INSTRUCTIONS for COMPILATION

Please fill in the User Agreement as required below:

1) Complete the user legal name and address. The user named must be a legal institution, or a department or section of a named legal institution, not an individual nor a project.

2) Add the signature, name of signee, title of signee and date of signature. The person signing this Agreement should be the person authorized by the institution for such signatures.

3) Each page of the agreement should also be initialed (the person signing the agreement should put his/her initials on each and all pages).

4) In Exhibit, you will find all necessary details about the Dataset.

Please send an original signed copy (Scanned copy) by email to:
olr19@cslt.org

On receipt of email, we will provide you with the information to download the Dataset.

Thank you.

****

Please find an example of how to fill in the form as following:

User Agreement

THIS USER AGREEMENT ("Agreement"), dated as of Jun 21, 2017, by and between Center for Speech and Language Technologies, Tsinghua University ("CSLT"), with office at Room 1-303, FIT Building, Tsinghua University, Haidian District, Beijing China and Speechocean Limited. ("Speechocean") with offices located at D801, Universal Center, 28# Chengfu Road, Haidian District, Beijing China in consideration of the mutual covenants contained herein, the parties agree as follows: ...

OLR Multilingual Database

User Agreement

THIS USER AGREEMENT ("Agreement"), dated as of _____, by and between _____ ("_____"), with office at _____ and Speechocean Limited ("Speechocean") with offices located at D801, Universal Center, 28# Chengfu Road, Haidian District, Beijing China in consideration of the mutual covenants contained herein, the parties agree as follows:

1. OLR Multilingual Database (the part provided by Speechocean), including Training Set (Development Set) & Testing Set, and other related documents, is subject to this AGREEMENT, with details listed in the exhibits, hereafter referred as "Database".
2. "Database" can only be used by applicants in the events of AP19-OLR challenge, hereafter referred as the "Event". For any other academic purpose, users shall notice Speechocean officially by Email and be permitted by Speechocean.
3. Applicant is limited to universities, colleges, labs, schools and other academic entities, or the research institute of commercial companies, and should be dedicating in the field of speech and language technologies, including but not limited to language phonetic and phonological analysis, speech recognition, speaker recognition, speech synthesis, language understanding and etc.
4. Applicant shall provide the linkage of the belonged organization, or any of the related papers published before in the previously mentioned research fields for qualification confirmation.
5. Applicant shall have no rights to publish, retransmit, disclose, display, copy, reproduce or redistribute the "Database" or part of the "Database" to any other entities, website or individuals, under any condition. The "Database" shall not be used for any commercial purpose.
6. Applicant shall give appropriate references to the "Database". No matter the whole "Database" or part of the "Database" is used, in any of the scholarly publications whenever and wherever they are mentioned. The following citation format is suggested:

- "KingLine Data Center, AP16-OL7 Multilingual Database, Speechocean Ltd. (www.speechocean.com), 2016."
- "Zhiyuan Tang, Dong Wang, Liming Song: AP19-OLR Challenge: Three Tasks and Their Baselines, APSIPA ASC 2019."

7. Applicant, who downloaded the “Database” but could not submit papers or challenge result to the “Event”, by any possible reasons, shall notice Speechocean officially by email.

8. All entities or individuals who do not participant the “Event” but want to use the “Database” will subject to data usage regulation of the Kingline Data Center, a data sharing and exchange data platform operated by Speechocean. Details can be found on http://kingline.speechocean.com/article.php?id=18.

9. Speechocean can be reached by any of the following manners:

   Email: contact@speechocean.com or songliming@speechocean.com

   Tel: 86-10-62660053

AUTHORISED SIGNATURES:

Title of Signee in User’s Organization:

On Behalf of:

OLR Multilingual Database

User Agreement

THIS USER AGREEMENT ("Agreement"), dated as of _____, by and between _____ ("_____"), with office at _____ and Center for Speech and Language Technologies, Tsinghua University ("CSLT") with offices located at Room 1-303, FIT Building, Tsinghua University, Haidian District, Beijing China in consideration of the mutual covenants contained herein, the parties agree as follows:

1. OLR Multilingual Database (the part provided by CSLT, M2ASR project), including Training Set (Development Set) & Testing Set, and other related documents, is subject to this AGREEMENT, with details listed in the exhibits, hereafter referred as "Database".

2. "Database" can only be used by applicants in the events of AP19-OLR challenge (referred as "Event" hereafter). This Database can be used for other research purposes, but commercial usage is not allowed.

3. Applicant is limited to universities, colleges, labs, schools and other academic entities, and should be dedicating in the field of speech and language technologies, including but not limited to language phonetic and phonological analysis, speech recognition, speaker recognition, speech synthesis, language understanding and etc.

4. Applicant shall provide the linkage of the belonged organization, or any of the related papers published before in the previously mentioned research fields for qualification confirmation.

5. Applicant shall have no rights to publish, retransmit, disclose, display, copy, reproduce or redistribute the "Database" or part of the "Database" to any other entities, website or individuals, under any condition.

6. Applicant shall give appropriate references to the "Database". No matter the whole "Database" or part of the "Database" is used, in any of the scholarly publications whenever and wherever they are mentioned. The following citation format is suggested:

- "Zhiyuan Tang, Dong Wang, Liming Song: AP19-OLR Challenge: Three Tasks and Their Baselines, APSIPA ASC 2019."

7. Applicant, who downloaded the "Database" but could not submit the challenge result to the "Event", by any possible reasons, shall notice the data provider.

8. All institutes or individuals who do not participant the “Event” but want to use the “Database” will subject to data usage regulation of the M2ASR project. Please contact the data provider:

CSLT@Tsinghua University, Northwest Minzu University, Xinjiang University, supported by M2ASR NSFC project (http://m2asr.cslt.org)

Email: wangdong99@mails.tsinghua.edu.cn

AUTHORISED SIGNATURES:

Title of Signee in User’s Organization:

On Behalf of:

EXHIBIT A:

* The following datasets are provided by SpeechOcean.

OL7 Multilingual Database		Training/Development Set (AP16-OL7)			Testing Set (AP17-OL7-test)			Testing Set (AP18-OL7-test)		
Language	Channel	No. of Speaker	Total Utterances	Recording Hours	Nos. of Speaker	Total Utterances	Recording Hours	Nos. of Speaker	Total Utterances	Recording Hours
Cantonese in China Mainland and Hongkong	Mobile	24	7676	10.19	8	2556	3.23	8	2,558	3.7
Mandarin in China	Mobile	24	7198	10.29	8	2400	3.37	8	2,400	3.1
Indonesian in Indonesia	Mobile	24	7671	10.64	8	2557	3.98	8	2,557	3.6
Japanese in Japan	Mobile	24	7659	7.98	8	2548	2.7	8	2,544	2.6
Russian in Russia	Mobile	24	7183	12.91	8	1796	3.27	8	2,394	2.3
Korean in Korea	Mobile	24	7195	7.91	8	2398	2.96	8	2,399	2.6
Vietnamese in Vietnam	Mobile	24	7197	11.4	8	2396	3.92	8	2,400	3.7
8 interference languages	Mobile	--						20*8	1,264	2.43

Note:

* Male and Female are balanced;
* There might be a minor discrepancy with the numbers of total utterances.

EXHIBIT B:

* The following datasets are provided by Tsinghua University, Northwest Minzu University and Xinjiang University, under the M2ASR project supported by NSFC.

OL3 Multilingual Database		**Training/Development Set** (AP17-OL3-train / AP17-OL3-dev)			**Testing Set** (AP17-OL3-test)			**Testing Set** (AP18-OL3-test)		
Language	Channel	No. of Speaker	Total Utterances	Recording Hours	Nos. of Speaker	Total Utterances	Recording Hours	Nos. of Speaker	Total Utterances	Recording Hours
Tibetan	Mobile	34	11100	12	34	1800	2	34	1800	2
Uyghur	Mobile	353	5800	14	353	1800	4	353	1800	4
Kazakh	Mobile	86	4200	10	86	1800	4	86	1800	4

Note:

* Male and Female are balanced;
* There might be a minor discrepancy with the numbers of total utterances.

EXHIBIT C:

* The following AP19-olr-dev is provided by both M2ASR project and SpeechOcean (the 3 dialects). AP19-olr-short is provided by both SpeechOcean and M2ASR project (Tibetan, Uyghur and Kazakh). AP19-olr-channel and AP19-olr-channel are provided by M2ASR project and SpeechOcean respectively.

AP19-OLR dev/test	**Development Set** (AP19-olr-dev)		**Testing Set** (AP19-olr-short)		**Testing Set** (AP19-olr-channel)		**Testing Set** (AP19-olr-zero)	
Language	Total Utterances	Recording Hours	Total Utterances	Recording Hours	Total Utterances	Recording Hours	Total Utterances	Recording Hours
Tibetan	500	3	1800	5	1800	10	-	-
Uyghur	500		1800		1800		-	-
Kazakh	-		1800		-		-	-
Mandarin in China	500		1800		1800		-	-
Japanese in Japan	500		1800		1800		-	-
Russian in Russia	500		1800		1800		-	-
Vietnamese in Vietnam	500		1800		1800		-	-
Cantonese in China Mainland and Hongkong	-	-	1800		-	-	-	-
Korean in Korea	-	-	1800		-	-	-	-
Indonesian in Indonesia	-	-	1800		-	-	-	-
Shanghai/Sichuan/Minnan Dialect	505*3	2	-	-	-	-	-	-
Catalan/Greek/Telugu	-	-	-	-	-	-	1810*3	8

Note:
* Male and Female are balanced;
* There might be a minor discrepancy with the numbers of total utterances.