

Voice & dialect conversion for low-resource Japanese dialects

Ding Ma, Aanchan Mohan, Luke Strgar, Xinglin Yu, Eden Wagari

Mentor: Shubham Bansal

Why Is This an Interesting Scientific Problem?

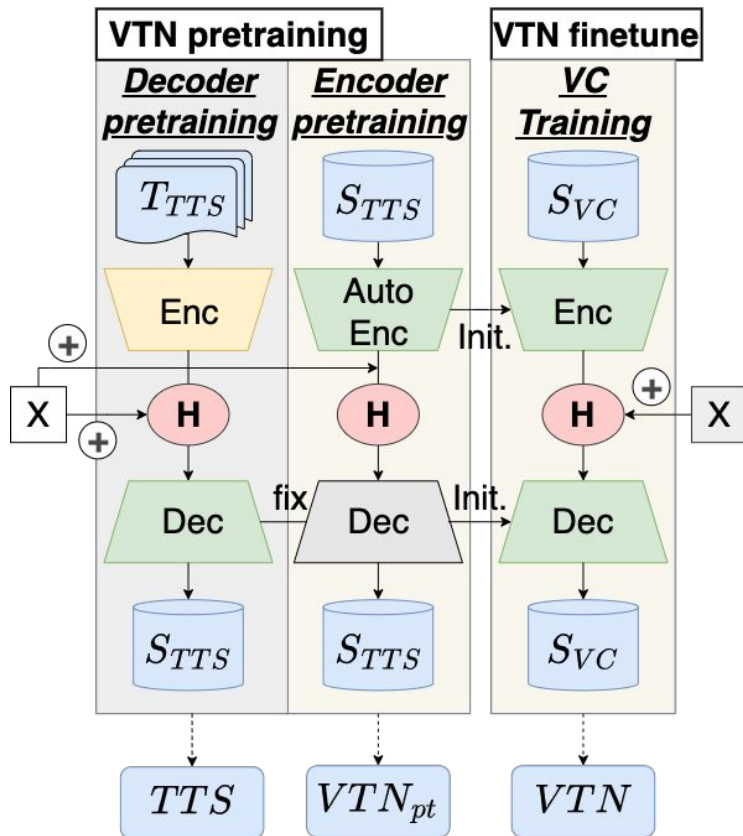
- (Inter)-dialect conversion occupies a unique space as a hybrid task of translation + voice and/or accent conversion.
- Successful systems must be sensitive to prosody, accent, speaker identity, and in some cases linguistic content.
- Designing system architecture and training objectives implicitly asks the extent to which different aspects of a speech signal are separable
 - In the raw waveform
 - In the language's socio-cultural context

Datasets

- We consider (standard) Japanese as our “high resource” language and use the JSUT dataset[2] in pre-training tasks.
- We finetune downstream voice and dialect conversion tasks on small parallel speech corpora of 5 Japanese regional dialects[1] - Kansai (OS), Kyushu (HCK), Tokyo (TK), Tohoku (KOU) and San-yo (HSY) **of just 100 utterances!**

1. [Yoshino et al. Parallel Speech Corpora of Japanese Dialects](#)
2. [Sonobe et al. JSUT CORPUS: FREE LARGE-SCALE JAPANESE SPEECH CORPUS FOR END-TO-END SPEECH SYNTHESIS](#)

Pre-Training for Voice & Inter-Dialect Conversion















- Pre-training stage: build rich hidden representations from TTS database.
 - Decoder pre-training: Text-to-speech (TTS) training.
 - Encoder pretraining: The input becomes speech set of TTS database instead of the text.
 - Leverage high-resource, non-dialectal transcribed corpus
- Finetune stage: A new low resource parallel dataset can be used for one-to-one, any-to-one, any-to-any, etc. VC

3. [Huang et al. Pretraining Techniques for Sequence-to-Sequence Voice Conversion](#)

One-to-One Voice Conversion (within dialect)

Source - HCK05	Target - HCK02	GL	PWG
			

One-to-One Inter-Dialect Conversion

Source - HCK02	Target - TK05	PWG JSUT PT	GL MAILABS PT
UTTID: 091 			
UTTID: 092 			
UTTID: 093 			

JSUT PT: JSUT Data set used for TTS pre-training (PT)






MAILABS PT: Judy with a US accent from MAILABS used for TTS PT

PWG: Parallel WaveGan Vocoder used for synthesis

GL: Griffin and Lim Vocoder used for synthesis

Any-to-Any Inter-Dialect Conversion

- To change speaker identity and convert the dialect from a low-resource accent e.g. speakers from the Kyushu (HCK) region to a high-resource dialect from the Tokyo (TK) region an any-to-any multi-speaker system was trained with **x-vectors** to change target speaker identity.

Source - HCK02	Target TK04 - Expected result	TK04 - Multi speaker JSUT PT + PWG	Target TK05 - Expected result	TK05 - Multi-speaker JSUT PT + PWG
				

Comparison to other solutions

- Dialect conversion is a sub task of Speech to Speech translation proposed in an end-to-end fashion in several recent works in [4,5]
- Majority of these works rely on large amount of data to achieve their objective.
- **Our work propose a framework to do VC and inter-dialect conversion in extremely low resource settings leveraging pre-training to achieve good qualitative results.**

4. [Jia et al. Direct speech-to-speech translation with a sequence-to-sequence model](#)

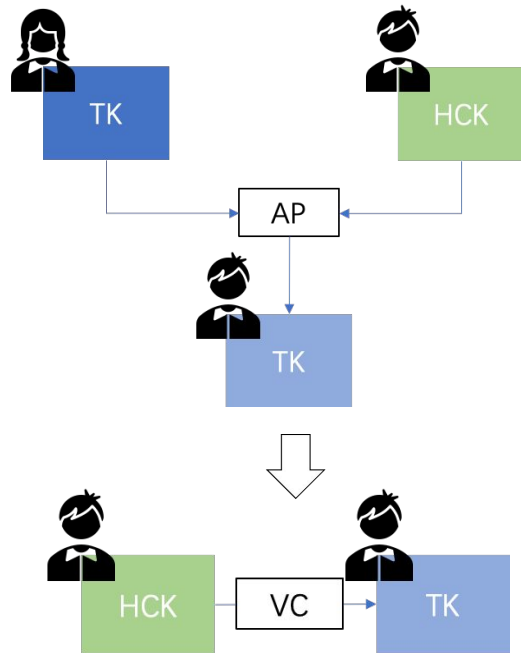
5. [Enaguama et al. Multilingual End-to-End Speech Translation](#)

Conclusion

- Voice and dialect conversion can facilitate technological advancement and inclusion.
- We develop a seq2seq style voice and dialect conversion system for low resource Japanese dialects and demonstrate good qualitative results.
- Our results are unique in that they are accomplished in a very low resource setting and bootstrapped via high resource pre-training
- Opportunities for future work:
 - More rigorous evaluation of VC and development of plausible baseline model for small dialectal corpus
 - Investigation of purely self-supervised representations (foundation models) as initialization for low-resource VC and dialect conversion systems
 - Addition of an adversarial loss term in seq2seq model optimization

Accent Conversion or Preservation

- In seq2seq dialect conversion models are fine-tuned with explicit supervised targets so source speaker accent is not preserved from source.
- Semantically identical utterances in two dialects may have slightly different linguistic contents
- We can simplify the dialect conversion task by generating synthetic data to compose a truly parallel corpus
- We can utilize use an unsupervised disentanglement framework to attempt source speaker accent preservation while changing speaker identity.



Project Formulation

- Initial project formulation asked: **how to leverage high resource language pre-training for low-resource intra-language voice conversion (VC)**
- Project evolution and dataset selection lead to: **how to leverage high resource language pre-training for low-resource intra and inter dialectal voice conversion.**

Instructions

- Please **edit directly on this google slide deck**. During the presentation, you will use a provided laptop for the presentation.
- The final presentation should consist of **3 min presentation + 1-2min QA from judges**. Please stick to the time as we will stop presentations that exceed 5 min.
- In your presentation please consider the following:
 - Goal of the project and what social or economic impact could it create
 - What it makes interesting and/or innovative ?
 - Challenges you have overcome
 - What have you learned from it ?
 - What makes the project special or gives your proposal an edge over similar solutions in the market ?

TIPS and guidelines

- Please do not copy the contents from other materials (if it is very difficult to redraw, it is acceptable with the appropriate citation information).
- It depends on the audience, but it is a good idea to spend some time clearly presenting the introduction/motivation/problem setups
- Use a simple picture to emphasize your method/concept
- Long sentences in slides are not a good idea
- If you are showing numbers, please extract important numbers or highlight important numbers
- Add a take-home message in your final part

VC with limited parallel data leveraging Transfer learning

- Please do not copy the contents from other materials (if it is very difficult to redraw, it is acceptable with the appropriate citation information).
- It depends on the audience, but it is a good idea to spend some time clearly presenting the introduction/motivation/problem setups
- Use a simple picture to emphasize your method/concept
- Long sentences in slides are not a good idea
- If you are showing numbers, please extract important numbers or highlight important numbers
- Add a take-home message in your final part

Multispeaker VC within dialect

- Please do not copy the contents from other materials (if it is very difficult to redraw, it is acceptable with the appropriate citation information).
- It depends on the audience, but it is a good idea to spend some time clearly presenting the introduction/motivation/problem setups
- Use a simple picture to emphasize your method/concept
- Long sentences in slides are not a good idea
- If you are showing numbers, please extract important numbers or highlight important numbers
- Add a take-home message in your final part

Multispeaker VC across dialect

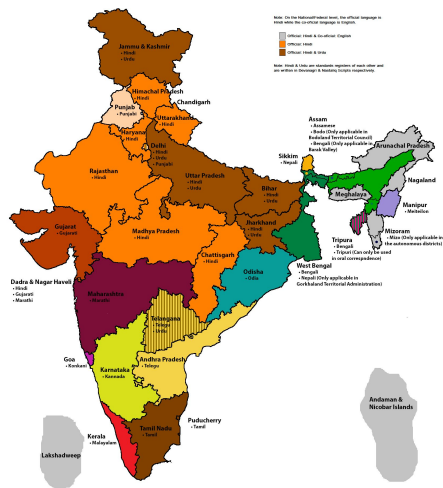
- Please do not copy the contents from other materials (if it is very difficult to redraw, it is acceptable with the appropriate citation information).
- It depends on the audience, but it is a good idea to spend some time clearly presenting the introduction/motivation/problem setups
- Use a simple picture to emphasize your method/concept
- Long sentences in slides are not a good idea
- If you are showing numbers, please extract important numbers or highlight important numbers
- Add a take-home message in your final part

Other things: Wav2vec, accent embedding / preservation

- Please do not copy the contents from other materials (if it is very difficult to redraw, it is acceptable with the appropriate citation information).
- It depends on the audience, but it is a good idea to spend some time clearly presenting the introduction/motivation/problem setups
- Use a simple picture to emphasize your method/concept
- Long sentences in slides are not a good idea
- If you are showing numbers, please extract important numbers or highlight important numbers
- Add a take-home message in your final part

Social Impact: More Inclusion and Harmony

- Automated systems can facilitate improved communication, harmony, knowledge sharing and inclusion between people of different dialect.



India: 22 Major languages but 700 related low resource dialects



Japan: 1 Major language but 16 related low resource dialects