech compared to the baseline for both models, with the largest relative improvement for the Conformer model (18.2%). Interestingly, adding pseudo-whispered speech also improved the WER on the normal wTIMIT speech was reduced for both models.

Adding the pseudo-whispered speech of the wTIMIT-n training set (**PW(TM+wTM-n)**) further improved recognition performance for the Hybrid-CTC model but performance for the Conformer model deteriorated for the normal and whispered speech. Recognition performance on the TM test set was again similar to the baseline models. Further adding the PW speech from LibriSpeech (**PW(TM +wTM-n+Libri100)**) gave the best recognition performance for normal wTIMIT speech, but it deteriorated the performance for the whispered speech. The best whispered speech results were obtained with the Conformer model trained with (only) the pseudo-whispered TIMIT speech added.

3.3. Analysis on different speaker groups

Comparing the recognition performance on normal and whispered speech for the two accent groups in the wTIMIT test set showed that Singaporean English normal and whispered speech is consistently worse recognised than US English. This performance gap is the largest for whispered speech.

Adding pseudo-whispered speech always improved the recognition performance of normal and whispered US and Singaporean English, even if the pseudo-whispered speech was based on US English only (PW(TM)). In fact, adding only the US English pseudo-whispered speech from TM gave the best result for $\mathbf{N_{SG}}$ and $\mathbf{W_{SG}}$ and reduced the performance gap with US English to 3.1% for normal speech and 7% for whispered speech for the Conformer, i.e., the smallest performance gap for whispered speech for the Conformer.

Interestingly, adding pseudo-whispered speech from Singaporean English did not further improve recognition performance for $\mathbf{N_{SG}}$ and $\mathbf{W_{SG}}$ for the Conformer, although it did further improve performance for the Hybrid-CTC model, giving the best results for normal and whispered Singaporean English for the Hybrid-CTC model.

When adding the pseudo-whispered US English from LibriSpeech (PW(Libri-100)) as additional training data, the performance on whispered US English ($\mathbf{W_{US}}$) improved to 30.7%, the best result, but it adversely affected the performance on $\mathbf{W_{SG}}$, widening the gap between US and Singaporean English to almost 20% for the Conformer model and even more for the Hybrid-CTC model. Thus, adding a large amount of pseudo-whispered speech based on US English negatively impacted the recognition of Singaporean English normal and whispered speech.

3.4. Exp. 3: Acoustic characteristics of whispered speech

Table 3 presents the results of the experiments on normal speech, real whispered speech, pseudo-whispered (PW) speech and the intermediate forms of whispered speech (see section 2.4), i.e., normal speech without glottal contributions (NG) and normal speech with widened formant bandwidth and shifted formant frequencies (WB). Note that to create PW, both NG and WB are applied.

When the model is trained on only normal speech, the gap between Normal and NG (>25%) is larger than the one be-

Table 2. WER (%) on the TIMIT and wTIMIT test sets when using pseudo-whispered training data generated from TIMIT, wTIMIT-n, and LibriSpeech-100h. Relative improvement (%) of the proposed method compared to the baseline is also reported. Results of the chosen baseline Hybrid-CTC and Conformer models are added (identical to those reported in Table 1.

Details				TM	wTIMIT						Relative Imp.	
Training Data	**Hours**	**Architecture**	**#Token**	**Test**	$\mathbf{N_{US}}$	$\mathbf{N_{SG}}$	$\mathbf{W_{US}}$	$\mathbf{W_{SG}}$	$\mathbf{N_{Avg}}$	$\mathbf{W_{Avg}}$	$\mathbf{N_{Avg}}$	$\mathbf{W_{Avg}}$
Baseline	166.3	Hybrid-CTC	100	44.6	41.2	55.8	44.1	65.9	47.4	53.4	-	-
		Conformer	100	34.1	34.9	41.8	37.7	53.5	37.9	44.4	-	-
+PW(TM)	175.8	Hybrid-CTC	100	46.5	40.1	52.4	41.2	59.9	45.4	49.2	4.2	7.9
		Conformer	100	36.4	32.1	**35.2**	33.4	**40.1**	33.4	**36.3**	11.9	**18.2**
+PW(TM+wTM-n)	253.6	Hybrid-CTC	100	43.4	36.1	44.3	38.0	51.8	39.6	43.9	**16.5**	17.8
		Conformer	100	34.6	32.4	36.6	33.6	42.1	34.2	37.2	9.8	16.2
+PW(TM+wTM-n+Libri100)	557.6	Hybrid-CTC	300	16.9	35.5	55.0	39.2	63.4	43.8	49.5	7.6	7.3
		Conformer	300	**11.0**	**26.8**	38.6	**30.7**	49.2	**31.8**	38.6	16.1	13.1

tween Normal and WB (5%). This indicates that performance is worse for speech without glottal contribution and that the widened formant bandwidth and shifted formant frequencies in whispered speech are less detrimental to recognition performance. Combining NG and WB into pseudo-whispered speech only shows a small deterioration compared to NG. This indicates that the effect of removing both glottal information and widening the formant bandwidth and shifting the formant frequencies is not entirely additive.

Not surprisingly, adding real whispered speech from wTIMIT-n greatly improves the recognition performance of real whispered speech. Recognition performance of pseudo-whispered speech and NG speech also greatly improves, to the level of that of real whispered speech. Performance on WB speech slightly deteriorates. This again indicates that the glottal information is the most important acoustic information to explain the whispered speech recognition performance.

Adding pseudo-whispered speech improves recognition performance of real and pseudo-whispered speech, indicating that the pseudo-whispered speech is close enough to real whispered speech for real whispered speech to benefit from the added data. Recognition performance of PW is actually better than that of real whispered speech which shows the benefit of adding matched training data. Speech without glottal information is now actually better recognised than WB speech, which shows that adding speech without glottal information is most beneficial for NG speech and that the benefit for WB speech is less great.

4. DISCUSSION AND CONCLUSION

This paper aims to deal with the data scarcity problem of whispered speech by generating artificial whispered speech to augment the training data for improved E2E whispered speech ASR and understand what acoustic characteristics of whispered speech have the largest effect on whispered speech ASR performance. Our proposed signal processing-based normal-to-whisper conversion method was used to create pseudo-whispered speech from three databases. Adding pseudo-whisper led to a relative WER improvement of 18% for whispered speech and only a small WER gap with normal speech. Performance was best for the US English accent group. Comparing our results to the state-of-the-art on wTIMIT shows that our WER on whispered speech is higher than in [13]; however, [13] does not report which accent group from wTIMIT they use in their evaluation. Assuming they only used the US English part of wTIMIT, considering that they used large amounts of US English data from LibriSpeech for training, our results are very close to theirs (29.4% vs. our 30.7%) on the US English whispered speech, but using far less data. Comparing our results to those of Chang *et al.* [12]: both their approach and ours showed relative improvements (their 44.4% in CER; our 18.2% in WER); however they only report phone and character error rates, making a direct comparison not feasible in the present work.

Our final experiment shows that the lack of glottal information in whispered speech has the largest impact on whispered speech recognition.

Table 3. WERs (%) of different test groups when the model is trained on normal speech (row TM+wTM-n in Table 1); normal and whispered speech (row TM+wTM-wn in Table 1); and normal, whispered, and pseudo-whispered speech (TM+wTM-wn+PW(TM+wTM-n)).

Training data	**Normal**	**Whisper**	**PW**	**NG**	**WB**
TM+wTM-n	51.2	101.9	79.7	78.0	56.5
TM+wTM-wn	56.3	63.5	65.5	65.2	59.2
TM+wTM-wn +PW(TM+wTM-n)	55.9	61.3	59.2	59.4	62.3

5. REFERENCES

[1] Vivien C Tartter, "What's in a whisper?," *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[2] Taisuke Ito, Kazuya Takeda, and Fumitada Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[3] Ramya Konnai, Ronald C Scherer, Amy Peplinski, and Kenneth Ryan, "Whisper and phonation: Aerodynamic comparisons across adduction and loudness," *Journal of Voice*, vol. 31, no. 6, pp. 773–e11, 2017.

[4] Nancy Pearl Solomon, Gerald N McCall, Michael W Trosset, and William C Gray, "Laryngeal configuration and constriction during two types of whispering," *Journal of Speech, Language, and Hearing Research*, vol. 32, no. 1, pp. 161–174, 1989.

[5] Beiming Cao, Myungjong Kim, Ted Mau, and Jun Wang, "Recognizing whispered speech produced by an individual with surgically reconstructed larynx using articulatory movement data," in *Workshop on Speech and Language Processing for Assistive Technologies*. NIH Public Access, 2016, vol. 2016, p. 80.

[6] Roger J. Ingham, Anne K. Bothe, Erin Jang, Lauren Yates, John Cotton, and Irene Seybold, "Measurement of speech effort during fluency-inducing conditions in adults who do and do not stutter," *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 5, pp. 1286–1301, 2009.

[7] Megan J Osfar, "Articulation of whispered alveolar consonants," 2011.

[8] Boon Pang Lim, *Computational differences between whispered and non-whispered speech*, University of Illinois at Urbana-Champaign, 2011.

[9] Shabnam Ghaffarzadegan, Hynek Boşil, and John H. L. Hansen, "Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5024–5028.

[10] Hamid Reza Sharifzadeh, Ian V. McLoughlin, and Martin J. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. e49–e56, 2012.

[11] Đorđe T. Grozdić and Slobodan T. Jovičić, "Whispered speech recognition using deep denoising autoencoder and inverse filtering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2313–2322, 2017.

[12] Heng-Jui Chang, Alexander H. Liu, Hung-yi Lee, and Lin-shan Lee, "End-to-end whispered speech recognition with frequency-weighted approaches and pseudo whisper pre-training," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 186–193.

[13] Prithvi RR Gudepu, Gowtham P Vadisetti, Abhishek Niranjan, Kinnera Saranu, Raghava Sarma, M Ali Basha Shaik, and Periyasamy Paramasivam, "Whisper augmented end-to-end/hybrid speech recognition system—cyclegan approac," *Proc. Interspeech*, pp. 2302–2306, 2020.

[14] Szu-Chen Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005, vol. 1, pp. I/1009–I/1012 Vol. 1.

[15] Beiming Cao, Myungjong Kim, Ted Mau, and Jun Wang, "Recognizing Whispered Speech Produced by an Individual with Surgically Reconstructed Larynx Using Articulatory Movement Data," in *Proc. 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2016)*, 2016, pp. 80–86.

[16] Stavros Petridis, Jie Shen, Doruk Cetin, and Maja Pantic, "Visual-only recognition of normal, whispered and silent speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6219–6223.

[17] 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," *NASA STI/Recon technical report n*, vol. 93, pp. 27403, 1993.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[19] Kuldeep Khoria, Madhu R. Kamble, and Hemant A. Patil, "Teager energy cepstral coefficients for classification of normal vs. whisper speech," in *European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1–5.

[20] Chang Feng, Xiaolong Wu, Mingxing Xu, and Thomas Fang Zheng, "Parameterization of dominant spectral peak trajectory for whisper speech recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 911–916.

[21] Marius Cotescu, Thomas Drugman, Goeric Huybrechts, Jaime Lorenzo-Trueba, and Alexis Moinet, "Voice conversion for whispered speech synthesis," *IEEE Signal Processing Letters*, vol. 27, pp. 186–190, 2019.

[22] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[23] Gunnar Fant, *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, Number 2. Walter de Gruyter, 1971.

[24] Paavo Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[25] Olivier Perrotin and Ian McLoughlin, "A spectral glottal flow model for source-filter separation of speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7160–7164.

[26] Parham Mokhtari, Brad Story, Paavo Alku, and Hiroshi Ando, "Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production," *Speech Communication*, vol. 104, pp. 24–38, 2018.

[27] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[28] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015.

[29] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[30] Rico Sennrich, Barry Haddow, and Alexandra Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, Aug. 2016, pp. 1715–1725, Association for Computational Linguistics.

[31] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki 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, 2017.

[32] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al., "Recent developments on espnet toolkit boosted by conformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5874–5878.