8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)
20-21 September 2019, Graz, Austria

Does Speaking Training Application with Speech Recognition Motivate Junior High School Students in Actual Classroom? – A Case Study

Satoshi Kobashikawa[1], *Atushi Odakura*[2], *Takao Nakamura*[1], *Takeshi Mori*[1]
Kimitaka Endo[3], *Takafumi Moriya*[1], *Ryo Masumura*[1], *Yushi Aono*[1] *and Nobuaki Minematsu*[4]

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]Head of Research and Development Planning, NTT Corporation, Japan
[3]AI and Robotics Business Headquarters, NTT Advanced Technology Corporation, Japan
[4]Graduate School of Engineering, The University of Tokyo, Japan

{satoshi.kobashikawa.he,atsushi.odakura.pa,takao.nakamura.vp}@hco.ntt.co.jp,
takeshi.mori.zb@hco.ntt.co.jp, kimitaka.endo@ntt-at.co.jp,
{takafumi.moriya.nd,ryo.masumura.ba,yushi.aono.dy}@hco.ntt.co.jp,
mine@gavo.t.u-tokyo.ac.jp

Abstract

This paper investigates the effectiveness of a speech training application with question answering problems based on speech recognition which is robust to noise in a classroom. In actual classrooms, since a lot of students speak at the same time, input speech waveforms include speech noise from neighboring students. To the best of our knowledge, no speech training study has examined speech recognition with the focus on actual noise corrupted speech input in classrooms. Since existing mobile applications assumed solitary voice input, they failed to evaluate the input as corrupted by speech noise. To maintain students' motivation even in noisy environments, our application ignores the insertion errors around the user's intended sentence. To adapt to junior high school students' speech and the background speech noise, we introduce unsupervised adaption with matched sentences by comparing the speech recognition results and target sentences candidates. We also improve the user interface by reflecting the feedback from teachers and students. The results of a two month trial with over 140 students in a public junior high school show that our speech recognizer improves accuracy and our application achieves a positive user experience.

Index Terms: computer assisted language learning, speech recognition, L2 learning

1. Introduction

Due to recent globalization trends, it is important to improve communication skill in English as a common language. In particular, speaking skill is required for effective communication. To improve students' speaking skill, several research institutes are supporting Spoken CALL Shared Task [1] to collect English speech from Swiss German teens. In our country, English is the dominant second language (L2) for Japanese students. However since Japanese and English are so dissimilar not only in pronunciation but also grammar, most Japanese students fail to speak English fluently; Japanese students often speak in Japanese-accented English even if they have good writing skill. To improve Japanese student's English ability, four English skills (reading, listening, writing, and 'speaking') will be introduced to the university entrance examination from 2020. In addition, the Tokyo metropolitan board of education also announced that an English speech test will be added to the public high school entrance examination from 2022. We should rapidly improve the English speaking skills of junior high school students.

There are a lot of training applications and evaluation services [2] intended to improve English language skills; examples include ELSA [3], Moby.Read [4], CARAMILLA [5], TipTopTalk! [6], Duolingo [7] Reading-while-Listening [8], RALL (Robot Assisted Language Learning) [9], Versant [10][11] and so on. Most applications and services were designed assuming personal use cases. Since they do not take the noise from surrounding speakers into account, they fail to work properly due to the competing speech sources in actual class rooms. As described by [12], the close-talking microphone is useful for reducing the surrounding noise, however it is difficult to eliminate the noise completely. Furthermore, since students' seats are very densely packed in actual class rooms, neighboring student's speech is loud. There are several noise reduction technologies, but excessive noise reduction distorts the true speech and degrades the speech recognition accuracy. Luan *et al.* focus on the surrounding speech noise to evaluate pronunciation with noise reduction [13]; however, their evaluation considered just GOP (Goodness Of Pronunciation) [14] and ignored speech recognition performance. Shidiev *et al.* evaluate speech recognition in classrooms, but they use speech recognition only for the teacher's speech [15]. Luo *et al.* use a smartphone to encourage English communication, but they don't mention speech recognition [16]. If school students are to be well supported by an English speaking training application with speech recognition, the problem of the noisy environments common in actual classrooms must be addressed.

This paper investigates the effectiveness of a noise robust English speech training application in an actual classroom of a public junior high school in Japan. Our speech recognizer uses both non-native and native English speech to recognize Japanese-English speech as [17][18]. To reduce the negative influence of background speech noise in classrooms, our application ignores the insertion errors before/after intended sentences in the automatic evaluation process. To improve accuracy further, we utilize unsupervised adaptation to handle background speech noise and junior high school student's speech. To improve the motivation of students and teachers, we revised several of the user interfaces based on user feedback. In a 2 month trial with 148 students, most students (75 %) enjoyed the English training application in the classroom. In particular, students who were not good at speaking English highly rated the application and noted that it reduced embarrassment and nervousness with regard to speaking English.

10.21437/SLaTE.2019-23

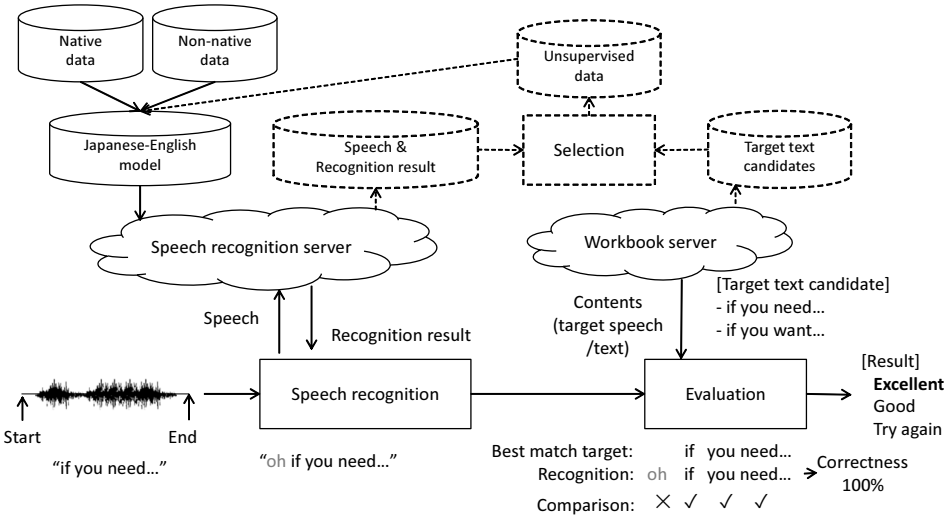

Figure 1: *Overview of proposed system*

The rest of this paper is organized as follows; the proposed approach is described in Section 2. Section 3 evaluates the effectiveness of the proposed approach. Our conclusion is drawn in Section 4.

2. Proposed approach

To adequately investigate the effectiveness of speech recognition for speaking training, we developed the application in two steps and conducted a two month trial. As the first step, we introduced speech recognition to a workbook application. We collected the voices of students and teachers during the first month, and used it as user feedback to improve the system in the 2nd step; the improved system was tested over a month in the last half of the trial. Fig. 1 shows an overview of the proposed system.

(a) *Initial system* (b) *Improved system*

Figure 2: *Application screen image of initial and improved system*

2.1. 1st step: Initial system

Smart device applications are more attractive than desktop/laptop applications [19]. Accordingly, we modified a workbook application "KNOUN" [20] for smart devices by introducing speech recognition. Fig.2a shows a screen image of the initial system.

Most Computer Assisted Pronunciation Training (CAPT) systems (e.g. [12][13][14]) give the correct sentence beforehand. Instead, since our system doesn't give the sentence as question answering, fill-in-the-blank, and word order problems, user's input waveform often has a hesitation speech while thinking. Most of smart device applications adopt VAD (Voice Activity Detection) [21] to detect end of speech. To reduce unintended end of speech detection automatically, our system use push-to-start and push-to-stop recognition approach which yields explicit end of speech tags. We don't use noise reduction to prevent from speech distortion, since the surrounding non-stationary speech noise is difficult to estimate.

In actual classrooms, the surrounding speech noise from neighboring students is significant. To reduce the influence of this speech noise, the uttered speech waveform is processed by ignoring the recognition errors created by the noise. Our system uses character correctness as it offers simple evaluation. However, as teachers demand simple explanations, the trial uses rule-based processing with regular expressions derived from content (teaching materials) providers. Both correctness and regular expression evaluations can ignore the unexpected inserted words generated from the surrounding irregular speech noise.

2.2. 2nd step: Improved system

To improve the user experience, we revised the application by reflecting the results of the first half trial, error analysis and user feedback. Fig.2b shows the application screen image of the improved system.

As predicted, the surrounding extraneous speech noise causes many recognition errors. To reduce these errors, we introduce unsupervised adaptation for the student's speech and the surrounding noise. After comparing speech recognition results and target candidate texts, we select the best match text from among the candidate texts. We use the selected text as unsupervised adaptation data, indicated by the dotted lines in Fig1, without transcription. By selecting the texts with high matching rate, unsupervised adaptation data approaches the accuracy of transcription data, but data sizes are smaller. By suppressing the insertion errors in calculating the matching rate, we can improve robustness to the surrounding noise; in terms of the selection criteria, correctness yields better performance than accuracy.

Table 1: *Problem, error analysis and action*

Problem	Error analysis	Action
1. Misrecognition due to surrounding noise	Main speech buried in noise	Add volume meter / adaptation
2. Low accuracy at the beginning	It requires some activation time to record	Put offset period information
3. Too many questions to teacher	Student doesn't know the correct speech	Add correct/record speech
4. Many requests to revise the evaluation	Short answers create errors	Add reference evaluations
5. Existence of skill gap	High/low skill students feel bored/difficult	Separate contents into two levels

We also revised the application user interface as follows. 1) We added a volume meter as some students' voices were buried in surrounding noise. 2) We set offset period information on the input form. This was needed because the speech at the beginning of the input sequences yielded low accuracy due to missing head of speech and activation noise from device. 3) We also add a function to replay target native speech and recorded student speech. When the student's pronunciation is really poor, the speech recognizer can't recognize the student' intended sentence. The result is that the student asks the teacher too many questions in the classroom. Since the replay function helps the students to learn by themselves, it can reduce the teacher's burden.

We prepared the original contents to suit the school's curriculum, and also revised the contents to reflect the user feedback as follows. 4) We added candidate alternatives as correct answers. This was needed since students and teachers wanted to select shorter expression. 5) We also separated the contents into two parts; easy and difficult. There are students with high and low skill in public junior school because no examination need be passed to enter the school. To motivate both types of students, we prepared 2-level contents.

Table 1 summarizes the problems, the error analysis results, and actions. We revised both user interfaces (1,2 and 3) as in Fig 2 and contents (3,4).

3. Evaluation

3.1. Evaluation of speech recognition

3.1.1. Settings of speech recognition

To investigate the effectiveness of our proposal, we collected Japanese English speech samples from 148 students via smart devices (iPad) and headset microphones in actual classrooms, see Fig.3. The collected samples totalled over 90 hours for the 2 month trial; each student used this application for 66 minutes and speaks 36 minutes in three classes. We selected 1.5 hours of Japanese English speech data to evaluate speech recognition performance; the speech data was transcribed by labellers who knew both English and Japanese-accented English. From the source data, the remaining Japanese English speech was used as the unsupervised training data for the acoustic model. In the language learning task, since we can forecast the target text by selecting the answer candidate sentences, we can use the best match sentence for training data instead of human transcriptions.

All speech data was recorded with 16 kHz sampling rate and 16 bit resolution in this experiment. The frame width and frame shift were 20 msec and 10 msec, respectively. VAD [21] and noise reduction were not used as they would have degraded the recognition rate degrade as mentioned in Section 2.2 in the preliminary experiments.

The acoustic model was a convolutional neural network with the network-in-network architecture (NiN-CNN) acoustic

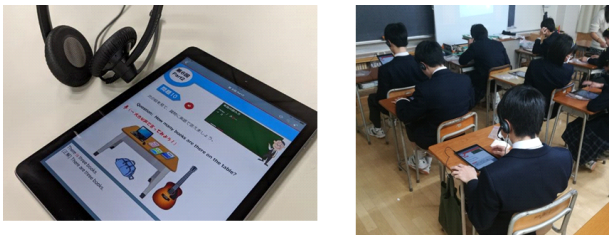

Figure 3: *Application with iPad and headset microphone and picture in class*

model [22]. Its structure is detailed in Table 2; "conv" is a convolutional layer, "pool" is a max pooling layer, and "fc" is a fully-connected layer. "conv{1,2}b" corresponds to the NiN architecture [22]. The sigmoid activation function was used for all hidden layers. The acoustic feature consisted of 40 log mel filterbank coefficients appended with delta and acceleration coefficients. Each static and dynamic component was spliced within 11 frames and treated as a feature map, i.e. 3 feature maps of 40x11 size were input to the acoustic model. The trained parameters were quantized from 32-bit floating-point precision into 8-bit fixed-point, and compressed by singular value decomposition (SVD) with fine-tuning for fast-decoding on the cloud server [23]. SVD was applied to all linear layers, i.e. "fc*" and "softmax" in Table2, which were decomposed into two matrices with 512 dimensional units.

We used a 3-gram language model with vocabulary size of 130,127 in all conditions. Decoding was performed by the WFST-based decoder VoiceRex [24][25].

Table 2: *Structure of the NiN-CNN acoustic model*

Layer	Filter size	Input size	#Feature maps
		40x11	3
conv1a	5x11	36x1	180
conv1b	1x1	36x1	180
pool1	2x1	18x1	180
conv2a	5x11	14x1	180
conv2b	1x1	14x1	180
pool2	2x1	7x1	180
fc1		2048	
fc2		2048	
softmax		2601	

3.1.2. Results of speech recognition

Table 3 shows recognition rate (word correctness and accuracy) for different amounts of training data. Columns of "Native", "Non-native", "Unsupervised", and "Transcription" show size of training data.

As in Table 3, the difference between correctness and accuracy is large (e.g. 11.05 points in ID=2. no-native model). Since its difference comes from the insertion errors created by

Table 3: *Performance of speech recognition*

ID	Correctness	Accuracy	Native	Non-native	Unsupervised	Transcription	Method
1.	42.01 %	31.39 %	820 h	-	-	-	native model
2.	88.37 %	77.32 %	890 h	890 h	-	-	non-native model
3.	89.32 %	79.72 %	890 h	890 h	25 h	-	2.+ accuracy=100%
4.	89.58 %	80.40 %	890 h	890 h	34 h	-	2.+ correct=100%
5.	**91.28 %**	**81.59 %**	890 h	890 h	**62 h**	-	2.+ **correct>80%**
6.	89.90 %	80.43 %	890 h	890 h	-	25 h	2.+transcription
7.	90.74 %	81.46 %	890 h	890 h	-	50 h	2.+transcription

Table 4: *Result of questionnaire survey*

Question	Yes	Yes if anything	No if anything	No	No answer	Positive	Negative
Are you good at English?	13 (8.78%)	50 (33.8%)	52 (35.1%)	32 (21.6%)	1 (.68%)	63 (42.6%)	85 (**57.4%**)
Was this speaking practice fun?	34 (22.9%)	77 (52.0%)	23 (15.5%)	14 (9.46%)	0 (.00%)	111 (**75.0%**)	37 (25.0%)
Was the amount spokenincreased?	34 (22.9%)	95 (64.2%)	10 (6.75%)	7 (4.73%)	2 (1.35%)	129 (**87.2%**)	19 (12.8%)

background speech noise, the conventional approach of strict evaluation based on accuracy would reduce the student's motivation. On the other hand, our system offers flexibility in evaluation as it uses correctness or regular expression to suppress insertion errors.

Adding non-native data (1→2), significantly improved the recognition rate. Conventional native model was not useful for learning applications based on speech recognition. Adding completely accurate recognition results (2→3) as unsupervised training data improves the recognition rate.

In the actual class room, since surrounding student speech noise is significant, the recognition results often have a lot of insertion errors. By selecting data with 100 % in correctness and training with the target sentence, the acoustic model can handle the surrounding speech noise and thus reduce insertion errors. By ignoring insertion errors (3→4) in unsupervised training data, the recognition rate is improved. Since it is difficult to recognize Japanese-accented English speech completely, the volumes of selected data aren't large. By adding data with slightly higher correct rates (correctness $>$ 80%, 4→5), the recognition rate is improved significantly due to the increase in data size. As a result, adding unsupervised data (2→5) improves the recognition rate significantly. The proposed unsupervised approach (4 and 5) could match the recognition rates possible with transcription (6 and 7) while eliminating manual transcribing costs.

3.2. Evaluation of user survey

We also evaluated the user impression of this application. We collect questionnaire responses from the 148 students. Table 4, Table 5 and Table 6 show the questionnaire items and their responses.

As Table 4 shows, although the number of English-poor students dominated the rest, most students well liked our application in the classroom. Our application yielded more English speech than the previous approach of practicing with teachers or classmates. To summarize, our application encourages students to speak more English more often.

Our computer-assisted training system can reduce the issue of concern from Table 5, because students can study by themselves using this system with the effect written in Table 6.

The trial itself was broadcasted as news on a television network and Youtube [26]; it contained the positive interview from the principal of the junior high school. After this two month trial, the number of applicants to the English proficiency test increased.

Table 5: *Questionnaire results about"why do you feel that you aren't good at English speaking?"*

Answer	Count	Rate
Can't find a proper English expression	74	50.0%
Have no confidence about pronunciation	59	39.8%
Feel nervous in front of people.	53	35.8%
Worried about communication	37	25.0%

Table 6: *Questionnaire results about "what's the good point of speaking training with tablet device?"*

Answer	Count	Rate
Learn at own pace	108	72.9%
Understand answer quickly	88	59.5%
Feel good even if wrong	77	52.0%
Not embarrassing & nervous	49	33.1%

4. Conclusion

We investigated the effectiveness of a iPad-based speech training application that used speech recognition in an actual classroom. Because many students speak at the same time in classrooms, input speech is often contaminated by speech noise from the other students. Few studies have focused on the effects of noise present in classrooms. Since the existing speaking applications assume that students use the application alone, they can't evaluate the input correctly in presence of classroom noise. To improve students' motivation, our application ignores the insertion errors around the user's intended sentence. We introduced unsupervised adaption with matched sentences by comparing the recognition results and target sentences. We also improved the user interface after considering the feedback from teachers and students. A two month trial with 148 students in a public junior high school found that the proposal improves speech recognition accuracy and most students had a positive opinion of our application.

5. References

[1] C. Baur, A. Caines, C. Chua, J. Gerlach, M. Qian, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2018 spoken call shared task," in *Interspeech*, 2018, pp. 2354–2358.

[2] G. W. Soad, N. F. D. Filho, and E. F. Barbosa, "Quality evaluation of mobile learning applications," in *IEEE Frontiers in Education Conference (FIE)*, 2016.

[3] X. Anguera and V. Van, "English language speech assistant," in *Interspeech*, 2016, pp. 1962–1963.

[4] J. Bernstein, J. Cheng, J. Balogh, and E. Rosenfeld, "Studies of a self-administered oral reading assessment," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 172–176.

[5] E. Gilmartin, J. Kim, A. Diallo, Y. Zhao, N. N. Chiarain, K. Su, Y. Huang, H. B. R. Cowan, and N. Campbell, "CARAMILLA – speech mediated language learning modules for refugee and. high school learners of English and Irish," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 138–143.

[6] C. Tejedor-García, D. Escudero-Mancebo, C. González-Ferreras, E. Cámara-Arenas, and V. C. noso Payo, "Evaluating the efficiency of synthetic voice for providing corrective feedback in a pronunciation training tool based on minimal pairs," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 25–29.

[7] D. Huynh, L. Zuo, and H. Iida, "An assessment of game elements in language-learning platform Duolingo," in *4th International Conference on Computer and Information Sciences (ICCOINS)*, 2018.

[8] E. Godde, G. Bailly, D. Escudero, M.-L. Bosse, M. Bianco, and C. Vilain, "Improving fluency of young readers: introducing a Karaoke to learn how to breath during a reading-while-listening task," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 127–131.

[9] A. Khalifa, T. Kato, and S. Yamamoto, "Measuring effect of repetitive queries and implicit learning with joining-in-type robot assisted language learning system," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 13–17.

[10] X. C. Sanchez, L. P. Gavilanez, and A. M. Fiallos, "The effects of using 'soundcloud' on speaking performance of Ecuadorian students," in *International Conference on eDemocracy & eGovernment (ICEDEG)*, 2018.

[11] R. Downey, M. Suzuki, and A. V. Moere, "High-stakes English-language assessments for aviation professionals: Supporting the use of a fully automated test of spoken-language proficiency," *IEEE Transactions on Professional Communication*, vol. 53, pp. 18–32, 2010.

[12] J. Yue, D. Saito, N. Minematsu, and Y. Yamauchi, "Development and maintenance of practical and in-service systems for recording shadowing utterances and their assessment," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, p. 189.

[13] Y. Luan, M. Suzuki, Y. Yamauchi, N. Minematsu, S. Kato, and K. Hirose, "Performance improvement of automatic pronunciation assessment in a noisy classroom," in *IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 428–431.

[14] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2, pp. 295–108, 2000.

[15] R. Shadiev, Y.-M. Huang, W.-Y. Hwang, and N. Shadiev, "Investigating the effectiveness of speech-to-text recognition application on learning performance in traditional learning environment," in *IEEE 15th International Conference on Advanced Learning Technologies (ICALT)*, 2015, pp. 441–445.

[16] B.-R. Luo, Y.-L. Lin, N.-S. Chen, and W.-C. Fang, "Using smartphone to facilitate English communication and willingness to communicate in a communicative language teaching classroom," in *IEEE 15th International Conference on Advanced Learning Technologies (ICALT)*, 2015, pp. 320–322.

[17] R. Masumura, S. Kabashima, T. Moriya, S. Kobashikawa, Y. Yamaguchi, and Y. Aono, "Relevant phonetic-aware neural acoustic models using native English and Japanese speech for Japanese-English automatic speech recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1435–1439.

[18] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Transfer learning based non-native acoustic modeling for pronunciation error detection," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2017, pp. 42–46.

[19] A. Nurhudatiana, A. N. Hiu, and W. Ce, "Should I use laptop or smartphone? a usability study on an online learning application," in *International Conference on Information Management and Technology (ICIMTech)*, 2018, pp. 565–570.

[20] "KNOUN," https://knoun.jp/.

[21] M. Fujimoto, S. Watanabeand, and T. Nakatani, "Frame-wise model re-estimation method based on gaussian pruning with weight normalization for noise robust voice activity detection," *Speech Communication*, vol. 54, no. 2, pp. 229–244, 2012.

[22] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Arakiand, and T. Nakatani, in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 436–443.

[23] T. Moriya, H. Kanagawa, K. Matsui, T. Fukutomi, Y. Shinohara, Y. Yamaguchi, M. Okamoto, and Y. Aono, "Efficient building strategy with knowledge distillation for small-footprint acoustic models," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 21–28.

[24] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "VoiceRex - spontaneous speech recognition technology for contact-center conversations," *NTT Technical Review*, vol. 5, no. 1, pp. 22–27, 2007.

[25] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1352–1365, 2007.

[26] "ANNnewsCH," https://www.youtube.com/watch?v=zIvBh5hM_0w.