Constructing Personalized Text-to-Speech Systems for Augmentative and Alternative Communication by the Speech Labeling and Modeling Toolkit

Chen-Yu Chiang, Yen-Ting Lin, Wu-Hao Li, Wei-Cheng Chen, Jen-Chieh Chiang*, Jia-Jyu Su,*
Cheng-Che Kao, Shu-Lei Lin, Pin-Han Lin, Wen-Yang Chang**

Speech and **M**ultimedia **S**ignal **P**rocessing **Lab**oratory,
Department of Communication Engineering,
National Taipei University, Taiwan

*AcoustInTek Co., Ltd, Taiwan

1. Introduction

The Speech Labeling and Modeling Toolkit version (SLMTK) is designed to facilitate constructing text-to-speech (TTS) systems with a knowledge-rich TTS framework. The SLMTK labels speech corpora with linguistic and prosodic-acoustic information. The labeled information can construct TTS models and be used for speech analysis.

The core SLMTK functions have been applied to constructing personalized TTS systems used in augmentative and alternative communication (AAC) for 20 amyotrophic lateral sclerosis (ALS) patients. Since August 2021, the personalized TTS systems for the 20 ALS patients have been available online. The patients can log in to the web-based TTS service as they would like to use their synthesized speech to communicate.

2. Overview of SLMTK

As shown in Fig. 1, users may upload the speech corpus recorded by a speaker to the SLMTK service. The corpus contains speech saved in *.wav and *.txt. The SLMTK then conducts speech labeling and modeling with the pre-trained prior models.

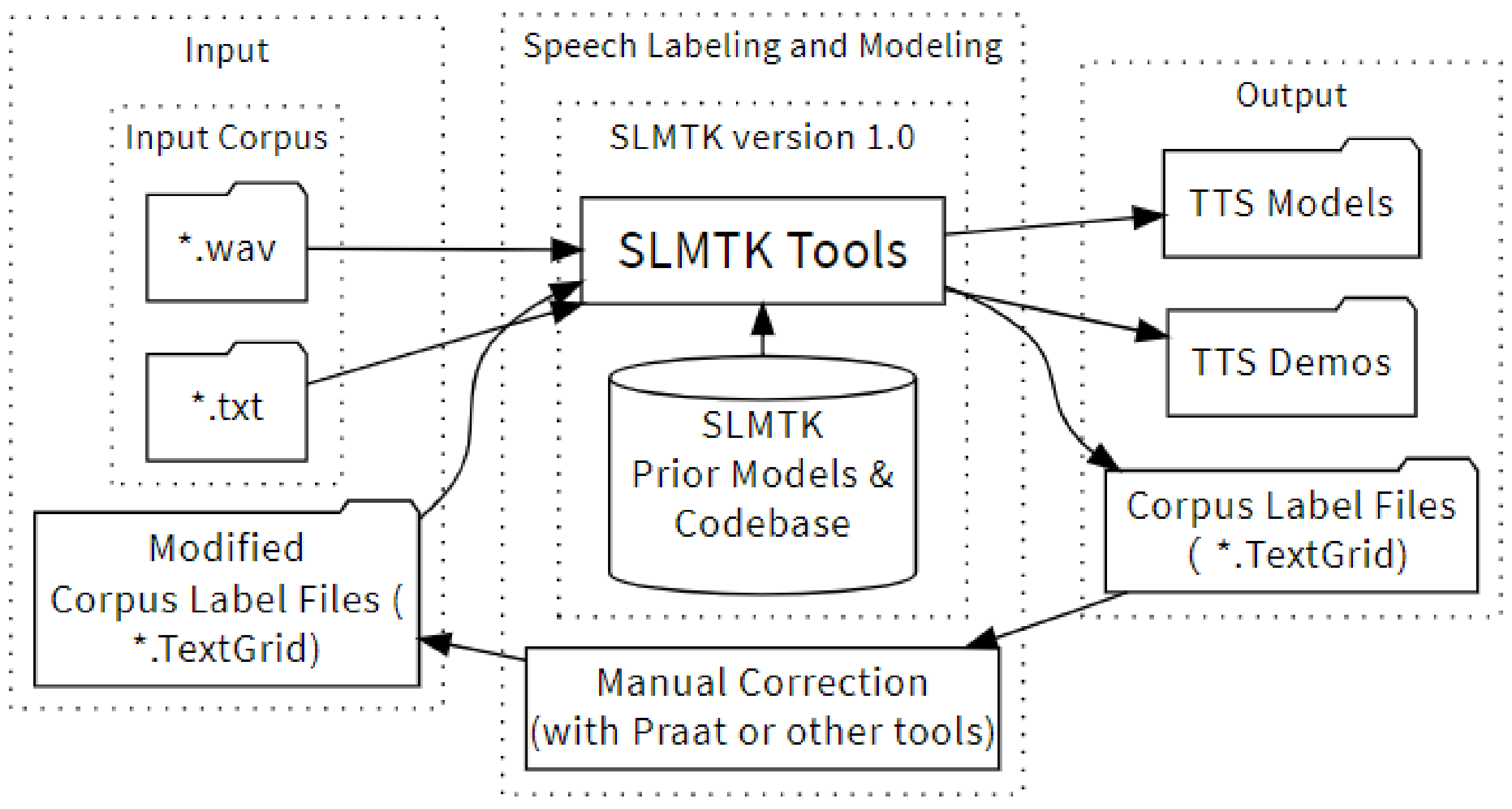

Input
Input Corpus
*.wav
*.txt
Modified Corpus Label Files (*.TextGrid)
Speech Labeling and Modeling
SLMTK version 1.0
SLMTK Tools
SLMTK Prior Models & Codebase
Manual Correction (with Praat or other tools)
Output
TTS Models
TTS Demos
Corpus Label Files (*.TextGrid)

Fig. 1: The Usage of the SLMTK

3. Speech Labeling & Modeling Framework

The SLMTK processes input corpus with the following seven steps:

1. text analysis (ta),
2. acoustic feature extraction (afe),
3. linguistic-speech alignment (lsa),
4. integration of syllable-based linguistic and prosodic-acoustic features (ilp),
5. prosody labeling (and modeling) (plm),
6. construction of prosody generation model (cpg), and
7. construction of acoustic models for speech synthesis with HMM-based and DNN-based synthesis (fine tuning module supported by AcoustInTek Co., Ltd, Taiwan)

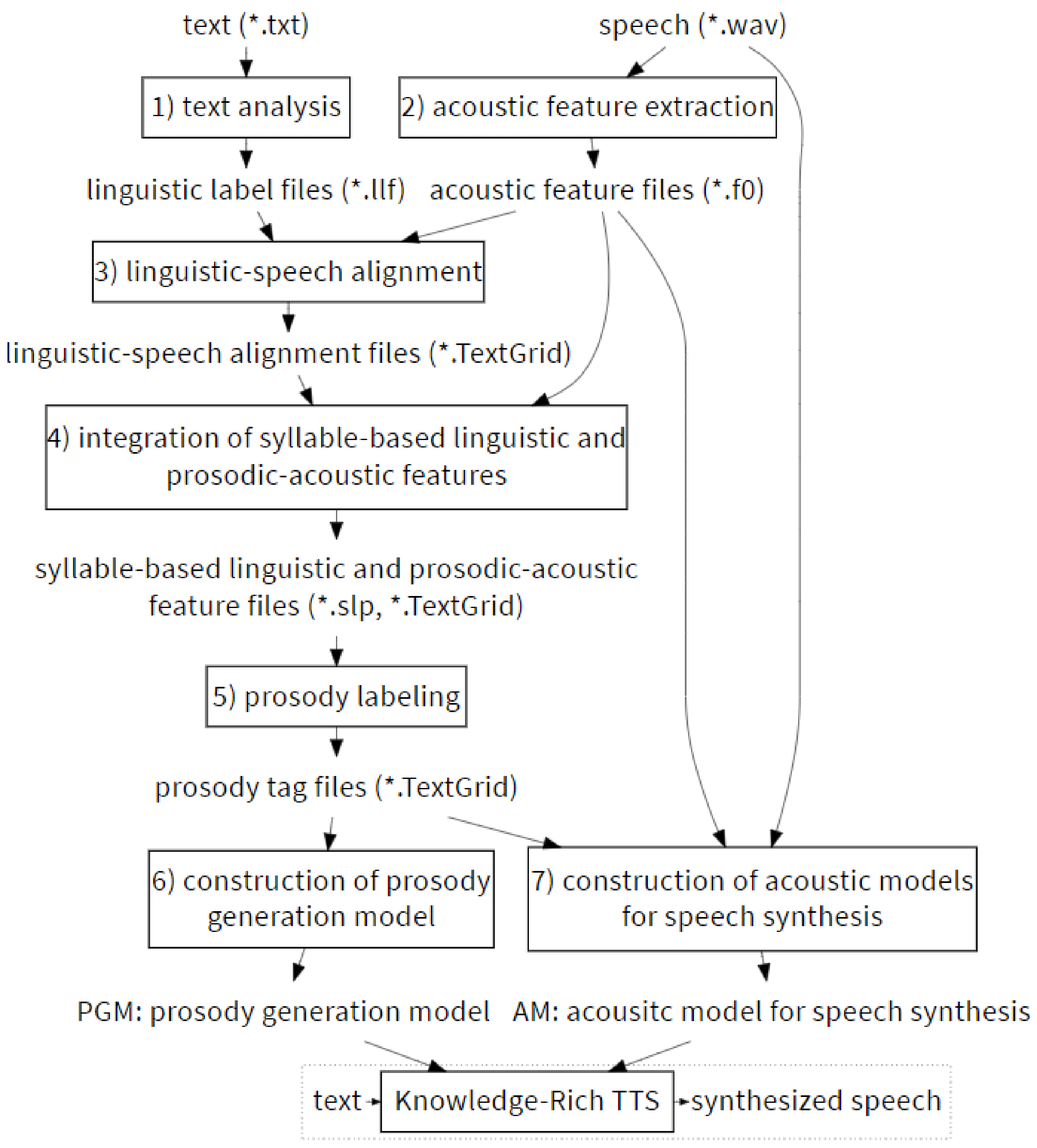

text (*.txt)
speech (*.wav)
1) text analysis
2) acoustic feature extraction
linguistic label files (*.llf)
acoustic feature files (*.f0)
3) linguistic-speech alignment
linguistic-speech alignment files (*.TextGrid)
4) integration of syllable-based linguistic and prosodic-acoustic features
syllable-based linguistic and prosodic-acoustic feature files (*.slp, *.TextGrid)
5) prosody labeling
prosody tag files (*.TextGrid)
6) construction of prosody generation model
7) construction of acoustic models for speech synthesis
PGM: prosody generation model
AM: acousitc model for speech synthesis
text
Knowledge-Rich TTS
synthesized speech

Fig. 2: The Framework of the SLMTK

4. Knowledge-Rich Text-to-Speech Framework

- The TA extracts linguistic information from the input text. The linguistic information contains lexical, syntactic, and partly semantic features.
- The PG produces prosodic information by the prosody generation model (PGM) given with the linguistic features extracted by the TA. The prosodic information considered in the SLMTK contains prosodic breaks, prosodic states, and syllable-based prosodic-acoustic features.
- The SG generates speech parameters that control the fundamental frequency and spectral envelope by the acoustic model for speech synthesis.
- Last, the WG produces speech signals by some vocoding models given the speech parameters.

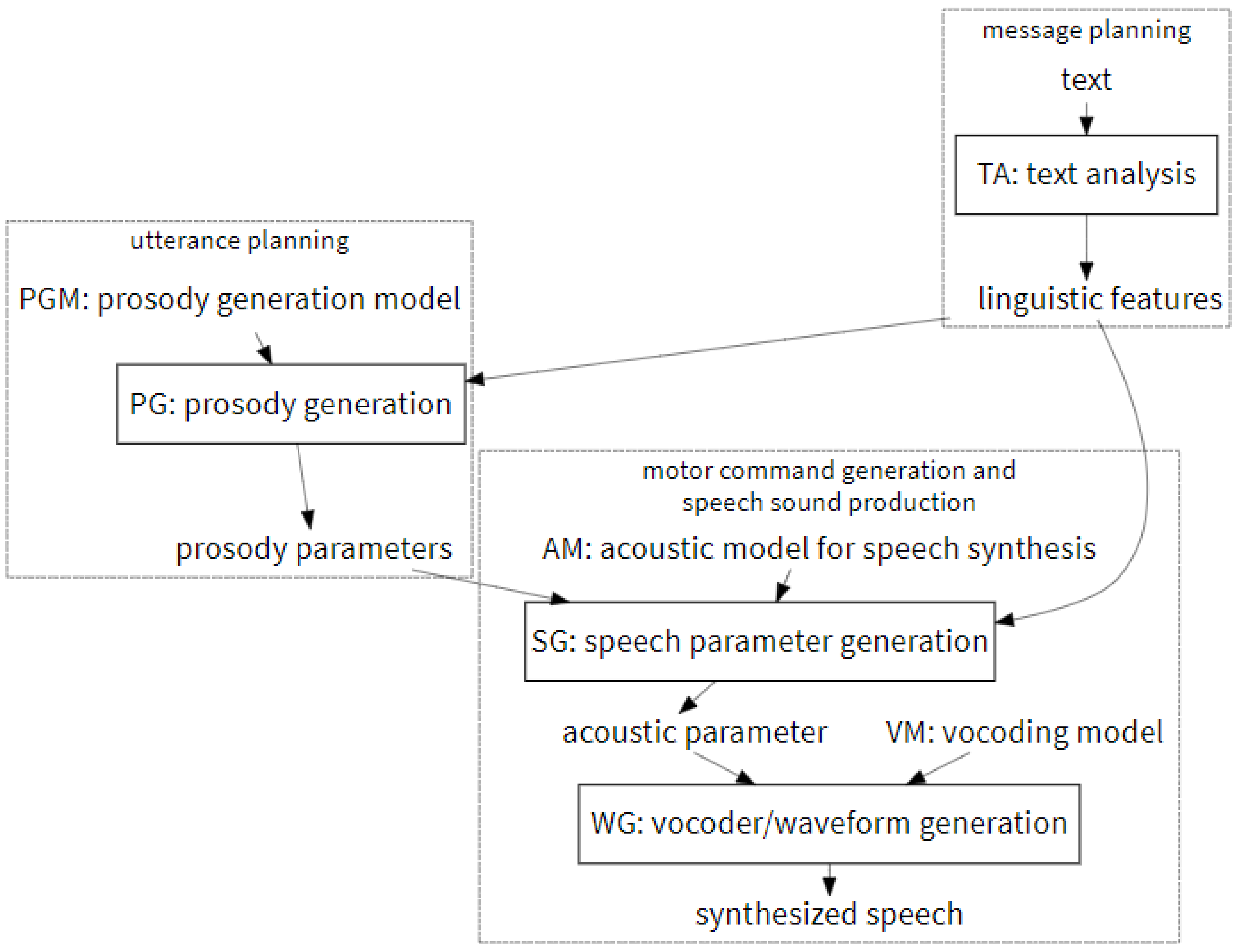
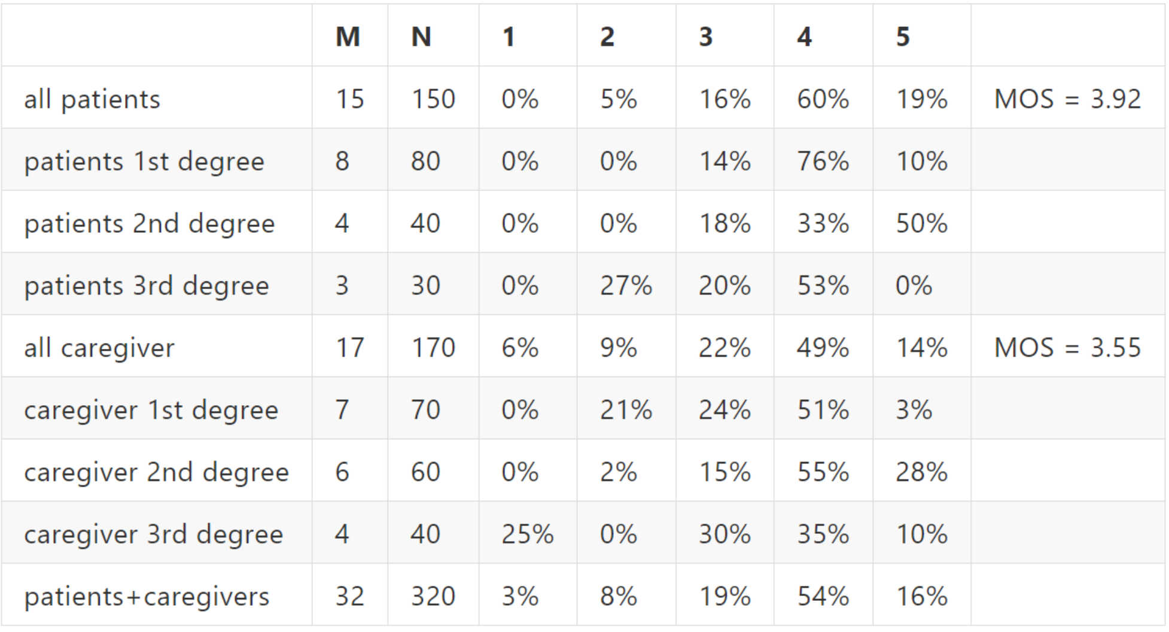

message planning
text
TA: text analysis
linguistic features
utterance planning
PGM: prosody generation model
PG: prosody generation
prosody parameters
motor command generation and speech sound production
AM: acoustic model for speech synthesis
SG: speech parameter generation
acoustic parameter
VM: vocoding model
WG: vocoder/waveform generation
synthesized speech

Fig. 3: Knowledge-Rich Text-to-Speech Framework

5. Evaluation

ASL patients and their caregivers were asked to listen to the ten synthesized speech utterances which are generated by the personalized TTS systems given with the text materials provided by the patients.

The rating is the 5-point mean opinion score: **1)** strongly disagree (synthesized speech sounded like others) , **2)** disagree, **3)** neutral, **4)** agree, and 5) strongly agree （synthesized speech sounded like the patient ）



	M	N	1	2	3	4	5	
all patients	15	150	0%	5%	16%	60%	19%	MOS = 3.92
patients 1st degree	8	80	0%	0%	14%	76%	10%	
patients 2nd degree	4	40	0%	0%	18%	33%	50%	
patients 3rd degree	3	30	0%	27%	20%	53%	0%	
all caregiver	17	170	6%	9%	22%	49%	14%	MOS = 3.55
caregiver 1st degree	7	70	0%	21%	24%	51%	3%	
caregiver 2nd degree	6	60	0%	2%	15%	55%	28%	
caregiver 3rd degree	4	40	25%	0%	30%	35%	10%	
patients+caregivers	32	320	3%	8%	19%	54%	16%	

M represents the number subjects; N represents the numebr of test synthesized utterances; the numbers 1-5 are 5 point MOS in speaker similarity.

6. Conclusion and Future Works

- It is encouraged to see that the synthesized speech made by the SLMTK could achieve fair MOS. The high scores of "Willingness to Use" and "Satisfaction" surveys were also reported by the enrolled patients and caregivers (details not shown here).
- Integrating with VoiceBank and providing a fully-online service that delivers personalized text-to-speech systems to users in 24hrs after recording is worthwhile doing.

Acknowledgments

The authors also would like to thank the Taiwan Motor Neuron Disease Association for coordinating speech recordings for the enrolled ALS patients and funding the development of VoiceBank.

References

1. C. -Y. Chiang et al., "The Speech Labeling and Modeling Toolkit (SLMTK) Version 1.0," 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Hanoi, Vietnam, 2022, pp. 1-5, doi: 10.1109/O-COCOSDA202257103.2022.9997860.
2. Y. -T. Lin and C. -Y. Chiang, "EGAN: A Neural Excitation Generation Model Based on Generative Adversarial Networks with Harmonics and Noise Input," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096801.