ENDING THE BLIND FLIGHT: ANALYZING THE IMPACT OF ACOUSTIC AND LEXICAL FACTORS ON WAV2VEC 2.0 IN AIR-TRAFFIC CONTROL

Alexander Blatt, Badr M. Abdullah, Dietrich Klakow

Saarland University, Saarland Informatics Campus, Germany

ABSTRACT

Transformer neural networks have shown remarkable success on standard automatic speech recognition (ASR) benchmarks. However, they are known to be less robust against domain mismatch, particularly with air traffic control (ATC) speech data. In the ATC domain, transformer-based ASR systems do usually not transfer across different datasets. The reasons for poor transferability across ATC datasets remain unclear. Our study investigates the influence of acoustic variability and lexical differences on the ASR performance across various ATC datasets. By fine-tuning and evaluating wav2vec 2.0 on synthetic ATC datasets, we examine the effect of acoustic variability on the model performance. Furthermore, we assess the effect of lexical differences by correlating language model perplexity with performance. Our findings reveal that a combination of acoustic and lexical mismatch causes the bad inter-dataset transferability and give insights on how to improve future ASR models for ATC.

***Index Terms*—** noise, lexical differences, air-traffic control, ASR, wav2vec 2.0

1. INTRODUCTION

Automatic speech recognition (ASR) is the first step in a speech-processing pipeline for air-traffic control (ATC) communication. ATC communication consists of instructions from an air-traffic controller (ATCO) to a specific pilot and read-back from that pilot [1]. In recent years, several corpora for ATC-ASR have been gathered [1]. But apart from the ATCOSIM corpus [2] and an one-hour chunk of the ATCO2 corpus [3], datasets are either not available without a fee or not publicly available at all. Since ATC communication is formalized and has a unique phraseology [4], out-of-domain (OOD) trained ASR models transfer poorly to ASR data [5, 6]. Despite these challenges, some ATC-ASR models have been developed in recent years. While earlier models rely on Kaldi [7], newer approaches are based on pretrained transformer models like wav2vec 2.0 [5]. Although those models are build on several hours of training data, that even incorporate non-publicly available data, high word error rate (WER) variations in-between different ASR benchmark corpora have been observed [7, 1, 5]. Previous works on ATC-ASR focused therefore on using newer or more parameter-rich models to increase the overall performance on the benchmark datasets. In contrast to these works, we will explore the causes for the poor transferability across the different ATC datasets at the example of wav2vec 2.0. This will not only give a better understanding on how to interpret the WERs reached on the individual benchmark datasets, but also allow to develop better ATC-ASR models in the future. In the following sections, we analyze the influence of the acoustic variability. Furthermore, we model the acoustic variability by adding Gaussian noise of different levels to text-to-speech (TTS) generated versions of the datasets. Regarding the lexical differences, we analyze intra and cross-dataset perplexities and out-of-vocabulary (OOV) rates. To get a better understanding of the wav2vec 2.0 adaptation to the ATC corpora, we additionally analyze the internal changes of the wav2vec 2.0 architecture during finetuning on the different ATC corpora. In the next section, we will elaborate related studies in the fields of ATC and explainability of transformer-based ASR.

2. RELATED WORK

Zuluaga-Gomez et al. have trained wav2vec 2.0 and XLS-R for ATC speech recognition and provide results over different ATC datasets [5], the resulting WERs of their best model still show a significant variation over the test datasets. One way to make wav2vec 2.0 more robust has been introduced by Zhu et al. [8]. They force the feature encoder to generate speech representations for a noisy speech input, that resemble representations for clean speech. The resulting model has a superior noise tolerance in comparison to the baseline wav2vec 2.0 model. This shows on the other hand the sensitivity of the standard transformer-based ASR models to noise. Hu et al. [9] have build on this work to develop a wav2vec 2.0 based model, that does speech enhancement without introducing artifacts that deteriorate the ASR performance. A method to deal with the low availability of labeled in-domain data has been proposed by Hsu et al. [10]. They have shown, that if there is no in-domain data available for finetuning, using unlabeled in-domain data during pretraining can give a significant performance improvement. For our wav2vec 2.0 feature analysis, we build on the following two previous

[1]Communication examples: https://wiki.flightgear.org/ATC_phraseology

979-8-3503-0689-7/23/$31.00

Table 1: Word and character error rates across the different ATC datasets depending on the training set. All scores are generated by finetuning and testing *wav2vec2-base* on the datasets, except for the last row, where *wav2vec2-base-960h* is used, which is already finetuned on LibriSpeech. Intra-dataset scores are marked blue.

Training Data	ATCO2		ATCOSIM		LiveATC	
	WER (%)	CER (%)	WER (%)	CER (%)	WER (%)	CER (%)
ATCO2	33.4	20.4	36.6	16.8	61.2	40.3
ATCOSIM	91.9	61.5	2.67	1.00	101.9	67.8
LibriSpeech	99.6	64.6	71.0	32.0	103.4	70.5

works. Phang et al. [11] have shown, that the centered kernel alignment (CKA) similarity scores of text-based transformer models show same-similarity clusters along the diagonal after they are fine-tuned. Choi et al. [12] have shown that the information encoded in an wav2vec feature encoder is analog to a spectrogram and that closer latent representations imply acoustic similarity.

3. EXPERIMENTAL SETUP

The ATC **datasets** used in the following experiments are listed in Table 2. The ATCOSIM corpus [2] consists of simulated conversations between air-traffic controllers and pilots. Since the recordings were done in a controlled environment, the speech is less noisy than for the following two corpora. The ATCO2 corpus [3] contains real ATC conversations from various, mostly European airports and was recorded during the ATCO2 project with VHF-receivers [2]. The LiveATC corpus consists of two subcorpora, LiveATC1 and LiveATC2 [13], both gathered during the ATCO2 project from the LiveATC web-page [3], a web-page broadcasting live ATC conversations. **Finetuning wav2vec 2.0** on ATC data is done by training `wav2vec2-base` [4] for 40 epochs on the train-split of the datasets in Table 2. After finetuning, the checkpoint model with the lowest WER score on the validation set is used for testing.

Table 2: Dataset splits used for the experiments. The mean utterance length for each dataset is roughly four seconds. In the last column, the mean SNR over the full dataset is given.

Dataset	Train	Val	Test	SNR
ATCO2	2739	342	343	13.1
ATCOSIM	2286	286	286	29.4
LiveATC	512	-	518	7.2

To generate the **text-to-speech (TTS)** versions of the aforementioned datasets out of the transcripts, we use the VITS model (Variational Inference with adversarial learning for end-to-end Text-to-Speech) [14] from the Coqui-AI library [5]. The model can be described as conditional variational autoencoder and produces natural sounding speech from text. The male speaker 226 is chosen out of the list of speakers, since it produces the most realistic ATC speech. To generate our synthetic noisy ATC data, we add Gaussian noise to the TTS versions of the datasets.

To overcome the problem of missing clean versions of the ATC datasets to calculate the **signal-to-noise ratio** (SNR), we use the WADA-SNR approach introduced by Kim et al. [15] to get a robust estimate for the SNR. To ensure consistency, all SNR values mentioned in this work are based on this method. Experimental validations on the synthetic noisy ATC datasets have shown that the WADA-SNR scores show just small deviations from the actual SNR values.

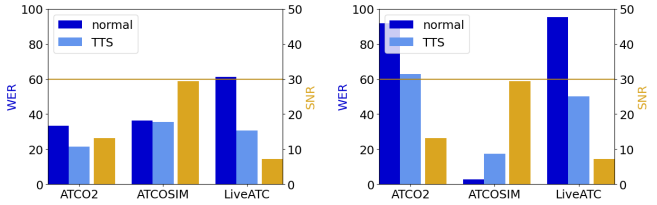
To measure the wav2vec 2.0 **feature similarities**, when finetuned on different datasets, we apply the centered kernel alignment (CKA) method as similarity measure, since it is well defined for small sample sizes, in contrast to other similarity measures like CCA and pwCCA [16]. The output-layer features of the convolutional blocks of the feature encoder and the dense-layer features of the transformer encoder are mean-pooled over the sentence length before comparison.

4. RESULTS

As already observed in previous works [7, 1, 5], the performance of an ASR model varies depending on the target and training dataset. Even if all datasets come from the same domain, namely air-traffic control, the word error rate and character error rate (CER) vary, as Table 1 shows.

However WER and CER correlate across all datasets and there are no dataset specific WER/CER ratios. Without including the intra-dataset scores, the lowest WER/CER ratios are reached on ATCOSIM followed by ATCO2 and LiveATC. This correlates inverse with the SNR values given in Table 2. The last WER column of Table 1 shows the importance of in-domain fine-tuning. Wav2vec 2.0 finetuned on ATCO2 reaches a WER 40-50% lower than the model finetuned on the OOD LibriSpeech corpus. Surprisingly, if wav2vec 2.0 is

[2]Receiver guide: https://ui.atc.opensky-network.org/intro
[3]LiveATC webpage: https://www.liveatc.net/
[4]Wav2vec 2.0 model: https://huggingface.co/facebook/wav2vec2-base
[5]Coqui-AI webpage: https://github.com/coqui-ai/TTS



(a) ATCO2 (b) ATCOSIM

Fig. 1: WER on the standard and TTS versions of the ATC datasets. All scores are generated by fine-tuning and testing *wav2vec2-base* on ATCO2 (a) and ATCOSIM (b) data. The border to clean speech SNR>30 is marked [17].

finetuned on ATCOSIM, this difference is much smaller. In the following, we will evaluate this and analyze which acoustic and lexical differences exist between the datasets and how wav2vec 2.0 reacts to them.

4.1. Acoustic Differences

As already discussed above, there seems to be a correlation between the noise level and the word error rate. To rule-out the influence out-of-vocabulary (OOV) words or other language, respectively lexical based features, we generate text-to-speech (TTS) versions of the datasets, as described in section 3, and compare the WERs reached on the datasets. Since they share the same transcripts, all differences between the TTS and non-TTS versions are due to acoustics. Figure 1 shows the WERs reached on the TTS and non-TTS versions together with the SNR values of the non-TTS versions taken from Table 1. For both training datasets, ATCO2 and ATCOSIM, the difference of the WERs between the TTS and non-TTS versions correlates inverse with the SNR value. This shows, that noise is a major cause for the performance degradation of the ASR models on ATC datasets.

In order to evaluate the effect of the noise over a broad range, we add Gaussian noise of different levels to the TTS versions of the datasets. The results are shown together with

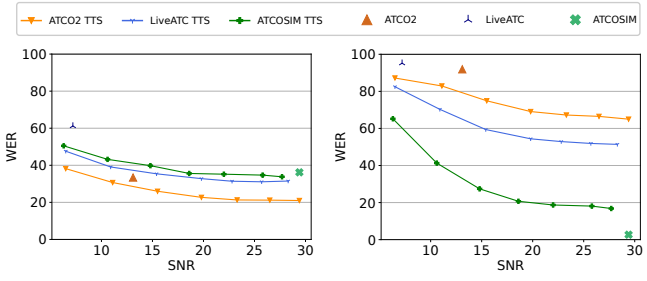
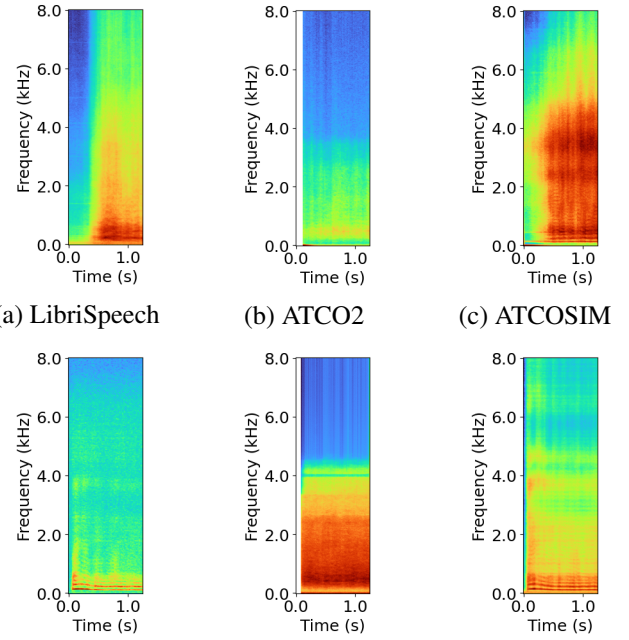


Fig. 2: WERs on the original and TTS data with Gaussian noise of different levels applied. Wav2vec 2.0 is trained on the original ATCO2 (left) or ATCOSIM (right) dataset.



(a) LibriSpeech (b) ATCO2 (c) ATCOSIM

(d) LiveATC Gauss (e) LiveATC (f) LiveATC TTS

Fig. 3: Overlay of spectrograms from 100 samples of the the different ATC datasets and LibriSpeech as reference. The *LiveATC Gauss* spectrogram is based on TTS data with Gaussian noise with an average SNR of 6.5 dB, which is close to the original noise level of LiveATC with 7.2 dB.

the WERs reached on the original datasets in Figure 2. There are four main observations. Firstly, the higher the noise, respectively the lower the SNR, the steeper is the gradient of the curves. For SNR levels over 25, the effect of the noise is negligible, which agrees with the definition of clean speech for SNRs>30 of Grimaldi et al. [17]. The second observation is, that the training dataset not only influences the overall WER reached on the test set, but also the noise sensitivity (gradient). The model trained on ATCO2 data (Figure 2 left) shows a significant lower sensitivity to noise than the model trained on the less noisy ATCOSIM data (Figure 2 right). The third observation is that for high noise levels with a SNR<10, the model trained on ATCO2 outperforms the ATCOSIM model on the ATCOSIM test data. This indicates that for high noise target datasets, matching the noise distribution during training can become more important than lexical similarities between the training and test set. The last observation is, that wav2vec 2.0 reaches slightly higher WERs on the non-TTS test sets of ATCO2 and LiveATC than on the TTS versions with the same noise level.

To examine this difference further, we overlay the spectrograms of 100 samples from each ATC dataset. To allow an overlay, each recording is trimmed to the same length and the spectrograms are normalized. Figure 3 shows the resulting spectrograms. The comparison of the datasets shows that

Table 3: Lexical diversity of the ATC datasets, measured with the moving average type-token ratio (MATTR) and the measure of textual lexical diversity (MTLD)

Dataset	MATTR	MTLD
ATCO2	0.635	29.5
ATCOSIM	0.585	26.6
LiveATC	0.581	23.3

each dataset has a unique noise characteristic. In the ATCO2 dataset, the harmonics of the voice stand far less out against the background noise than the harmonics in the ATCOSIM dataset. Additionally, the LiveATC and ATCO2 dataset spectrograms show a low-pass characteristic, with a loss of signal power over 4 kHz. Additionally the LiveATC dataset shows a narrow-band signal loss exactly at 4 kHz. The spectrogram of the LiveATC TTS data with Gaussian noise (SNR = 6.5), noticeably differs from the standard LiveATC spectrogram (SNR =7.2). Meaning that the WADA-SNR scores do not reveal the complexity of the noise. This explains why the WER curves on the TTS datasets in Figure 2 are lower bounds for the WERs reached on the original datasets. To reproduce the original noise for each dataset, more complex noise types, like band-pass or low-pass filters must be included. In the next section we will evaluate the lexical differences between the datasets.

4.2. Lexical differences

We have shown that there exists a correlation between the noise and the WER reached on the datasets. In this section, we will evaluate if there is a similar correlation for the lexical features. To get a better understanding for the complexity of the datasets, the lexical diversity (LD) is measured via moving average type-token ratio (MATTR) and the measure of textual lexical diversity (MTLD), which are better estimates for the lexical diversity than other measures, as shown by Tager-Flusberg et al. [18]. Table 3 shows the diversity scores of the datasets. The LiveATC dataset has the lowest MATTR and MTLD score, indicating that it has the lowest lexical diversity. But the small difference of just 9% to the highest MATTR score, measured on the ATCO2 dataset, shows that the three datasets have a quite similar lexical diversity.

Table 4: Cross (black) and intra-dataset (blue) perplexities. 4-gram language models are generated for each training dataset.

Training Data	Perplexity on test data		
	ATCO2	ATCOSIM	LiveATC
ATCO2	24.8	138.0	88.2
ATCOSIM	**417.2**	4.8	276.5
LiveATC	144.6	120.4	25.6

Table 5: Cross and intra-dataset (blue) OOV rates in percent.

Training Data	OOV rate on test data (%)		
	ATCO2	ATCOSIM	LiveATC
ATCO2	2.05	7.24	5.20
ATCOSIM	**27.0**	0.65	18.02
LiveATC	11.17	12.90	3.23

To find more substantial lexical differences, we calculate the cross and intra-dataset perplexities using 4-gram language models (LM). All LMs are generated on the train-splits and tested on the test-splits of the datasets, Table 4 shows the results. The highest cross-dataset perplexities are found on the ATCO2 test dataset, indicating the worst transferability of an ASR model trained on the other datasets to this dataset. For the intra-dataset perplexities, the LiveATC and ATCO2 dataset have similar scores, while the ATCOSIM $\rightarrow$ ATCOSIM perplexity is five times lower. This shows, that the simulated scenarios in ATCOSIM do not have the variability of the operational recordings found in the ATCO2 and LiveATC corpora. This could also be due to the case, that the ATCO2 and LiveATC datasets cover multiple airspaces as stated in section 3. If ATC conversations are recorded in different airspaces for different datasets, this has consequences on the vocabulary. Each airspace has different waypoints, is targeted by different airlines and uses different communication frequencies, to just name a few differences. This also shows in the OOV rates, which can be seen in Table 5. The comparison of Table 4 and Table 5 shows that the perplexities and the OOVs correlate, with one exception. On the ATCOSIM test data, lower OOV rates are reached with ATCO2 source data, than with LiveATC source data, while it is the other way around for the perplexity. An inspection of the OOVs shows, that in the case LiveATC $\rightarrow$ ATCOSIM, the OOVs contain many German words, like airline names, city names and greetings. These OOVs are missing in the ATCO2 $\rightarrow$ ATCOSIM case, likely due to the recordings from Swiss airspaces in the ATCO2 dataset.

Since both, perplexity and OOV rates show a lexical mis-

Table 6: Relative WER drop in percent (%), when using a 4-gram LM generated on the train-split of the target dataset. Testing is done on the test-splits of the target datasets. Mean scores over all target-source dataset combinations are given for TTS and non-TTS versions. The absolute difference is given in brackets.

Source Data	rel. WER drop on target data (%)	
	normal	TTS
normal	21.6 (53.42-44.95)	**27.6** (36.4-27.3)
TTS	11.9 (89.01-79.71)	22.8 (19.55-15.58)

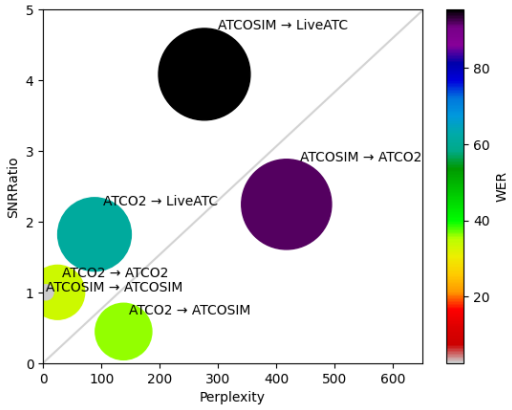


Fig. 4: WER depending on the relative difference between test and training SNR and the perplexity of a LM generated from training data and evaluated on test data.

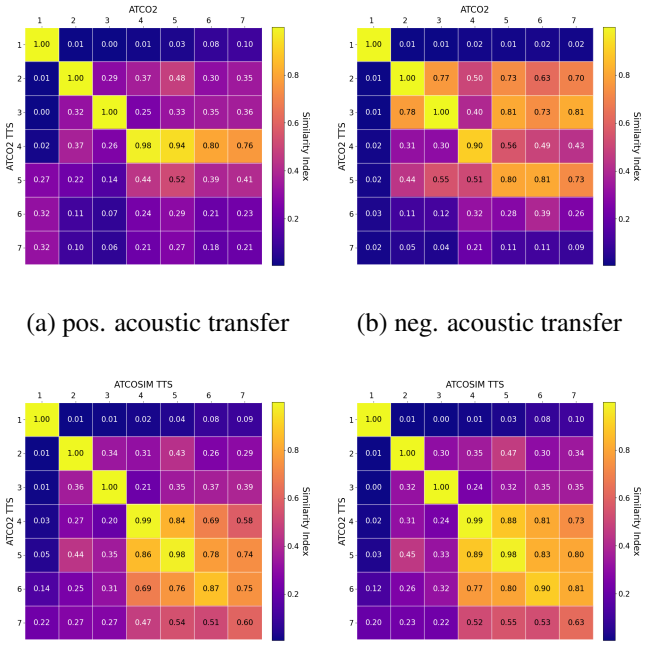
match, we want to quantify to which extend this can be fixed by using a 4-gram LM trained on the train-split of the target dataset. Table 6 shows the mean results over all source and target dataset combinations, using ATCO2 and ATCOSIM as source data and ATCO2, ATCOSIM and LiveATC as target data. For both, source a target data, either TTS or non-TTS versions of the datasets are used.

The resulting scores show that adding the LM on top of wav2vec 2.0 results in the highest improvement for the non-TTS (train) $\rightarrow$ TTS (test) setting. Interestingly, the relative improvement for TTS $\rightarrow$ TTS and non-TTS $\rightarrow$ non-TTS is nearly equivalent. This shows that even if there is an acoustic mismatch, adaptation to the target airspaces via LM can bring a big improvement. In the worst case scenario, TTS $\rightarrow$ non-TTS, where wav2vec 2.0 has never seen noisy data during training, there is still more than 11% improvement.

Since the influences of lexical differences and noise variability have been laid out, the question is, if there is an overall clear dependency of the WER on the ratio between the lexical differences and the noise differences. To evaluate this, we plot the WER in dependence of the ratio between the source-target LM perplexity and the source-target SNR-ratio. The resulting Figure 4 shows, that the aforementioned dependency exists. This explains the different WERs reached on the datasets, depending on the selection of the training and test set. It furthermore opens the door for future research on predicting the WER for unknown (ATC) benchmark datasets. While we have focused mostly on dataset features until now, we will look also at wav2vec 2.0 features in the next section.

4.3. wav2vec adaption

To better understand how wav2vec 2.0 adapts to the lexical differences and the acoustic variability between the different ATC datasets, we use CKA to compare the features of the different parts of the model. We examine four different cases.



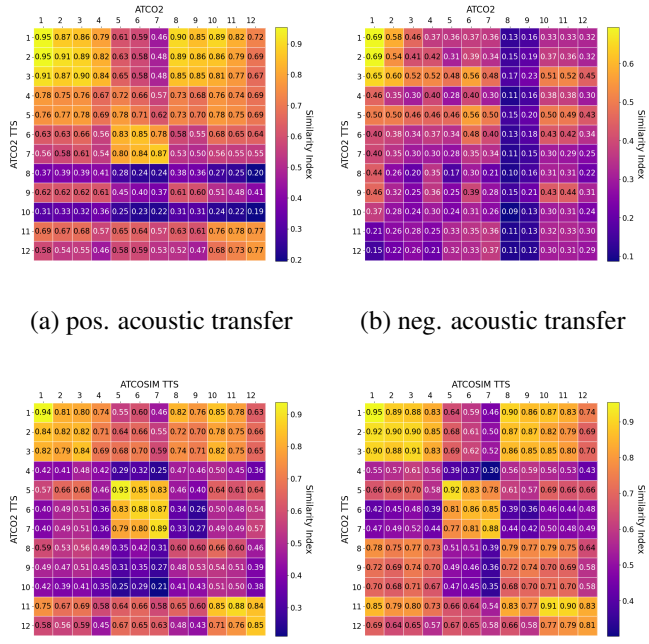
(a) pos. acoustic transfer (b) neg. acoustic transfer

(c) pos. lexical transfer (d) neg. lexical transfer

Fig. 5: CKA analysis on the adaptation of the wav2vec feature encoder to acoustic and lexical changes. The CKA scores in (a) are produced on ATCO2 TTS data and the scores on (b) on ATCO2 data. The CKA scores in (c) are produced on ATCOSIM TTS data and the scores in (d) ATCO2 TTS data. All scores are given on the output layers of each convolutional layer of the feature encoder.

In the first two cases, we look at the feature similarity between two models, when one of them encounters a dataset with new acoustic properties during testing. For the positive acoustic transfer, we compare the CKA scores of wav2vec 2.0 finetuned on ATCO2 and ATCO2 TTS data and tested on ATCO2 TTS data. This is labeled as positive transfer case since wav2vec 2.0 trained on ATCO2 reaches a WER of 21.5% on the unseen ATCO2 TTS data, which is a significant decrease from the 33.4 % WER on ATCO2 test data. For the negative acoustic transfer, the wav2vec 2.0 model finetuned on ATCO2 TTS data encounters a new dataset. Wav2vec 2.0 trained on ATCO2 TTS reaches a WER of 5.6 % on ATCO2 TTS test data but the score increases about a factor of 17 to a WER of 96.7% on the unseen ATCO2 dataset.

Figure 5 shows the CKA similarity scores of the wav2vec 2.0 feature encoder in the positive (a), respectively negative acoustic transfer case (b). Interestingly, the initial and the intermediate layers show even a higher similarity for the negative acoustic transfer case. But the similarity score on the final layer of the feature encoder reaches 0.21 in the positive scenario, while for the negative scenario, the similarity



(a) pos. acoustic transfer

(b) neg. acoustic transfer

(c) pos. lexical transfer

(d) neg. lexical transfer

Fig. 6: CKA analysis: Adaptation of the wav2vec transformer encoder layers to acoustic (a) and (b) and language changes (c) and (d). The test datasets are equal to Figure 5.

score is considerably lower with just 0.09. This difference also propagates through the dense transformer encoder layers as Figure 6 (a) and (b) show. Even in the first layer of the transformer encoder, the scores differ already significantly with 0.95 and 0.69. Towards the final layer, the difference further increases. Additionally, the CKA plot of the negative acoustic transfer does not show the typical clusters of similar representations, which can be found along the diagonal after finetuning, as observed by Phang et al. [11]. For the positive acoustic transfer, there are three non-symmetric clusters visible. This higher similarity shows, that if wav2vec2.0 is trained on noisier data, it is still able to produce good output features on the cleaner dataset.

For the last two cases, we look at the similarity scores for the case, that one model encounters a dataset with different lexical properties during testing. To exclude acoustic influences, we purely use TTS data. For both, negative and positive lexical transfer, we plot the CKA similarity scores for wav2vec 2.0 finetuned on ATCOSIM TTS and ATCO2 TTS. If wav2vec 2.0 gets finetuned on ATCO2 TTS, the WER on ATCO2 TTS is 5.6%, while the WER on ATCOSIM TTS is 17.4%, which is an increase of a factor of 3, but still an above average WER for an ATC dataset as Table 1 and Figure 1 show. We therefore use this scenario as positive lexical transfer. In contrast, if wav2vec 2.0 gets finetuned on ATCOSIM TTS, the WER on ATCOSIM TTS is 2.1%, while the WER on ATCO2 TTS is 47.0%, which is more than 20 times higher. This case is therefore labeled as negative lexical transfer. The comparison of the similarity scores of the feature encoder, Figure 5(c) and (d), shows that there is no significant difference between the positive and negative case. In other words, the feature encoder is agnostic to lexical differences. For the transformer encoder layers, there are evident visual differences between the positive and negative lexical transfer. The fact, that the differences are not as big as for the acoustic transfer needs further investigation. The comparison between lexical and acoustic transfer however shows, that without the presence of noise, a cluster of similar representations in the intermediate layers of the transformer encoder is forming, which is more prominent for the positive transfer case. Since this cluster is also partially forming in the positive, but not in the negative acoustic transfer, it could be a possible candidate to indicate a good lexical and acoustic transferability.

5. CONCLUSION

Pretrained transformer-based speech recognition models, like wav2vec 2.0 have shown a remarkable performance on low-resource domains. But for the air-traffic control domain, a highly variable transferability across different datasets has been observed. In this paper, we have presented an empirical study to identify the causes of this phenomenon. We demonstrated that each ATC dataset has specific noise characteristics. Nevertheless, adding Gaussian noise to clean air-traffic control data can be used to get a lower WER bound for different noise levels. This is an effective way to estimate the robustness of the ATC-ASR model. We have furthermore shown that there are significant lexical differences between the datasets and that the transferability correlates with cross-dataset language model perplexities as well as with the OOV rates. Dominant OOV entities are airspace-dependent cities, greetings and airlines. A target-dataset specific language model on top of wav2vec 2.0 was identified as an effective method to significantly reduce lexical mismatch and therefore the WER, even for very noise target data. With various source and target-dataset pairings, we have provided evidence for the dependency of the WER on the ratio between the source-target LM perplexity and the source-target SNR-ratio. A final wav2vec 2.0 feature analysis demonstrated, that the feature encoder is agnostic to lexical changes while adapting to different noise scenarios. Finally, we identified a same similarity cluster between the intermediate-layer-transformer-encoder features of the target and source-data-finetuned wav2vec 2.0 models as indicator for a good transferability of the source-model to the target data. The insights of this work not only allow the development of better ATC-ASR models, but also better ASR models for other domains, where poor cross-dataset transferability is observed.

6. REFERENCES

[1] Juan Zuluaga-Gomez, Petr Motlicek, Qingran Zhan, Karel Vesely, and Rudolf Braun, "Automatic speech recognition benchmark for air-traffic communications," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, vol. 2020-Octob, pp. 2297–2301.

[2] Konrad Hofbauer, Stefan Petrik, and Horst Hering, "The ATCOSIM corpus of non-prompted clean air traffic control speech," in *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, 2008, pp. 2147–2152.

[3] Juan Zuluaga-Gomez, Karel Veselý, Igor Szöke, Petr Motlicek, Martin Kocour, Mickael Rigault, Khalid Choukri, Amrutha Prasad, Seyyed Saeed Sarfjoo, Iuliia Nigmatulina, Claudia Cevenini, Pavel Kolčárek, Allan Tart, and Jan Černocký, "ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications," pp. 1–29, 2022.

[4] Hartmut Helmke, Michael Slotty, Michael Poiger, Damián Ferrer Herrer, Oliver Ohneiser, Nathan Vink, Aneta Cerna, Petri Hartikainen, Billy Josefsson, David Langr, Raquel García Lasheras, Gabriela Marin, Odd Georg Mevatne, Sylvain Moos, Mats N. Nilsson, and Mario Boyero Pérez, "Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04," in *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. dec 2018, vol. 2018-Septe, Institute of Electrical and Electronics Engineers Inc.

[5] Juan Zuluaga-Gomez, Amrutha Prasad, Iuliia Nigmatulina, Seyyed Saeed Sarfjoo, Petr Motlicek, Matthias Kleinert, Hartmut Helmke, Oliver Ohneiser, and Qingran Zhan, "How Does Pre-Trained Wav2Vec 2.0 Perform on Domain-Shifted Asr? an Extensive Benchmark on Air Traffic Control Communications," *2022 IEEE Spoken Language Technology Workshop, SLT 2022 - Proceedings*, pp. 205–212, 2023.

[6] Aravind Krishnan, Jesujoba Alabi, and Dietrich Klakow, "On the N-gram Approximation of Pre-trained Language Models," 2023.

[7] Martin Kocour, Karel Veselý, Alexander Blatt, Juan Zuluaga Gomez, Igor Szöke, and Jan Černocký, "Boosting of contextual information in ASR for air-traffic call-sign recognition," pp. 2993–2997, 2021.

[8] Qiu Shi Zhu, Jie Zhang, Zi Qiang Zhang, Ming Hui Wu, Xin Fang, and Li Rong Dai, "a Noise-Robust Self-Supervised Pre-Training Model Based Speech Representation Learning for Automatic Speech Recognition," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, no. 62101523, pp. 3174–3178, 2022.

[9] Yuchen Hu, Chen Chen, Qiushi Zhu, and Eng Siong Chng, "Wav2code: Restore Clean Speech Representations via Codebook Lookup for Noise-Robust ASR," pp. 1–12, 2023.

[10] Wei Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 3, pp. 2123–2127, 2021.

[11] Jason Phang, Haokun Liu, and Samuel R. Bowman, "Fine-Tuned Transformers Show Clusters of Similar Representations Across Layers," *BlackboxNLP 2021 - Proceedings of the 4th BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 529–538, 2021.

[12] Kwanghee Choi and Eun Jung Yeo, "Opening the Black Box of wav2vec Feature Encoder," 2022.

[13] Juan Zuluaga-Gomez, Karel Veselý, Alexander Blatt, Petr Motlicek, Dietrich Klakow, Allan Tart, Igor Szöke, Amrutha Prasad, Saeed Sarfjoo, Pavel Kolčárek, Martin Kocour, Honza Černocký, Claudia Cevenini, Khalid Choukri, Mickael Rigault, and Fabian Landis, "Automatic Call Sign Detection: Matching Air Surveillance Data with Air Traffic Spoken Communications," *Proceedings*, vol. 59, no. 1, pp. 14, dec 2020.

[14] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," 2021.

[15] Chanwoo Kim and Richard M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2598–2601, 2008.

[16] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton, "Similarity of neural network representations revisited," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 6156–6175, 2019.

[17] Vincent Grimaldi, Gilles Courtois, and Hervé Lissek, "Objective evaluation of static beamforming on the quality of speech in noise," , no. c, pp. 369–374, 2018.

[18] Helen Tager-Flusberg, "The Development of English as a Second Language With and Without Specific Language Impairment: Clinical Implications," *Journal of Speech, Language, and Hearing Research*, vol. 24, no. 2, pp. 1–14, 2015.