Formosa Speech Recognition Challenge 2020 and Taiwanese Across Taiwan (TAT) Corpus

Taiwanese Across Taiwan

Lugang Longshan Temple

Speaker：Chia-Yu Chang

Authors：Yuan-Fu Liao, Hak-Khiam Tiun, Huang-Lan Su,
Hui-Lu Khoo, Jane S. Tsay, Le-Kun Tan, Peter Kang,
Tsun-Guan Thiann, Un-Gian Iunn, Jyh-Her Yang, Chih-Neng Liang

OUTLINE

- Introduction
  - Endangered Taiwanese
- Taiwanese Across Taiwan (TAT) Project
  - 3 years (2019~2021) project, 300 hours * 6 microphones, 600 speakers
- Status Report
  - TAT-Vol1~2, in total 100 hours
- Formosa Speech Recognition Challenge 2020
  - Free TAT-Vol1, Lexicon, Baseline Recipe
  - Everyone is Welcome to Participate

INTRODUCTION (1/2)

- Endangered Taiwanese
  - Since 1980's, Mandarin became dominant in Taiwan
    - Banned by KMT
  - Most people cannot speak Taiwanese fluently nowadays
  - Especially younger generation
- Difficult Situation
  - Only one Taiwanese TV channel (started last year)
- Hope it's not yet too late to save Taiwanese!

INTRODUCTION (2/2)

- Taiwanese is not Min-Nan
- Taiwanese is Unique

- **From Japan**
  - 以日語形式傳入臺灣(In form of Japanese introduced into Taiwan)
    - 生魚片「さしみ（刺身）」
  - 日本漢字傳入臺灣(Kanji introduced into Taiwan)
    - 招牌 = khang-pang：源自日本語「かんばん（看板）」
  - 日文漢字臺語發音(Same as Kanji's pronunciation)
    - 便當（餐盒） = piān-tong；源自日本語「べんとう（弁当）」
    - 病院（醫院） = pēnn/pīnn-īnn；源自日本語「びょういん（病院）」
  - 日語外來語傳入臺灣(Foreign words in Japanese introduced into Taiwan)
    - 混凝土 - コンクリート，英語 concrete
    - 麵包 - パン，葡萄牙語 pão

- **From Aboriginal**
  - 「臺灣」這詞即來自於南臺灣原住民西拉雅族的「Taian」或「Tayan」
    ("Taiwan" in Taiwanese was came from aboriginal)
  - 茫然不知實情 = a-se（阿西）：源自南部平埔語assey（不明白、不知）

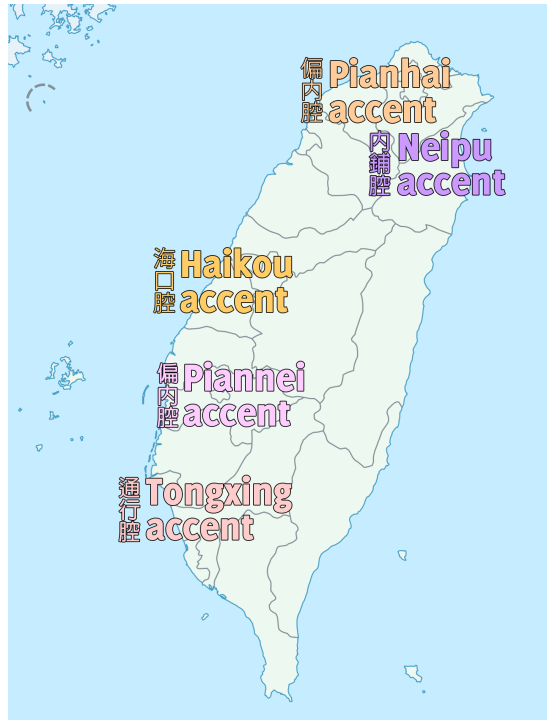
Taiwanese Across Taiwan (TAT) Project

- Project Period
  - 3 years, 2019~2021
- Purpose
  - Taiwanese Speech Recognition
- Target
  - 600 speakers
  - 300 hours * 6 microphones
- Formosa Speech Recognition Challenge 2020
  - Free Taiwanese Speech Corpus
  - Free Lexicon
  - Free Kaldi Baseline Recipe

Taiwanese Across Taiwan

Lugang Longshan Temple

Difficulties



- Most People cannot Speak Taiwanese Fluently
  - Especially Young Generations
- Too many Regional Variations
  - At least 5 variations
- No Standard Writing System
  - Traditional Chinese characters (繁體中文字/Chinese)
    - 現在是晚上八點 (It is eight o'clock now)
  - Taiwanese Southern Min Recommended Characters (台文正字/Taiwanese Hàn-jī)
    - 這馬是暗時八點 (It is eight o'clock now)
  - Taiwan Minnanyu Luomazi Pinyin Fang'an (台羅拼音/Tâi-lô)
    - Tsit4-ma2 si7 am3-si5 peh4-tiam2 (It is eight o'clock now)

Strategy

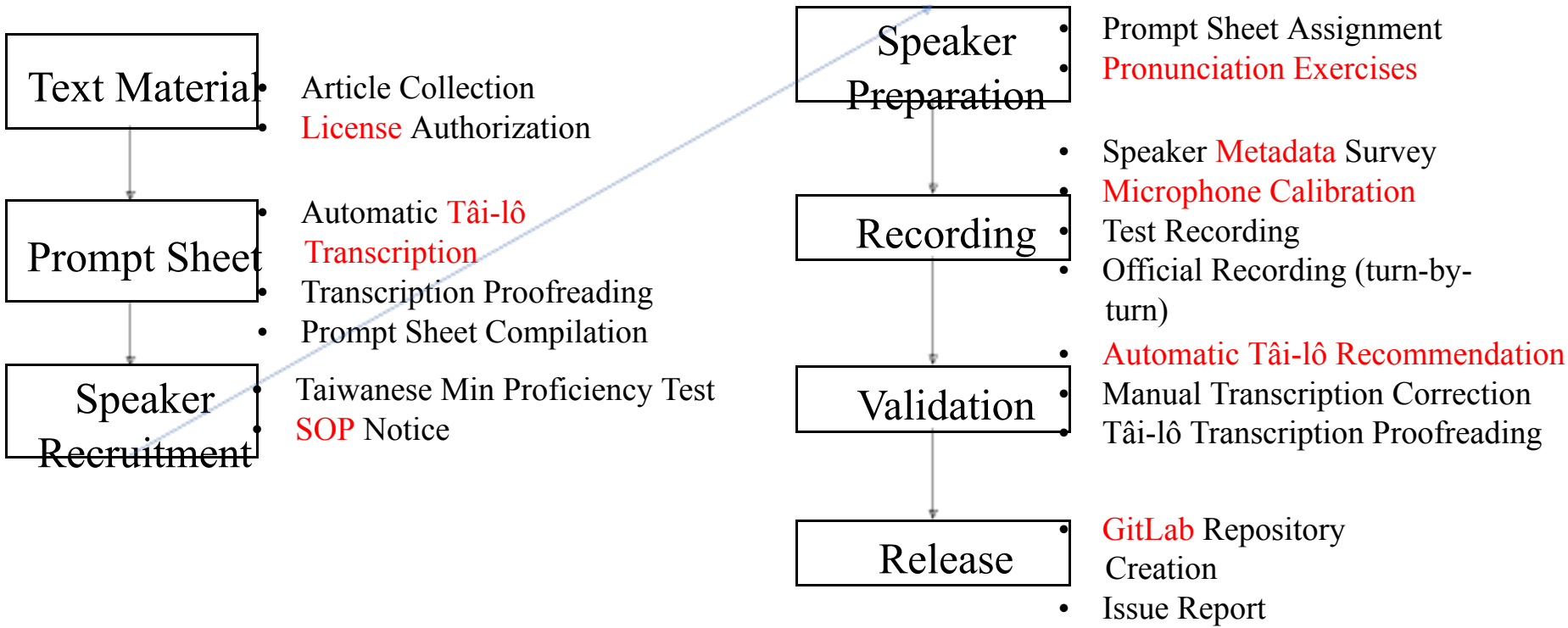
- Recruit Speakers across Taiwan Taiwanese across Taiwan (TAT)
- Adapt Native Taiwanese Article for Prompt Sheets
  - Taiwanese Southern Min Recommended Characters (台文正字/Taiwanese Hàn-jī)
    - 這馬是暗時八點 (It is eight o'clock now)
  - Taiwan Minnanyu Luomazi Pinyin Fang'an (台羅拼音/Tâi-lô)
    - Tsit4-ma2 si7 am3-si5 peh4-tiam2 (It is eight o'clock now)
- Record Reading Speech

  Avoid Transcribing Spontaneous Speech

Partners Across Taiwan

- Hui-lu Khoo, NTNU
- Un-Gian Iunn, Tsun-guan Thiann, NTCU
- Janes S. Tsay, NCCU
- Le-kun Tan, NCKU
- Hak-khiam Tiun, Huang-Lan Su, NTU
- Peter Kang, NDU
- Tân Hong-hūi, 李江卻 Taiwanese Culture and Education Foundation
- Sih4-sing5 hong, Ì-thuân kho-ki

Recording Protocols



Recording Configuration (1/3)

- Native Taiwanese Prompt Sheets

Daily Conversation

1
運動顧健康
ūn-tōng kòo kiān-khong

2
Tsiânn久無見面，你看--起來有khah瘦，
tsiânn kú bô kìnn-bīn, lí khuànn-khí-lâi ū khah sán,

3
koh比進前ke tsiok有元氣！
koh pí tsìn-tsîng ke tsiok ū guân-khì!

4
瘦是
sán

5
幾
kuí-

6
Hó
hóo

Date, Number, Address,…

1
辦公室地址是新莊區中平路439號
pān-kong-sik tē-tsí/tuē-tsí sī Sin-tsng-khu tiong-pîng lōo 439 hō

2
我2016年對大學畢業
guá jī khòng it liok nî uì tāi-hák pit-giap

3
上元節是佇舊曆正月十五
siōng-guân tsiat sī tī kū-lik tsiann--guėh/tsiann--gėh/tsiann-guėh/tsiann-gėh tsap-gōo

4
七點半上北的車敢閣有票？
tshit tiám-puànn tsiūnn-pak ê tshia kám koh ū phiò?

Short Article (~6000 characters)

1
That車
that-tshia

2
啊，頭前毋知閣that偌長？
ah, thâu-tsîng m̄ tsai koh that guā-tn̂g?

3
Tshuā一隻烏貓，騎掃梳飛，真出名的阿琪：
tshuā tsi̍t tsiah oo-niau, khiâ sàu-se/sàu-sue pue/pe, tsin tshut-miâ ê a-kî:

4
長袂過你頭殼內看袂透尾--的，
tn̂g bē/buē kuè/kè lí thâu-khak-lāi khuànn bē/buē thàu-bué-ê,

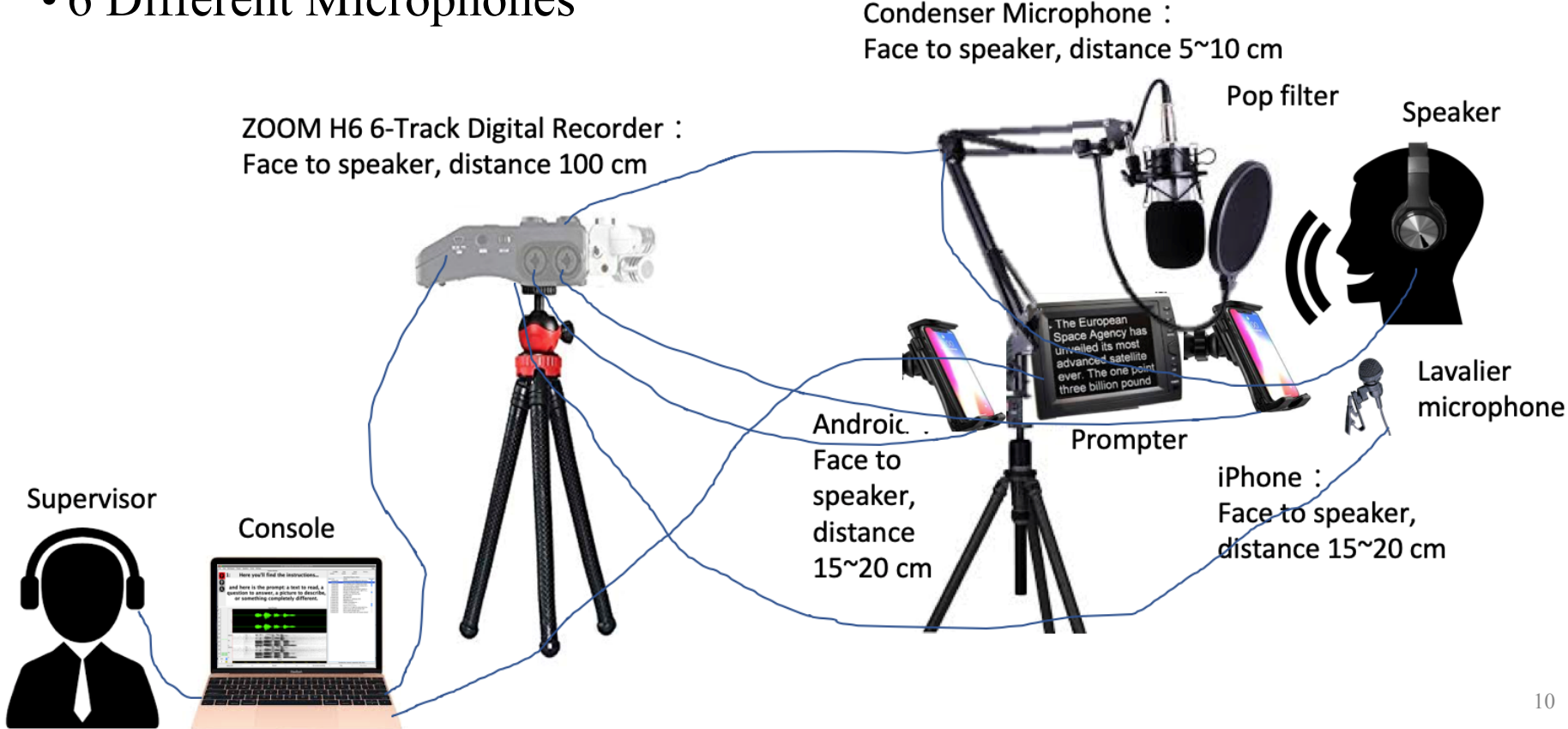
5
遐--的急欲愛的ǹg望！
hia-ê kip beh/bueh ài ê ǹg-bāng!

6
外口便若that車，
guā-kháu piān-nā that-tshia,

7
一寡人的頭殼內，

Recording Configuration (2/3)

- 6 Different Microphones

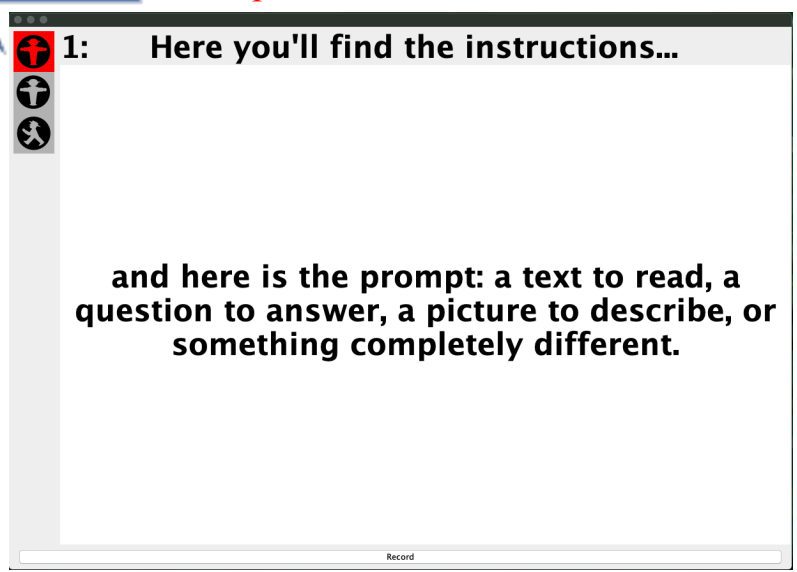


Recording Configuration (3/3)

- SpeechRecorder: https://www.bas.uni-muenchen.de/forschung/Bas/software/speechrecorder/
- Dual Monitors
- Recording Scripts

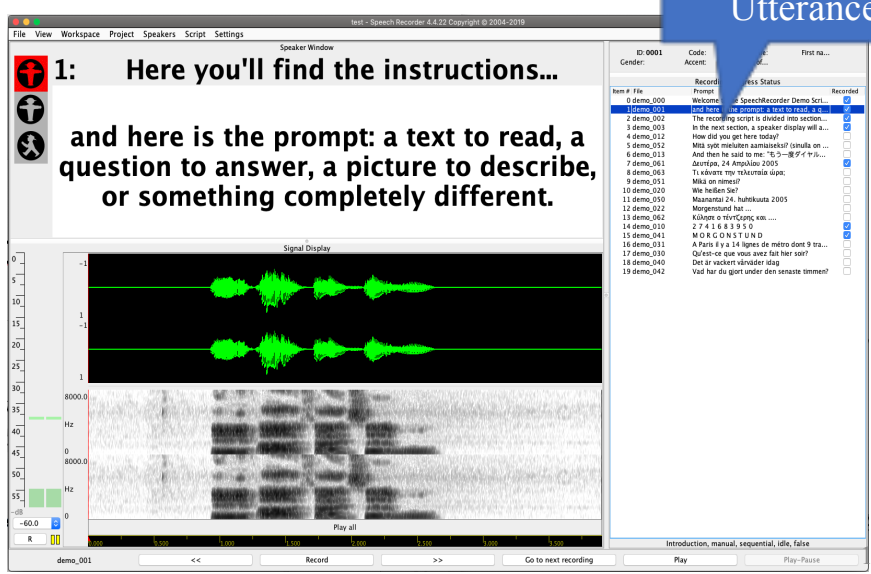
Status

Speaker Monitor

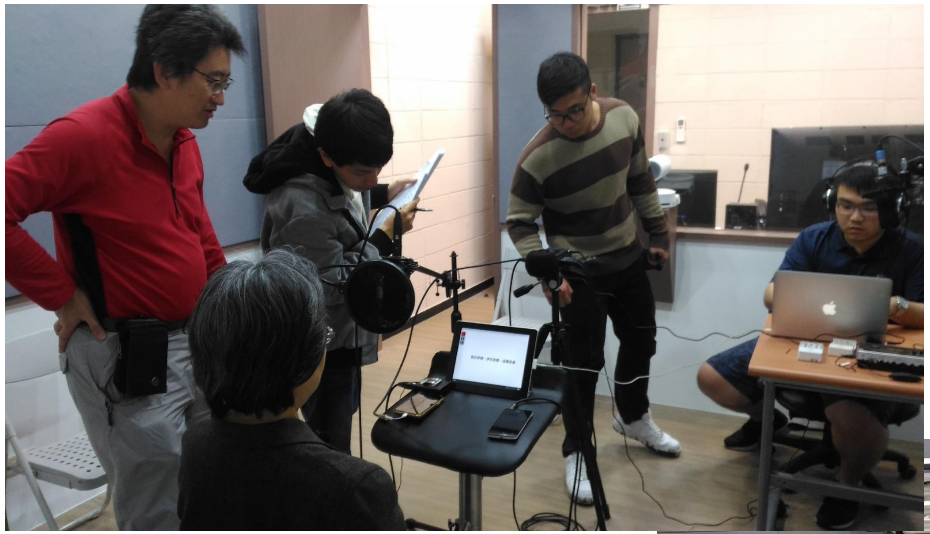


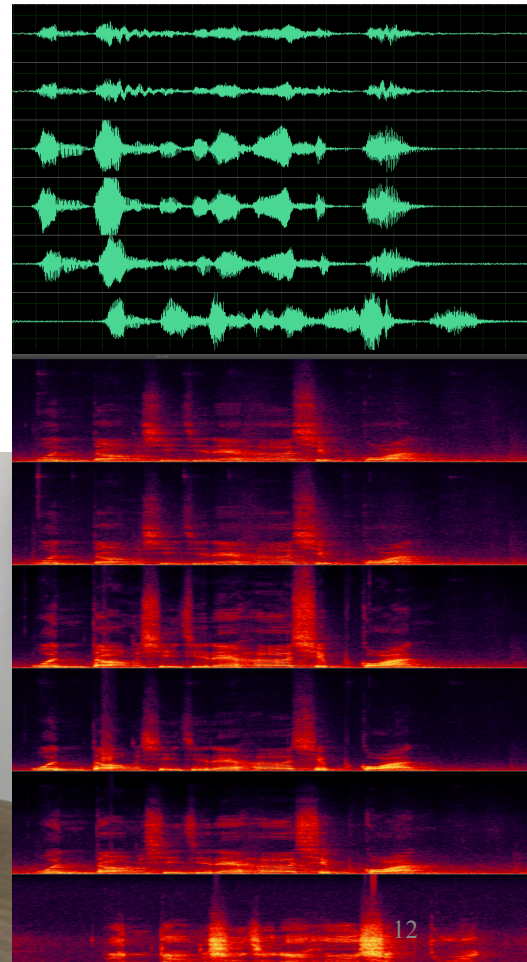
Producer Monitor

Utterances



Test Trial





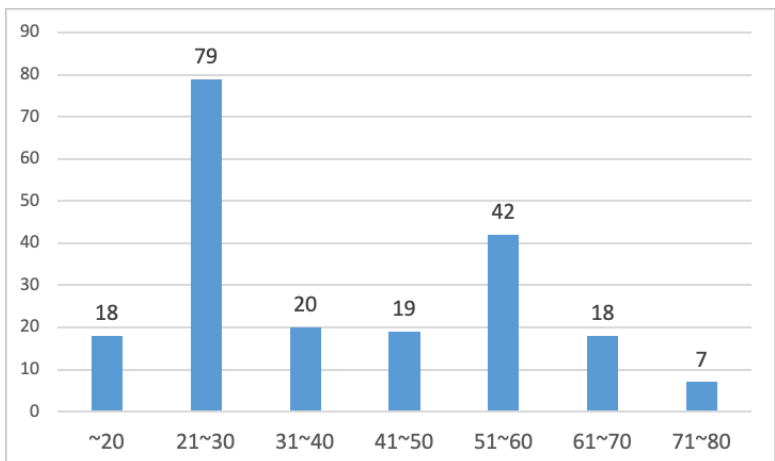
STATUS REPORT (1/3)

- TAT-Vol1~2 available, in total 100 hours * 6 microphones

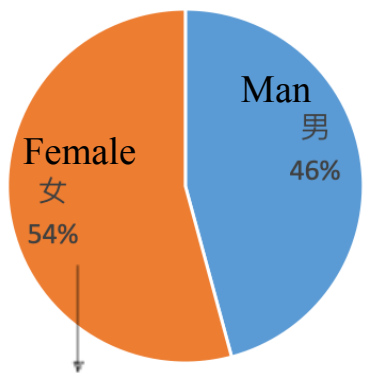
TAT-Vol1				
	Speakers	Sentences	Characters	Hours
Train	80	23,104	271,772	41.76
Evaluation	10	2,943	34,426	5.02
Test	10	2,786	33,394	5.16
TAT-Vol2				
	Speakers	Sentences	Characters	Hours
Train	80	23,216	272,671	42.39
Evaluation	10	2,951	35,951	4.76
Test	10	2,811	31,985	5.27
Total				
	Speakers	Sentences	Characters	Hours
Total	200	57,811	680,199	104.36

STATUS REPORT (2/3)

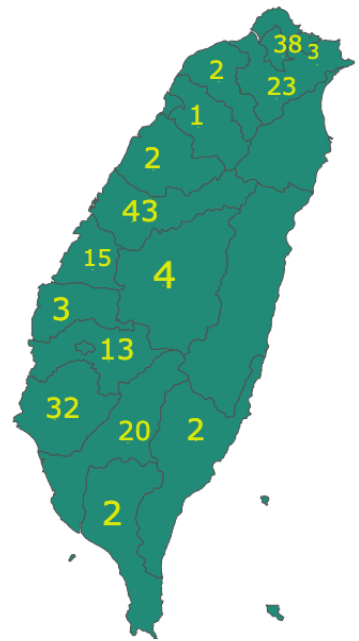
- Speaker Distribution
  - 200 native speakers , each produced about 30 minutes speech



Age Distribution



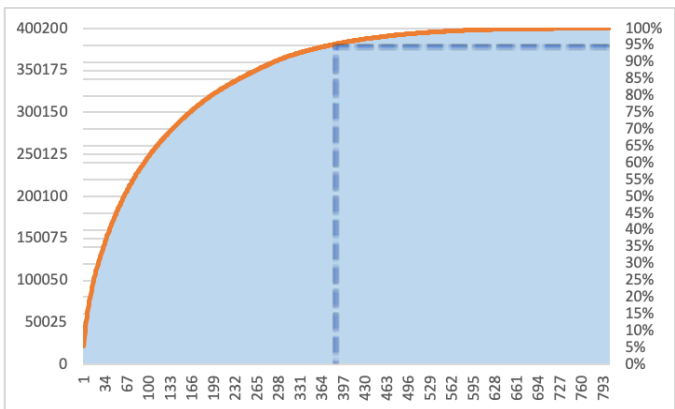
Gender Distribution



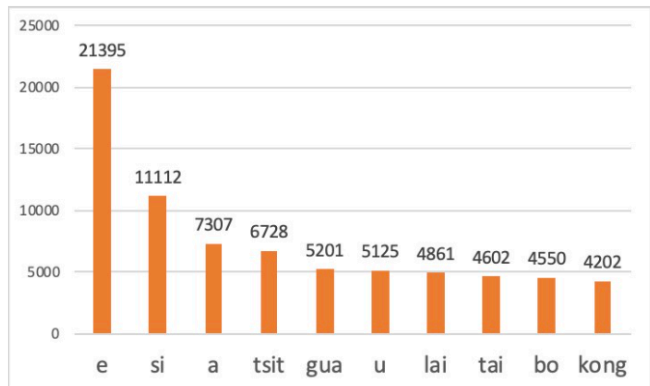
Region Distribution

STATUS REPORT (3/3)

- Accumulated Histogram of Syllables
  - 750 out of 803 syllables had been covered
  - 53 missing syllables should be fixed as soon as possible
  - 388 syllables could cover 95% of the running speech
  - /e/ and /si/ are the highest-frequency syllables



Accumulated histogram of the syllables in TAT



The highest-frequency syllables in TAT corpus

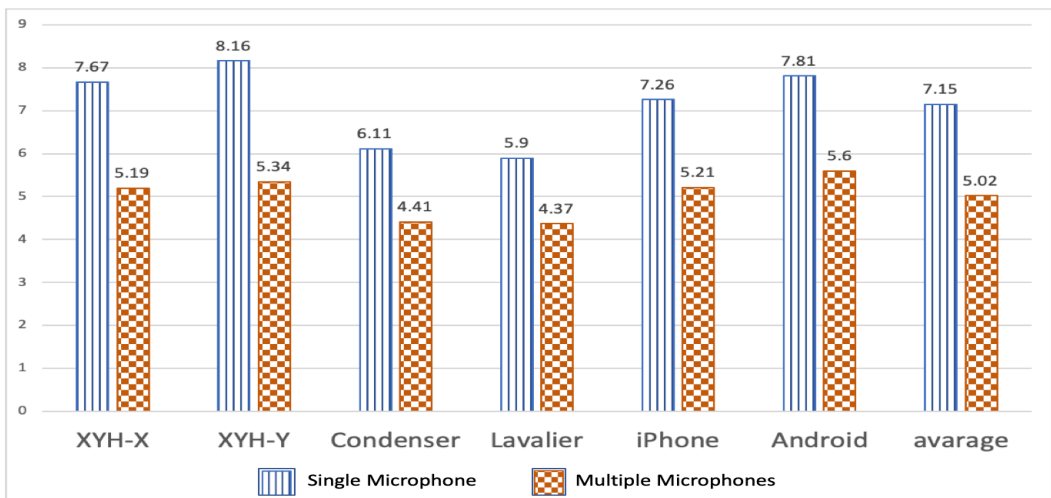
EXPERIMENTS (1/2)

- Baseline
  - Kaldi Recipe available
  - Hybrid HMM/TDNNs (chain model)
- Corpus
  - Single Microphone (lavalier)
  - Train: TAT-Vol1-train (40 hours)
  - Test: TAT-Vol1-eval & TAT-Vol1-test (5 hours each)
- Results
  - 台羅拼音 (Tâi-lô)
  - 台文正字 (Taiwanese Hàn-jī)

Database	TAT-Vol1-eval	TAT-Vol1-test
Tâi-lô(SER)	12.81%	11.24%
Hàn-jī(CER)	21.39%	16.81%

EXPERIMENTS (2/2)

- Corpus
  - All Six different Microphones
  - Train: TAT-Vol1~2-train (480 hours)
  - Test: TAT-Vol1~2-eval & TAT-Vol1~2-test (60 hours each)
- Results
  - 台羅拼音 (Tâi-lô)
- Coupus works well in speech recognition



FORMOSA SPEECH RECOGNITION CHALLENGE 2020 - TAIWANESE ASR

- Homepage
  - https://bit.ly/319EHCy

- Features
  - Free TAT-Vol1 Corpus (50 hours)
  - Free Tâi-lô lexicon
  - Baseline Recipe available: https://bit.ly/3k2vSSu

- Schedule
  - 2020/06/01 --- Registration Open
  - 2020/09/01 --- Pilot-Test (dry-run)
  - **2020/12/01 --- Registration Close (Still open, Welcome to Participate)**
  - 2021/01/01 --- Final-Test

FORMOSA SPEECH RECOGNITION CHALLENGE 2020 - TAIWANESE ASR

- Track 1
  - Traditional Chinese characters (繁體中文字/Chinese)
    - 現在是晚上八點 (It is eight o'clock now)
- Track 2
  - Taiwanese Southern Min Recommended Characters (台文正字/Taiwanese Hàn-jī)
    - 這馬是暗時八點 (It is eight o'clock now)
- Track 3
  - Taiwan Minnanyu Luomazi Pinyin Fang'an (台羅拼音/Tâi-lô)
    - Tsit4-ma2 si7 am3-si5 peh4-tiam2 (It is eight o'clock now)

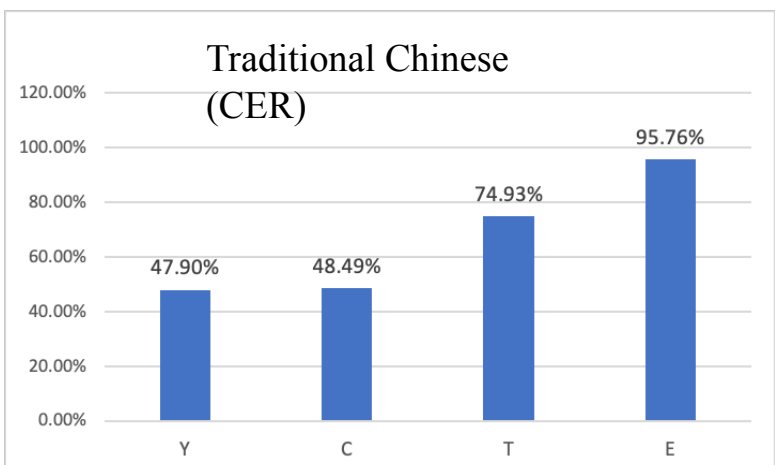
FSR-2020 Pilot-Test Results (1/4)

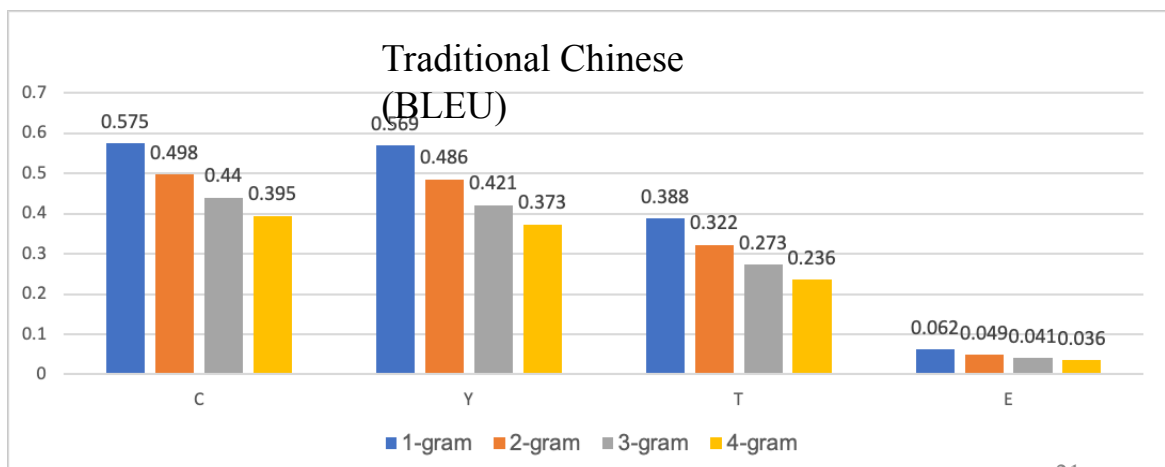
- Corpus
  - Train: TAT-Vol1-Train
  - Test: TAT-Vol1-Eval

- Participants
  - In Total 25 Teams now!
    - Track 1：4 teams
    - Track 2：12 teams
    - Track 3：8 teams

FSR-2020 Pilot-Test Results (2/4)

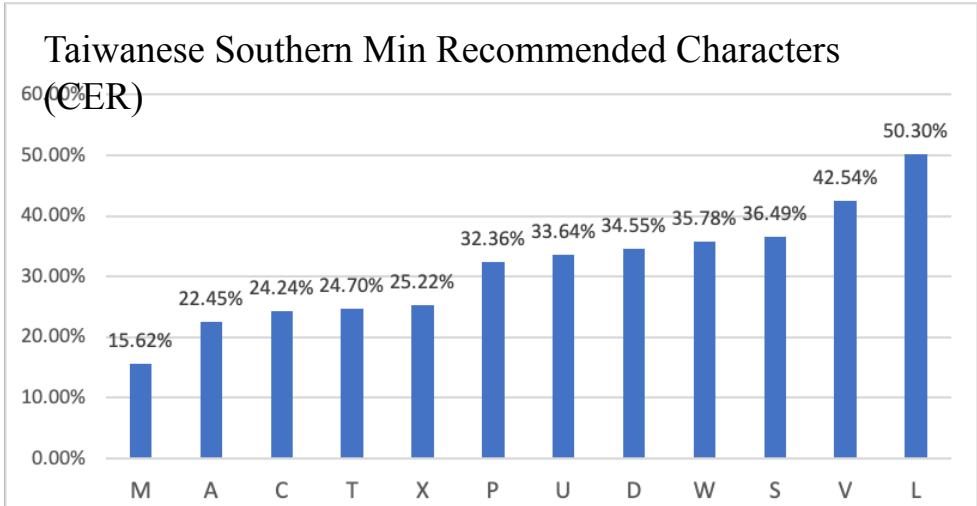
- Track 1: Traditional Chinese characters (繁體中文字/Chinese)
  - Metrics
    - CER in % (Character error rate)
    - BLEU (Bilingual Evaluation Understudy)
  - Very difficult





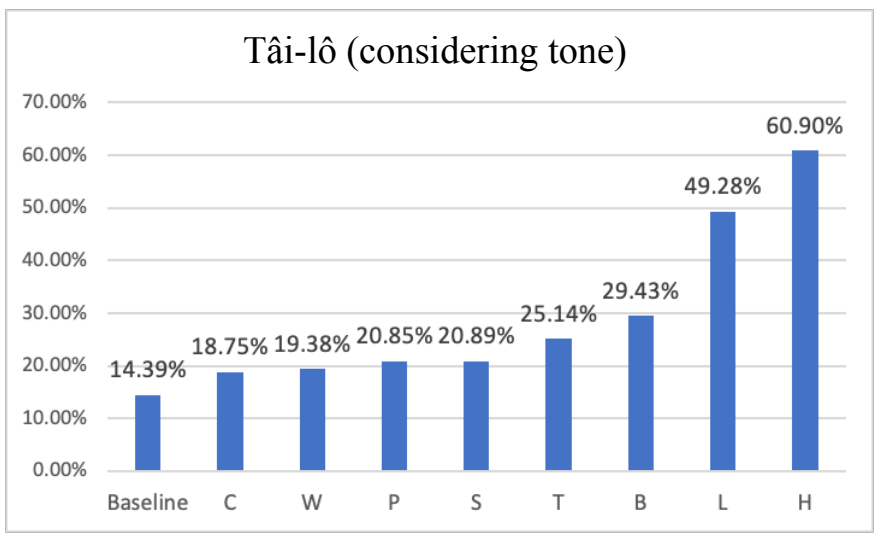
FSR-2020 Pilot-Test Results (3/4)

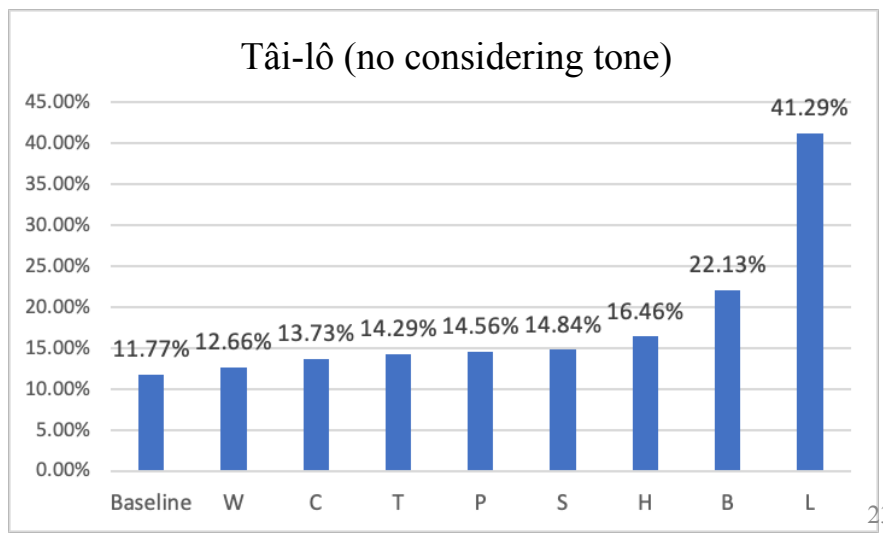
- Track 2: Taiwanese Southern Min Recommended Characters (台文正字/ Taiwanese Hàn-jī)
  - Metrics
    - CER in % (Character error rate)
  - Achieved 15.62% in Taiwanese Southern Min Recommended Characters



FSR-2020 Pilot-Test Results (4/4)

- Track 3: Taiwan Minnanyu Luomazi Pinyin Fang'an (台羅拼音/Tâi-lô)
  - Metrics
    - SER in % (Syllables error rate)
  - Achieved 11.77% in Tâi-lô (no considering tone)





CONCLUSIONS

Taiwanese Across Taiwan

- Taiwanese Across Taiwan (TAT) Project
  - For Taiwanese speech Recognition
  - 2019~2021
  - Native Taiwanese Prompt Sheet/Reading Speech
  - Target: 300 hours * 6 microphones
  - 100 hours (TAT-Vol1~2) released

Lugang Longshan Temple

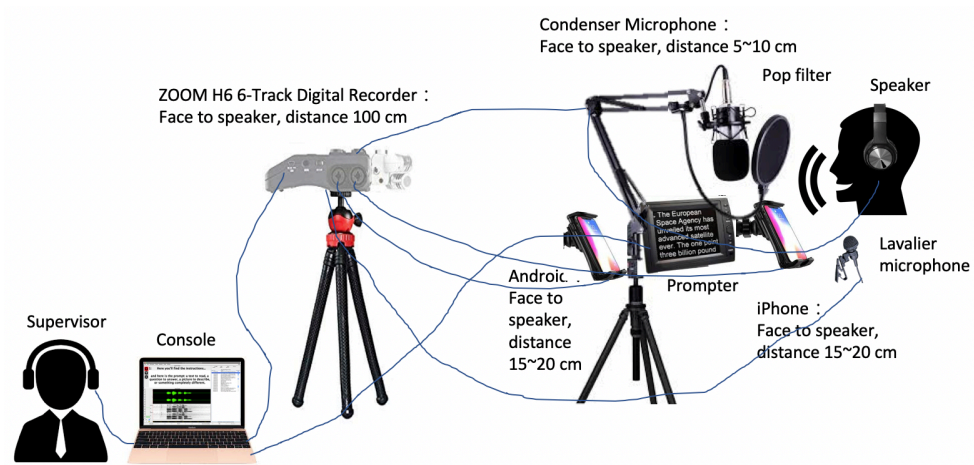
- Welcome to Participate in Formosa Speech Recognition Challenge 2020
  - Free TAT-Vol1, 50 hours
  - Free Tâi-lô lexicon
  - Baseline Recipe Available
  - Registration open till 2020/12/01

Taiwanese Across Taiwan (TAT) Corpus

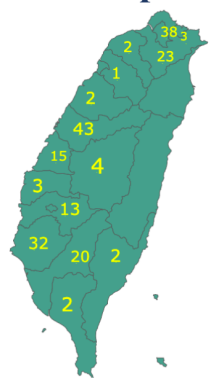
- **PI: Prof. Yuan-Fu Liao, National Taipei University of Technology**
- **Large Scale Taiwanese Across Taiwan (TAT) Corpus**
- 2019~2021
- **Native Taiwanese Articles/Reading Speech**
- Target: 300 hours * 6 microphones
- **100 hours (TAT-Vol1~2) released**
- **License**
- https://bit.ly/319473f
- **Protocol**

Taiwanese Across Taiwan

300 hours * 6 mic.



Dist. of Speakers



Dist. of Ages