SPEECH-TO-SPEECH TRANSLATION BETWEEN UNTRANSCRIBED UNKNOWN LANGUAGES

Andros Tjandra[1], Sakriani Sakti[1,2], Satoshi Nakamura[1,2]

1) Nara Institute of Science and Technology, Japan 2) RIKEN AIP, Japan

1. Introduction

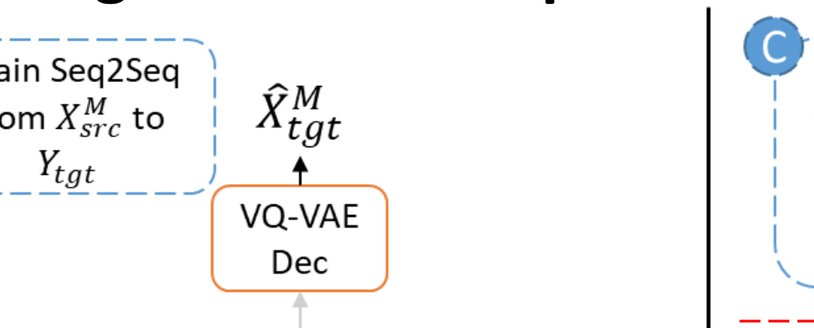
- Speech-to-speech translation overcomes language barrier
- Require paired speech-text -> only several languages can be handled
- Most works only cover speech-to-text translation

We proposed:

- **Direct speech-to-speech translation for unk. languages**
- **No transcription required**

2. Unsupervised Unit Discovery

Speech signal can be disentangled into {contexts, speaking style}



speech chain

Multispeaker VQ-VAE
[de Oord et al., 2017]

$Enc_\theta(x) = q_\theta(y|x)$

$Dec_\phi(y,s) = p_\phi(x|y,s)$

Codebook $E = [e_1, .., e_K]$

Speaker vec $V = [v_1, .., v_L]$

Continuous speech (harder target)

Discrete symbol (easier target)

3. Training & Inference Speech2Code

Training :

A) Train VQ-VAE to get discrete unit representation & Code2Spec to invert the code back to spectrogram

B) Train Seq2Seq to translate source spectrogram X^M_{src} -> target code Y_{tgt}

C) Inference : Source spectrogram X^M_{src} translated into target code Y_{tgt} and recovered into spectrogram $\hat{X}^R_{tgt}$

4. Experiment & Result

Dataset: BTEC (160k train, 510 test), single speaker

Pair: French -> English & Japanese -> English

Speech feature: MFCC (13 dim + Δ + ΔΔ)

Model:

- **Baseline** (direct spectrogram-to-spectrogram)
- **Proposed** SP2C (C=codebook size, T=time reduce)
- **Topline** (speech src-> text tgt*-> speech tgt, ***requires text transcription during training**)

Model	BLEU4	METEOR
Baseline (FR-EN & JA-EN)	Not converged	
SP2C FR-EN C=64, T=12	25	23.2
Topline FR-EN (Cascade) *	47.4	41.2
SP2C JA-EN C=128, T=8	15.3	15.3
Topline JA-EN (Cascade) *	37.4	32.8

5. Discussion & Conclusion

Model	Transcription
Groundtruth	how long are you going to stay
SP2C FR-EN	how long are you going to stay
SP2C JA-EN	how long will it take
Groundtruth	please tell him to call me as soon as he comes in
SP2C FR-EN	please tell him to call me back
SP2C JA-EN	please tell him that i called

incomplete transcription

Based on the example, 1) gives quite close result
However, 2) SP2C result left out the latter part

- We proposed a novel approach for training speech-to-speech translation w/o transcription
- Experiments was performed on French-English & Japanese-English

More samples: **https://sp2code-translation-v1.netlify.com/**