Synthetic Audio Generation Framework for Air Traffic Control Speech Recognition

*Raphaël Bagat[1], **Zhe Zhang[2]**, Junichi Yamagishi[2], Irina Illina[1], Emmanuel Vincent[1]*
1 Université de Lorraine, CNRS, Inria, LORIA – F-54000 Nancy, France
2 National Institute of Informatics, Tokyo, Japan

Introduction & Problem Statement

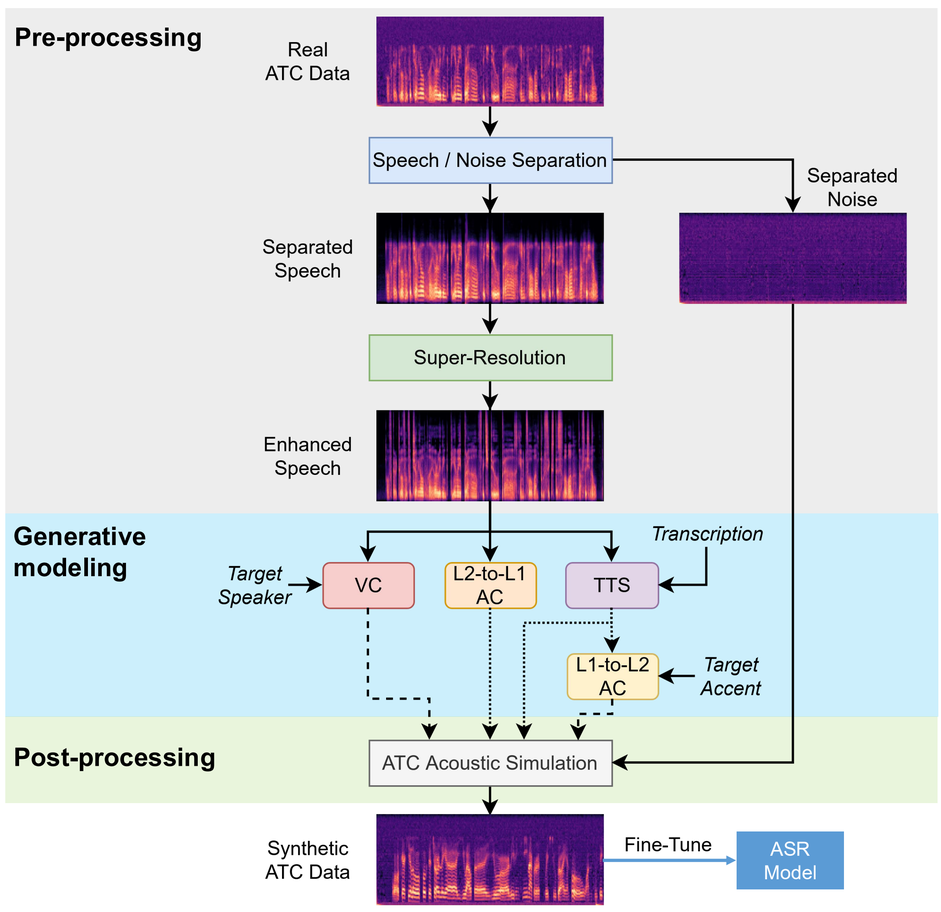
- ATC ASR must handle non-native accents, radio noise and rapid, specialized phraseology with little transcribed data.



- Data augmentation can mitigate the scarcity of transcribed ATC speech.
- Traditional methods modify existing signals, e.g., speed perturbation, noise injection and SpecAugment.
- We generate additional speech from existing recordings and transcripts, varying accents, speakers and acoustic conditions.
- Check out our **paper** and **code**!

Objective: generate synthetic data from existing real data to train an ASR system that better recognizes ATC speech.

Proposed Methodology

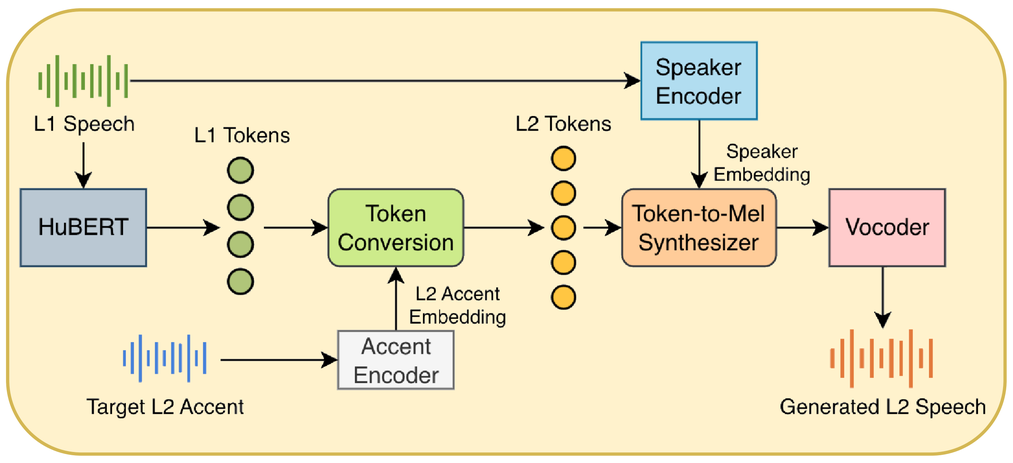


Generative Modeling

- **TTS** (F5-TTS [3]) generates L1 speech from ATC text and a voice reference.
- **VC** (kNN-VC [4]) varies speaker identity using L2-ARCTIC target speakers.
- **L2-to-L1 AC** (TokAN [5]) normalises non-native speech towards native.
- **L1-to-L2 AC** (ours) converts speech to a controllable target L2 accent.

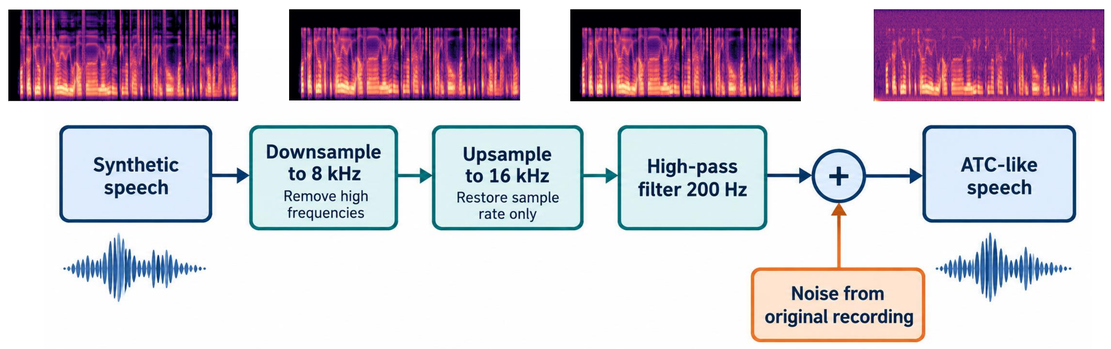
L1-to-L2 Accent Conversion

- A novel **controllable L1-to-L2 accent conversion** module.
- An optional speaker reference enables joint voice conversion. The L2 reference controls the target accent.



ATC Acoustic Simulation (AAS)

- Makes the generated audio match ATC acoustic conditions.



Experiments, Results & Conclusion

Experimental settings

- Whisper-small, evaluated on real ATCO2 speech [6].
- 4-fold cross-validation on 4 h of transcribed speech.
- 72 min AC training, 72 min ASR training, 36 min validation and 60 min testing.
- Synthetic-only or Real+Synth training (50/50 per batch).

Method	AAS	Target Accent	VC (L1→L2)	WER (%)	
				Synth-only	Real+Synth
TTS	×	-	-	53.88	22.90
TTS	✓	-	-	33.77	**21.69**
VC	×	-	-	26.46	22.45
VC	✓	-	-	**24.18**	22.54
$AC_{L2\rightarrow L1}$	✓	-	-	41.60	25.92
$AC_{L1\rightarrow L2}$	×	L2-ARCTIC	Yes	38.22	**21.64**
$AC_{L1\rightarrow L2}$	✓	ATCO2	No / Yes	33.84 / 31.40	22.60 / **22.00**
$AC_{L1\rightarrow L2}$	✓	L2-ARCTIC	No / Yes	32.96 / 31.48	**22.16** / **21.95**
VC + TTS	✓	-	-	34.39	**22.00**
$AC_{L1\rightarrow L2}$ + TTS	✓	L2-ARCTIC	Yes	28.90	**21.69**
$AC_{L1\rightarrow L2}$ + VC	✓	L2-ARCTIC	Yes	32.20	**22.07**
$AC_{L1\rightarrow L2}$ + VC + TTS	✓	L2-ARCTIC	Yes	28.05	22.91
Out-of-the-box	-	-	-	63.32	
Fine-tuning with real data only	-	-	-	22.69	

Results

- Synthetic-only training improves on the 63.32% baseline. VC + AAS performs best (24.18% WER).
- AAS improves synthetic-only TTS: 53.88% to 33.77% WER. L2-to-L1 AC worsens Real+Synth WER to 25.92%.
- L1-to-L2 AC gives the best Real+Synth WER: 21.64% vs. 22.69% real-only (4.6% relative reduction). Several methods tie statistically.
- **The best result uses L2-ARCTIC accent + VC, without AAS. Mixing all generators adds no gain.**

Conclusion

- Generative augmentation improves ATC ASR without collecting additional real recordings.
- It introduces a **controllable L1-to-L2 accent conversion** module to increase the accent diversity of ATC speech.
- Mixing real data with L1-to-L2 accent-converted speech, or with TTS + AAS, **significantly outperforms** fine-tuning on real data alone.

This work was funded by the DeepMAUVES project supported by DGA and CNRS, by the Inria-NII TrustedSpeech Associate Team, and MEXT KAKENHI Grants (24K21324). It was granted access to the HPC resources of IDRIS under the allocation 2024-AD011015024 made by GENCI.

[1] Y.-X. Lu et al., "DAIEN-TTS: Disentangled audio infilling for environment-aware text-to-speech synthesis," in ICASSP 2026.
[2] H. Liu et al., "AudioSR: Versatile audio super-resolution at scale," in ICASSP 2024.
[3] Y. Chen et al., "F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching," in ACL 2025.
[4] M. Baas et al., "Voice conversion with just nearest neighbors," in Interspeech 2023.
[5] Q. Bai et al., "Accent normalization using self-supervised discrete tokens with non-parallel data," in Interspeech 2025.
[6] J. Zuluaga-Gomez et al., "ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of air traffic control communications," Journal of Data-centric MLR, 2024.