How can speech technologies support learners to improve their skills of speaking, listening, conversation, and more?

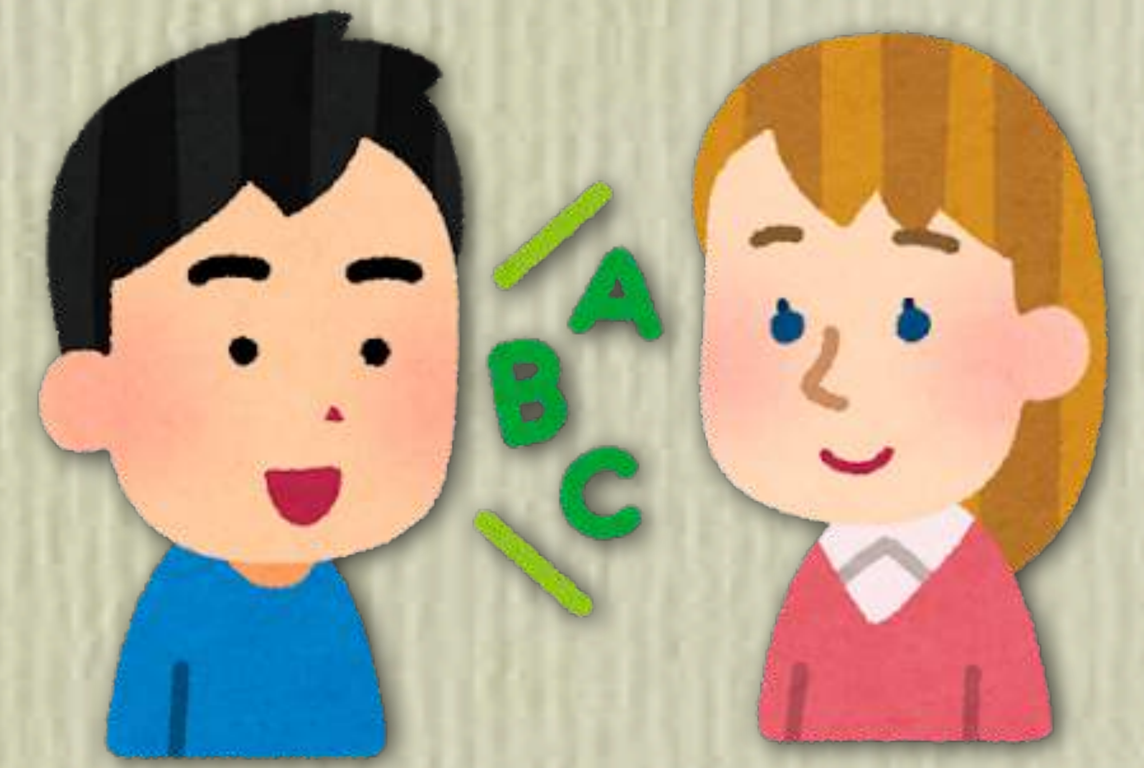
Nobuaki MINEMATSU
Graduate School of Engineering,
The University of Tokyo

Language learning at schools and society

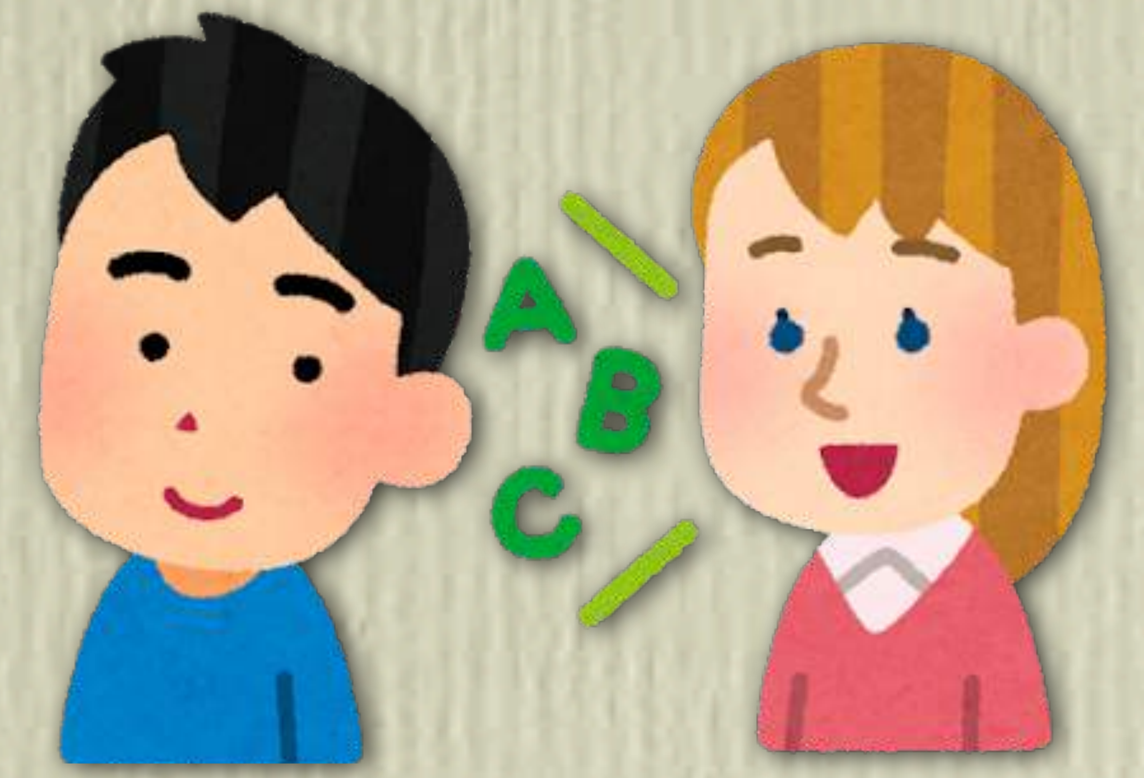
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

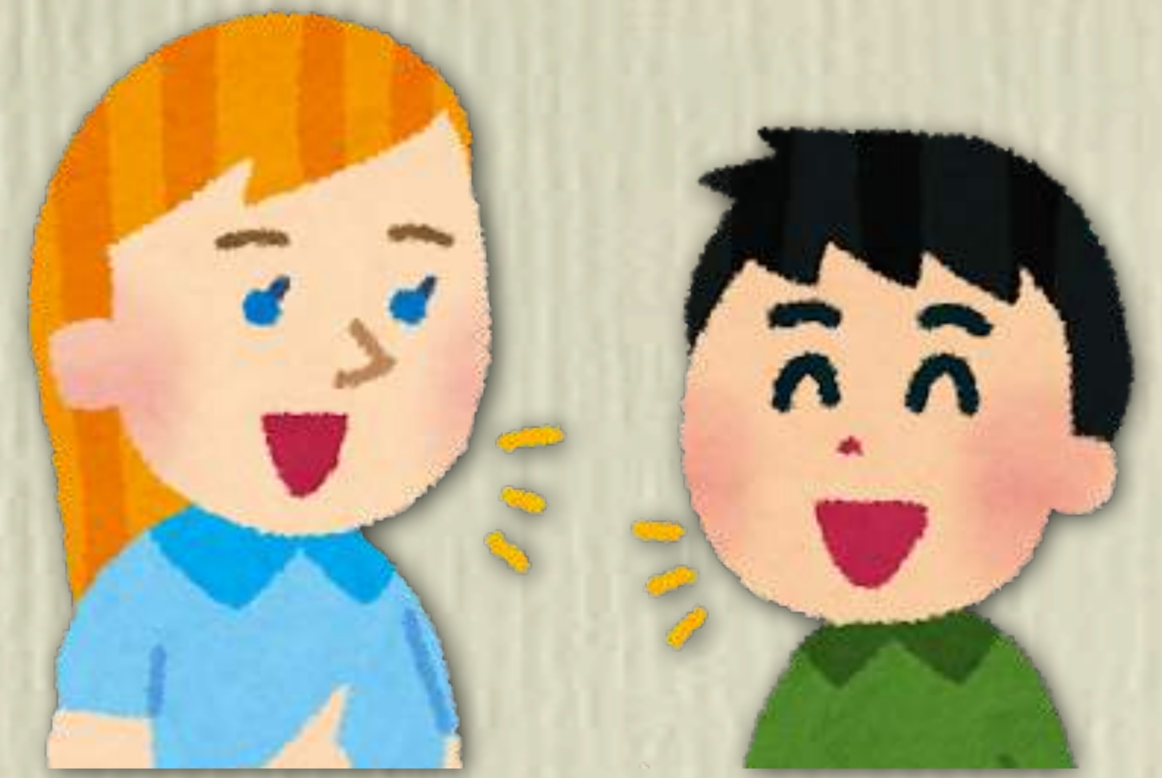
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



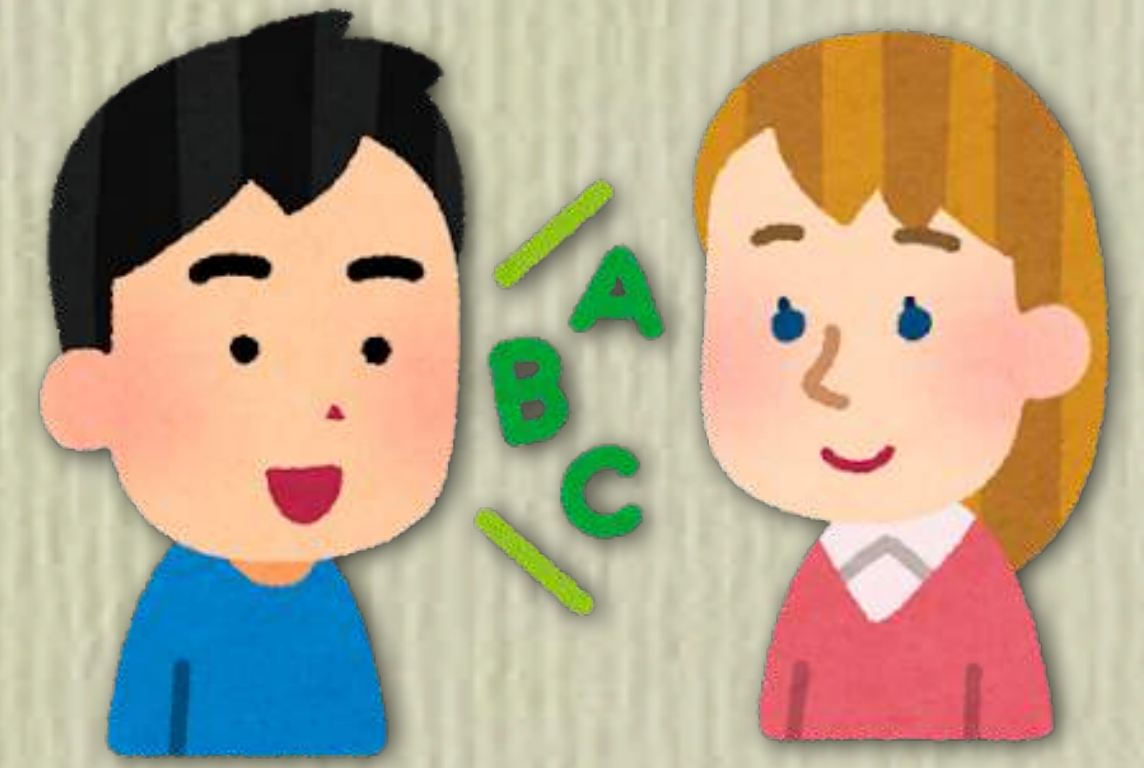
with speech recognition technologies

with new speech technologies being developed in our new project

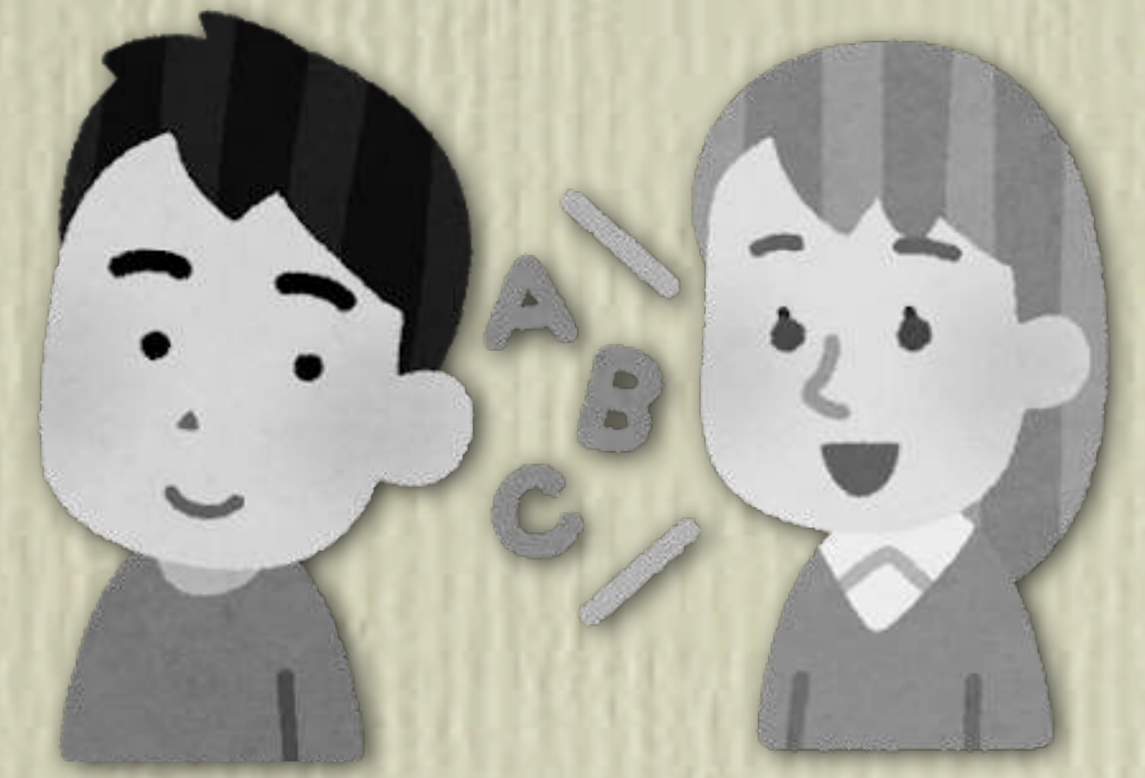
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

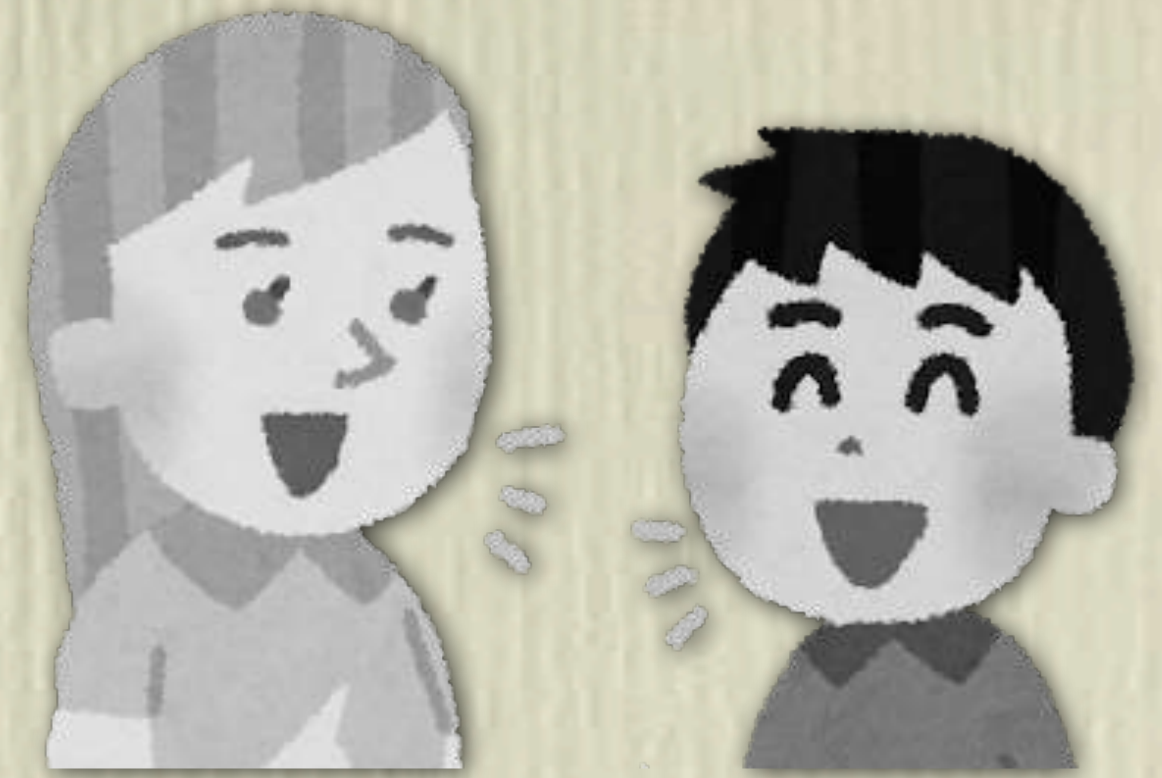
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies ✔



with speech analysis technologies



with speech recognition technologies

with new speech technologies being developed in our new project

Word accent of Japanese and its control while speaking

Japanese word accent is pitch accent (H/L accent).

- The pitch value (H/L) has to be controlled and changed according to context.
- Prosody control, including word accent control, is rarely taught in classes.

Examples of accent changes when speaking

- A noun + another = a compound noun
  - あか ＋ えんぴつ → あかえんぴつ
- Verb conjugation
  - あるく → あるきます,　あるいて,　あるいた,　あるかない
- A bunsetsu + another = an accentual phrase
  - わたしは ＋ たべる → わたしはたべる　　かれは ＋ たべる → かれはたべる

さんがつ	ひこーき	かんごふ	いもーと	おはなみ
頭高型 initial high	中高型 middle high		尾高型 tail high	平板型 unaccented
起伏式 accented				平板式 unaccented
1型 type 1	2型 type 2	3型 type 3	4型 type 4	0型 type 0
−4型	−3型	−2型	−1型	

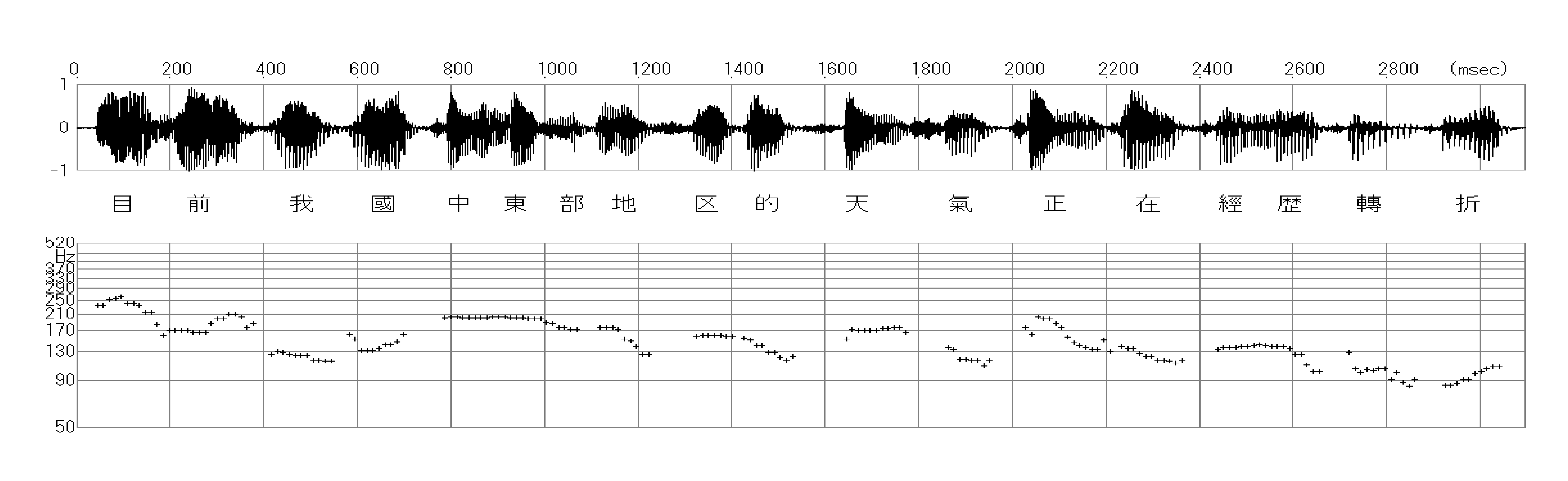
Word accent control of Japanese is
SOOOO MYSTERIOUS!!

N. Minematsu, et al., “Automatic … nese speech Conversion,” Trans. IEICE, E86-D, 3, 550-557, 2003

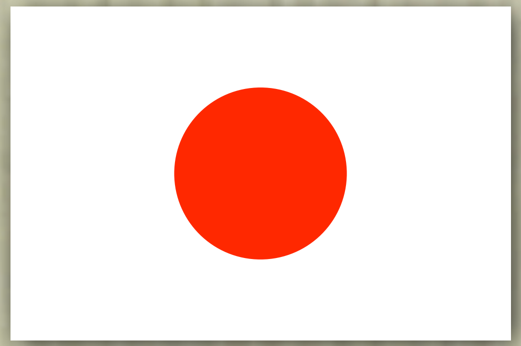
Differences in controlling phrase intonation bet. C and J

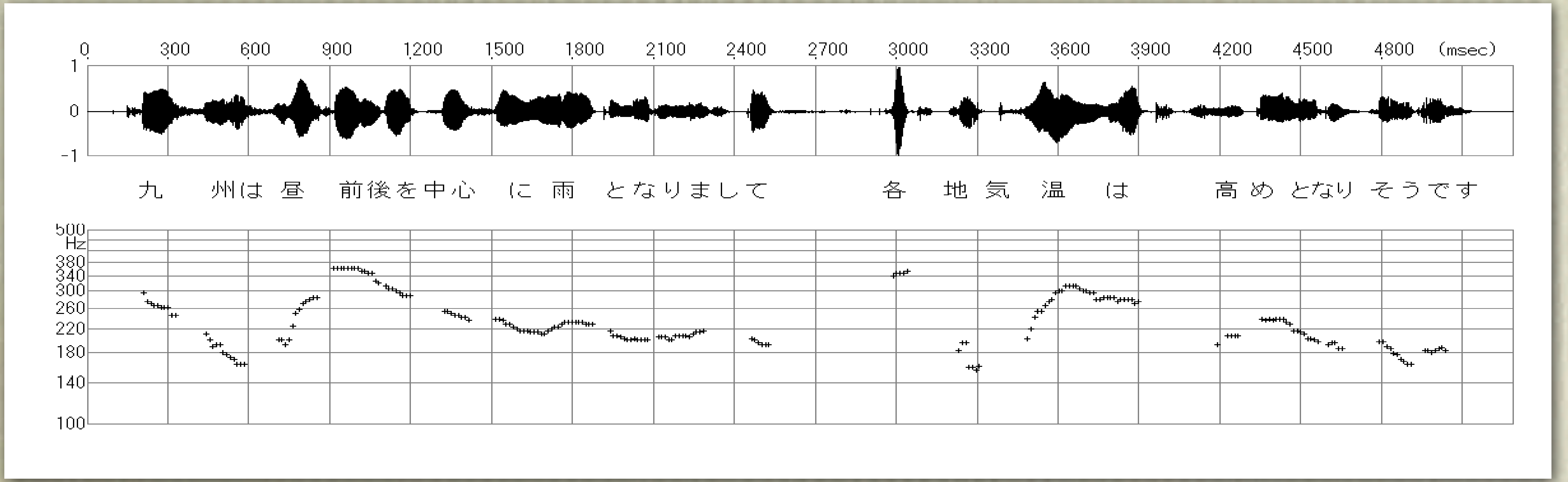
An interesting example of comparison between Chinese and Japanese

- Pitch changes acoustically observed in a Chinese utterance (weather forecast)



- Pitch changes acoustically observed in a Japanese utterance (weather forecast)

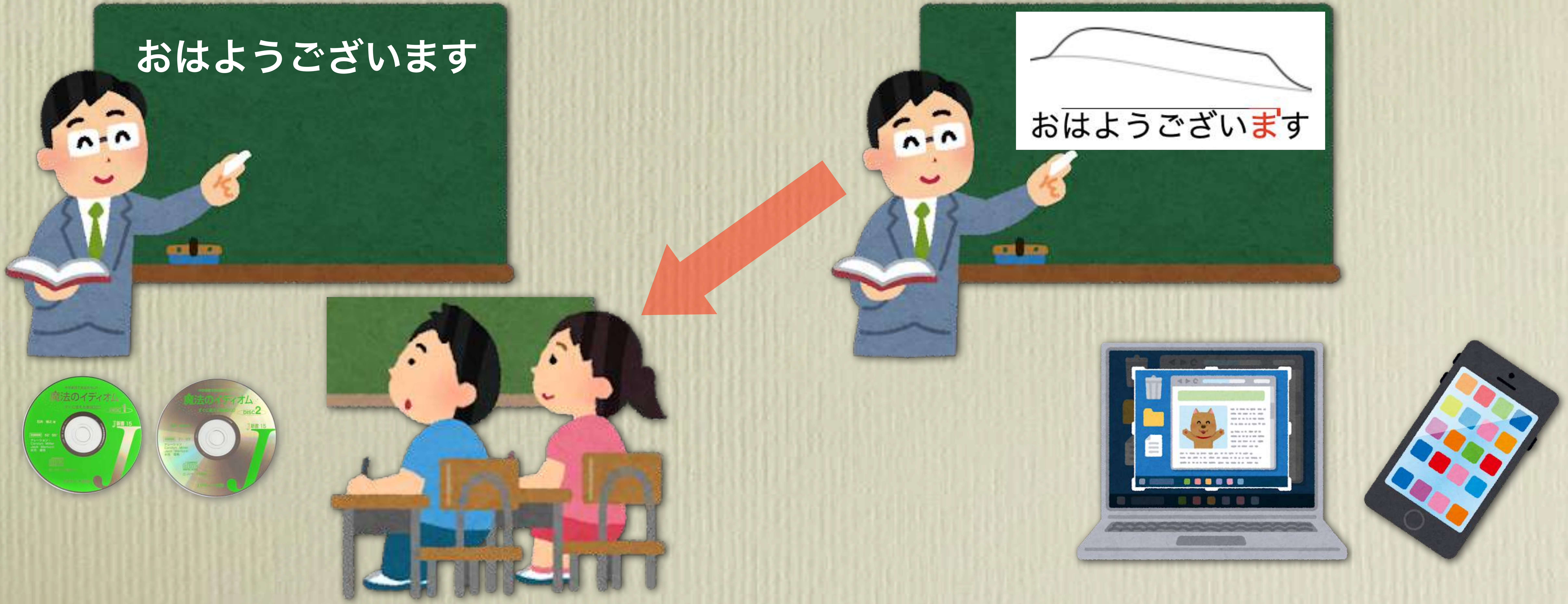




Two kinds of language teachers

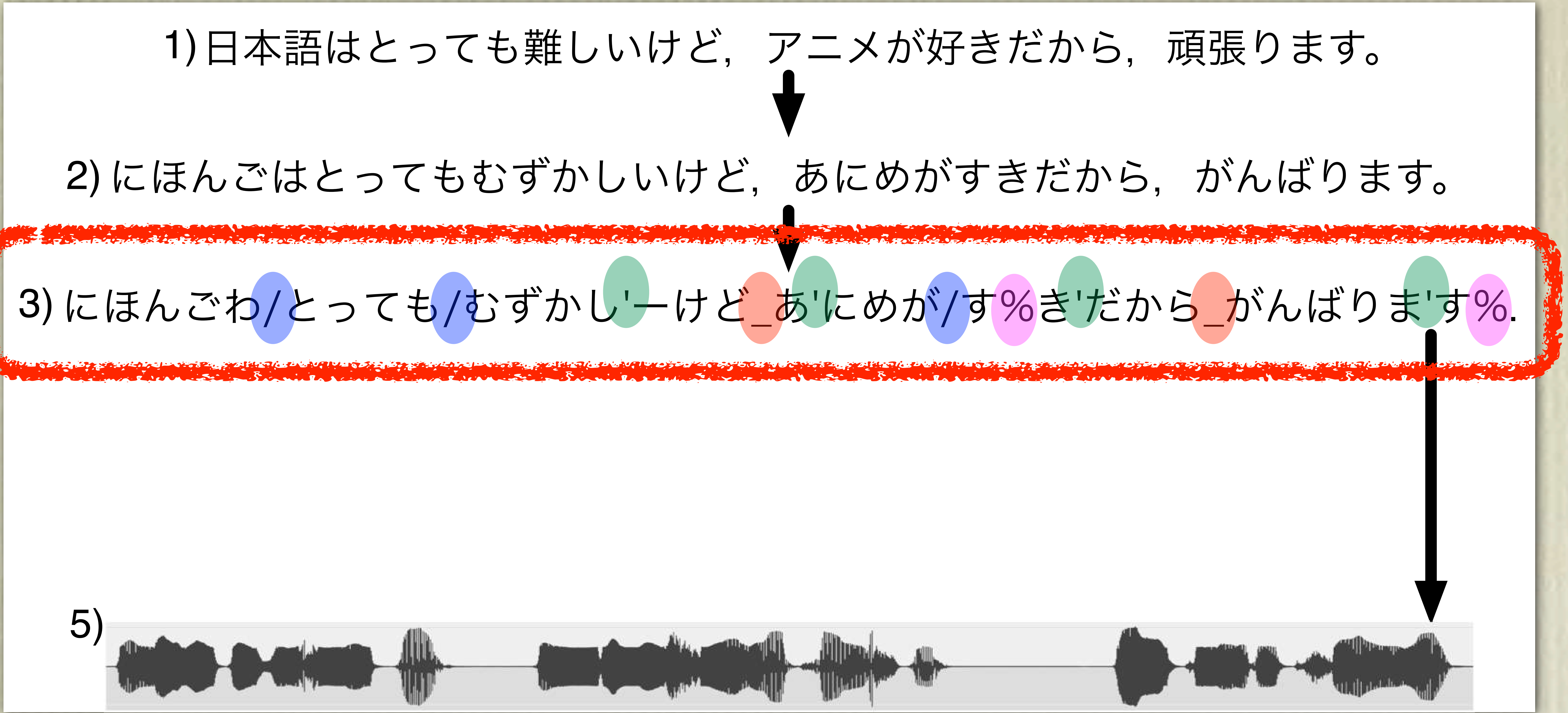
Those teaching to human learners and those to machine learners

- More-than-50-year history of teaching Japanese prosody to machine learners



TTS technologies are effectively introduced.

Visualization of prosodic control for speaking in Tokyo Japanese.



M. Suzuki, et al., “Accent sandhi estimation of Tokyo dialect of Japanese using conditional random fields,” Trans. IEICE, E100-D, 4, 655-661, 2017 (IEICE ISS Paper Award)

N. Minematsu, et al., “Development and evaluation of online infrastructure to aid teaching and learning of Japanese prosody,” Trans. IEICE, E100-D, 4, 662-669, 2017 (IEICE ISS Paper Award, PSJ Academic Award)

1.5-min promotion video for Suzuki-kun of OJAD

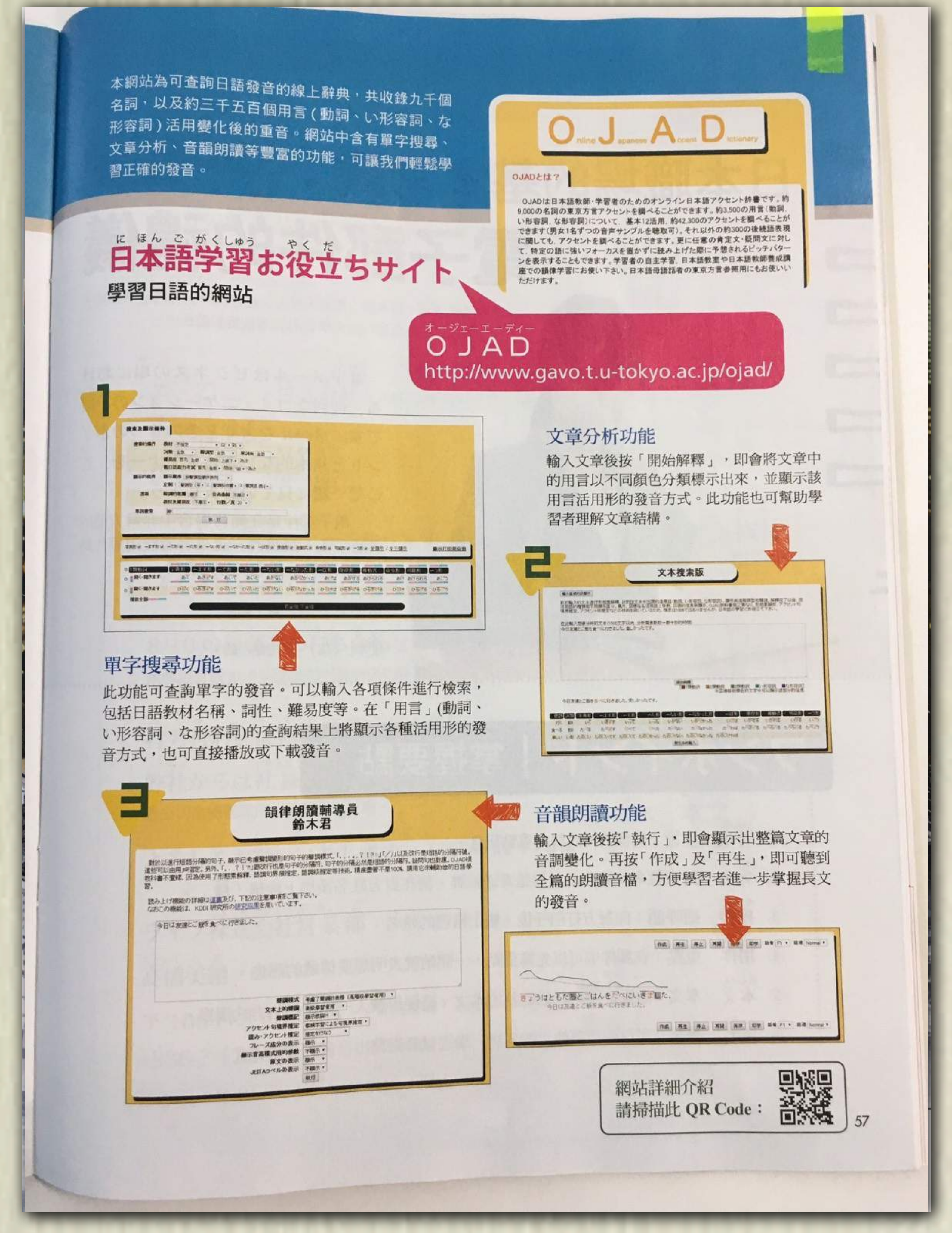
Suzuki-kun = prosodic reading tutor of Tokyo Japanese in OJAD

- "The first and only teaching material to explain prosodic control of TJ for any given text."

Thank you for learning Japanese with OJAD!

Many teachers and learners are using OJAD all over the world.

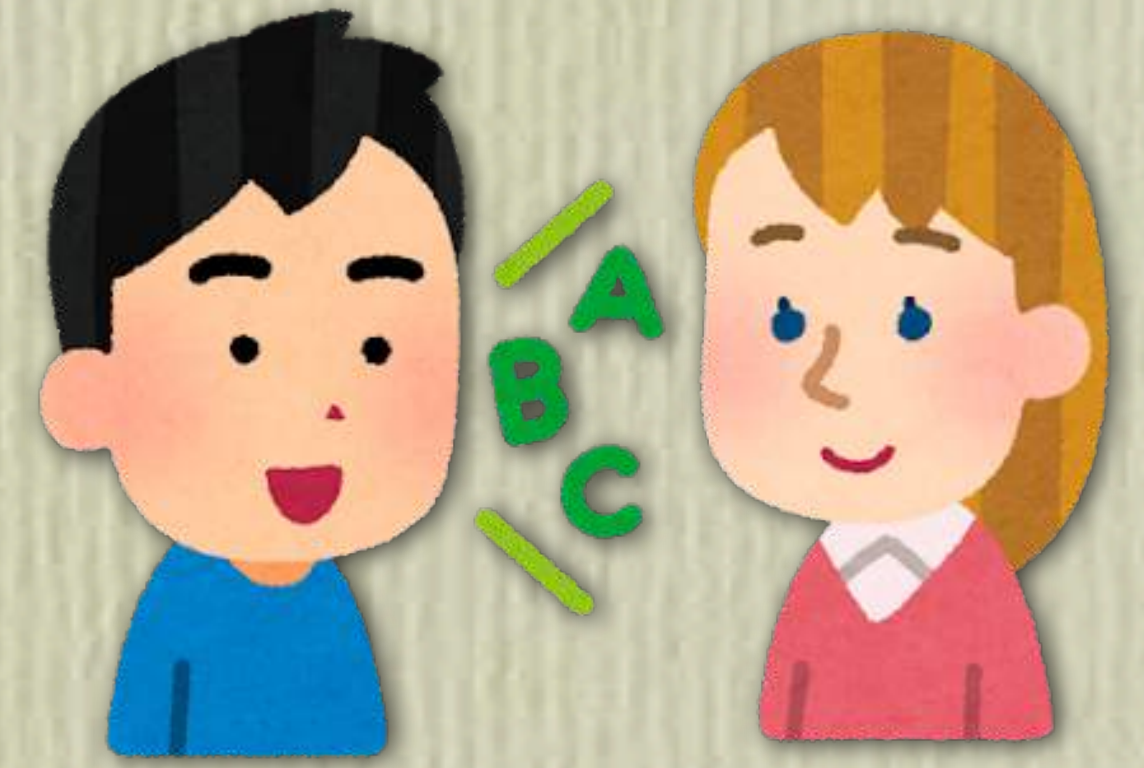
- “OJAD is MUST to participate in a speech contest and win the championship!!”



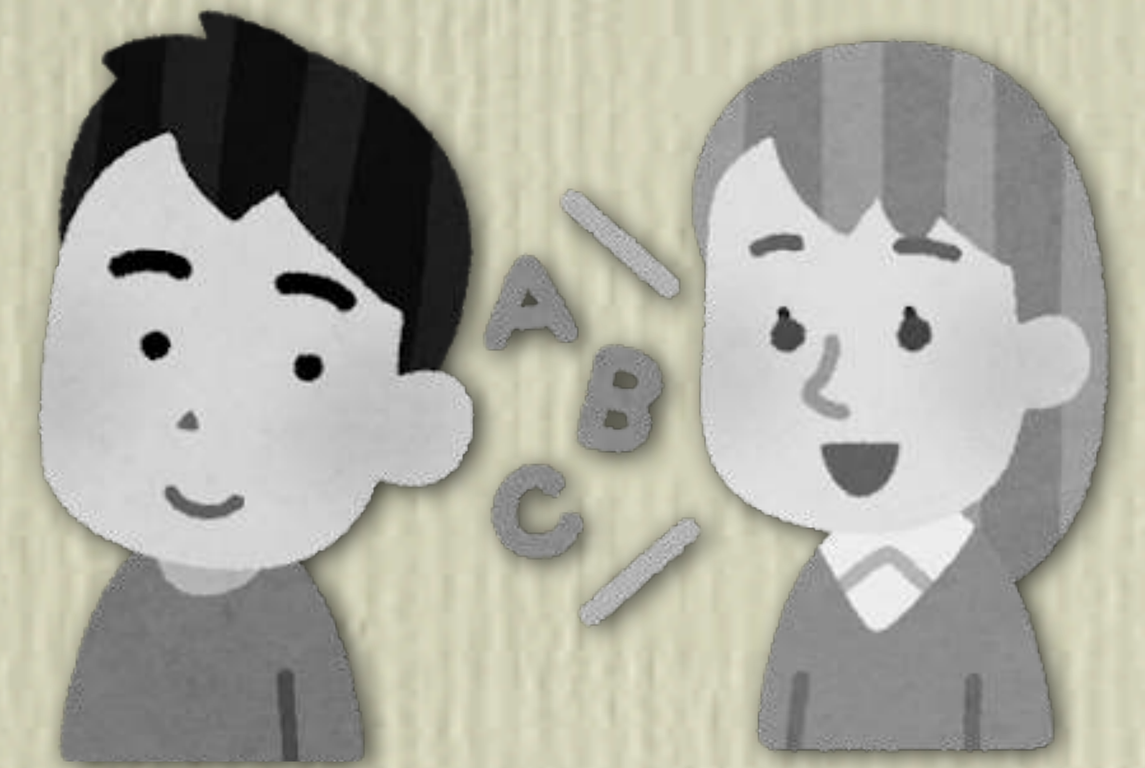
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

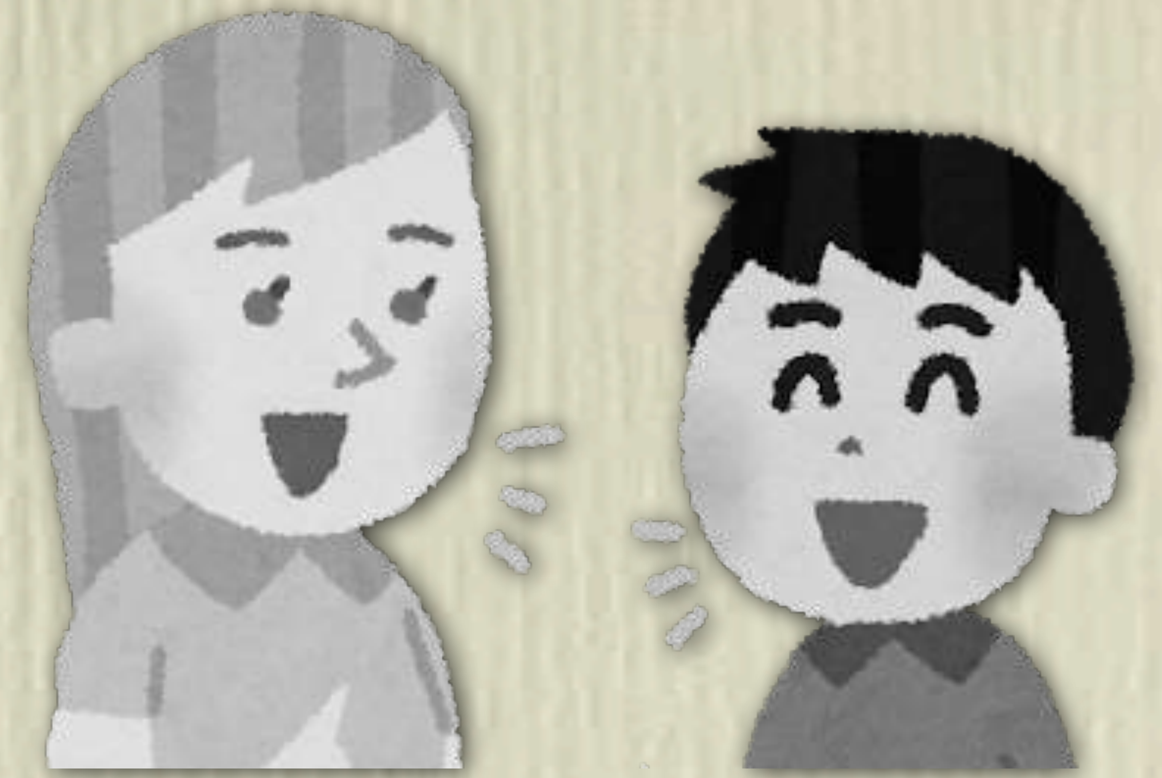
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies ✓



with speech analysis technologies



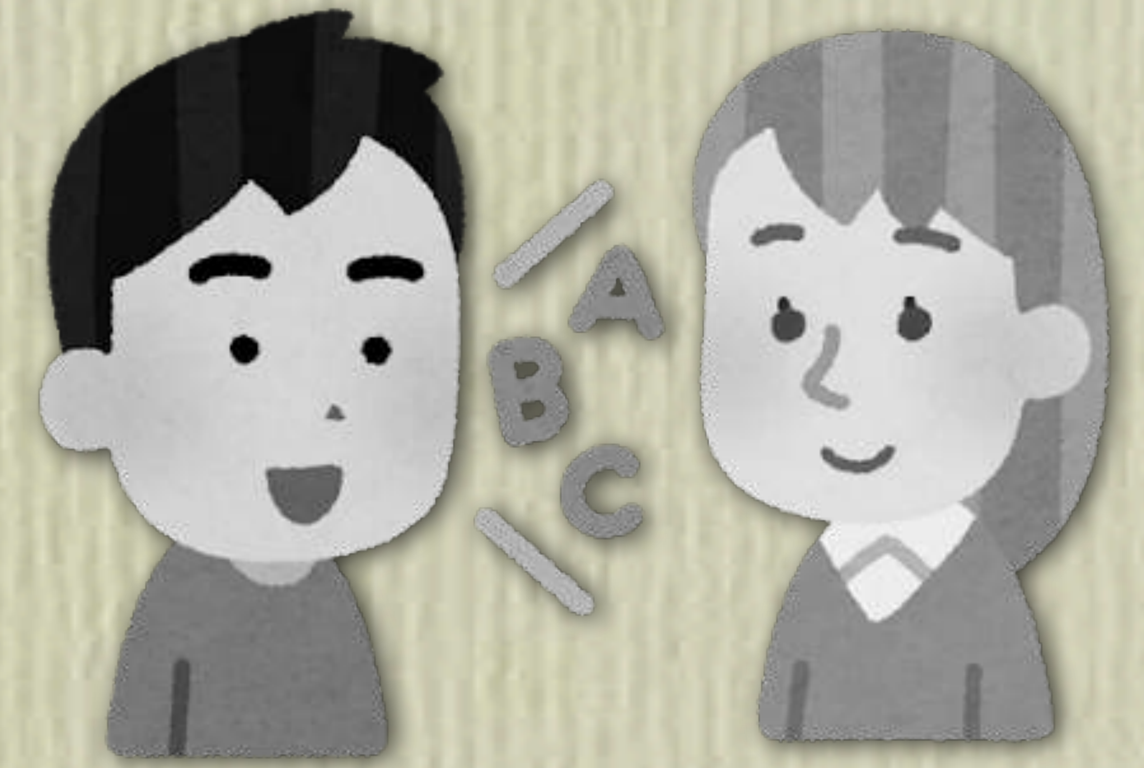
with speech recognition technologies

with new speech technologies being developed in our new project

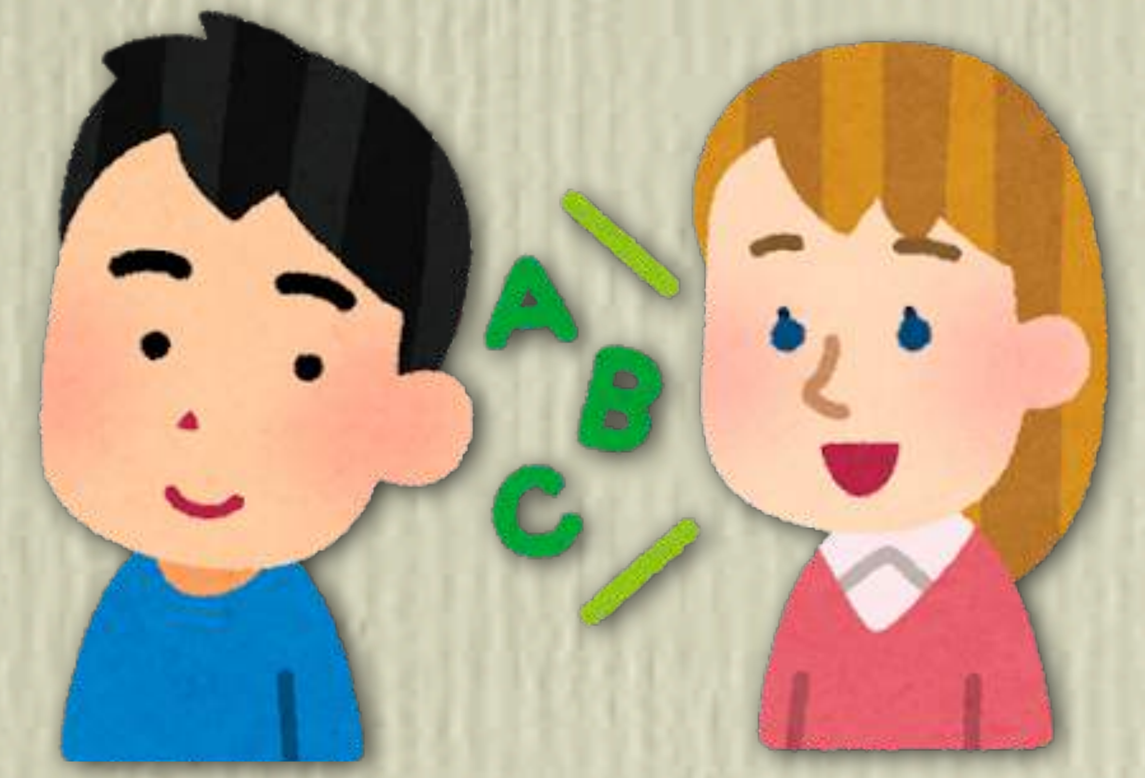
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

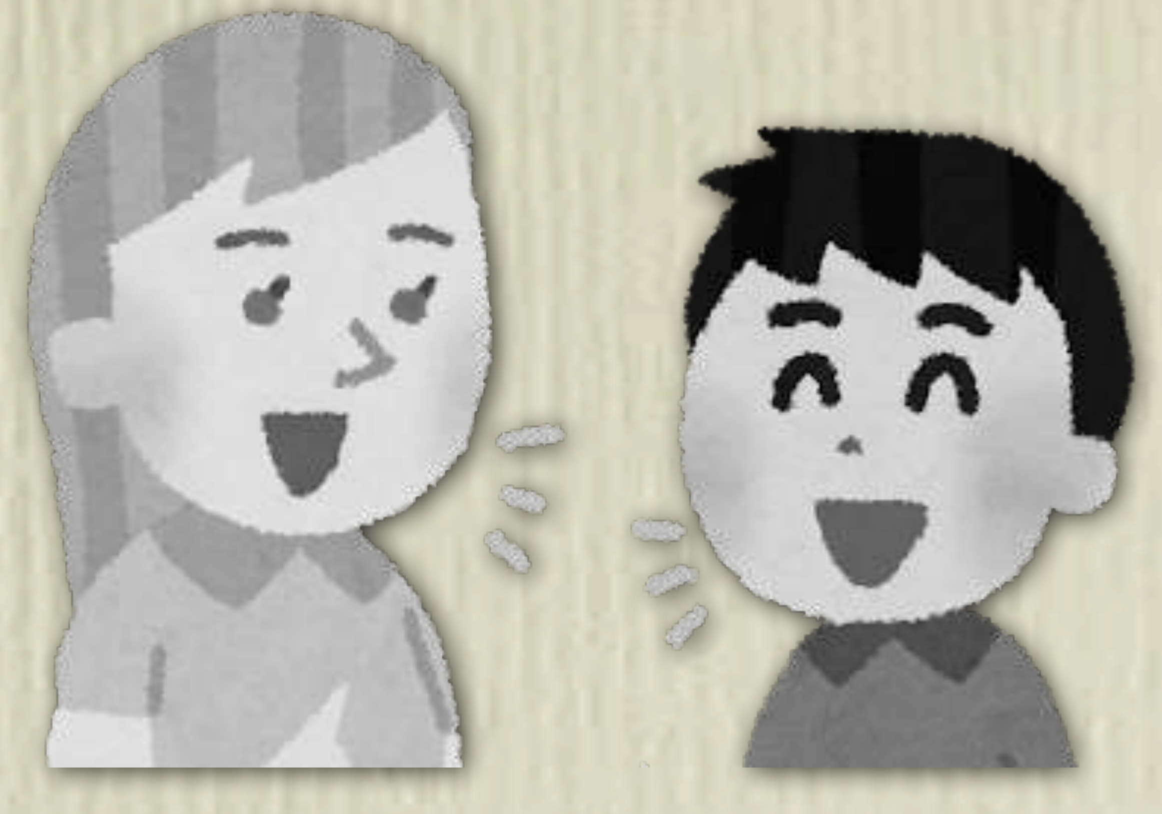
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



with speech recognition technologies

with new speech technologies being developed in our new project

A honest confession from a young Japanese pilot

命がけのリスニング

何か新しいことにチャレンジしてみようってことで、軽い気持ちで飛行機の練習をはじめたのが半年前。
泥沼にはまりつつも、なんとかもう少しで取れそうなところまで来た。

まだ終わったわけではないのだけど、一番大変だったのは無線交信。
予想以上にパイロットはしゃべる仕事だということを痛感。

特に、私が練習している空域は、北米でも有数の混雑地帯で、無線交信の量がはんぱじゃない。
飛行中は、常に誰かがしゃべっているという感じ。そして、管制塔の指示がわからなかったら、最悪の場合、衝突の可能性もあるわけで、まさに命がけ。

それなのに、

・無線なのでノイズが大きい
・管制官が早口で訛っている
・パイロットも訛っている
・エンジンの音が大きい
・無線機がボロくて、たまに聞こえなくなる
・エリアによっては妨害電波が出ているみたい
・コクピットの中はただでさえ緊張して、頭の中が白くなる

Desperate efforts needed for listening

After becoming a pilot, I realized that a pilot has to talk always with air traffic controllers, and it is under severely degraded conditions.

- machine noses
- communication (channel) noises
- regional and foreign accents
- speaking very fast
- etc

Listening robustness is needed even in daily conversations!!

- Trains, cars, buses, restaurants, airports, telephones, big halls, etc
- Different places may cause different types of acoustic degradation.

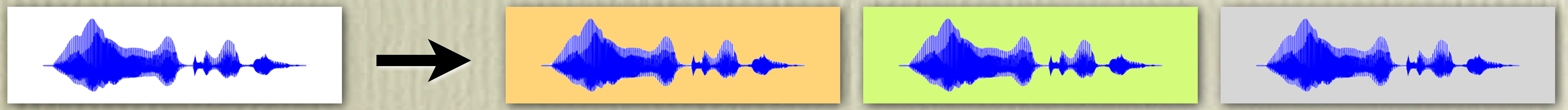
A training method for robust listening

High Variability Phonetic Training (HVPT)

- Listening training using speech samples with acoustically high variability
  - Speakers, speaking style, gender, age, accents, background noises, etc
- Many articles showed the effectiveness of HVPT.
  - Lively+1993, Masuda+2012, Wong+2014, Hwang+2015
- Teachers often collect various audio samples *manually*.

Technically-enhanced variability in HVPT

- Speech analysis-resynthesis technologies
  - can convert a single utterance into acoustically various versions with its message unchanged.
- HVPT with artificially converted audio samples



....

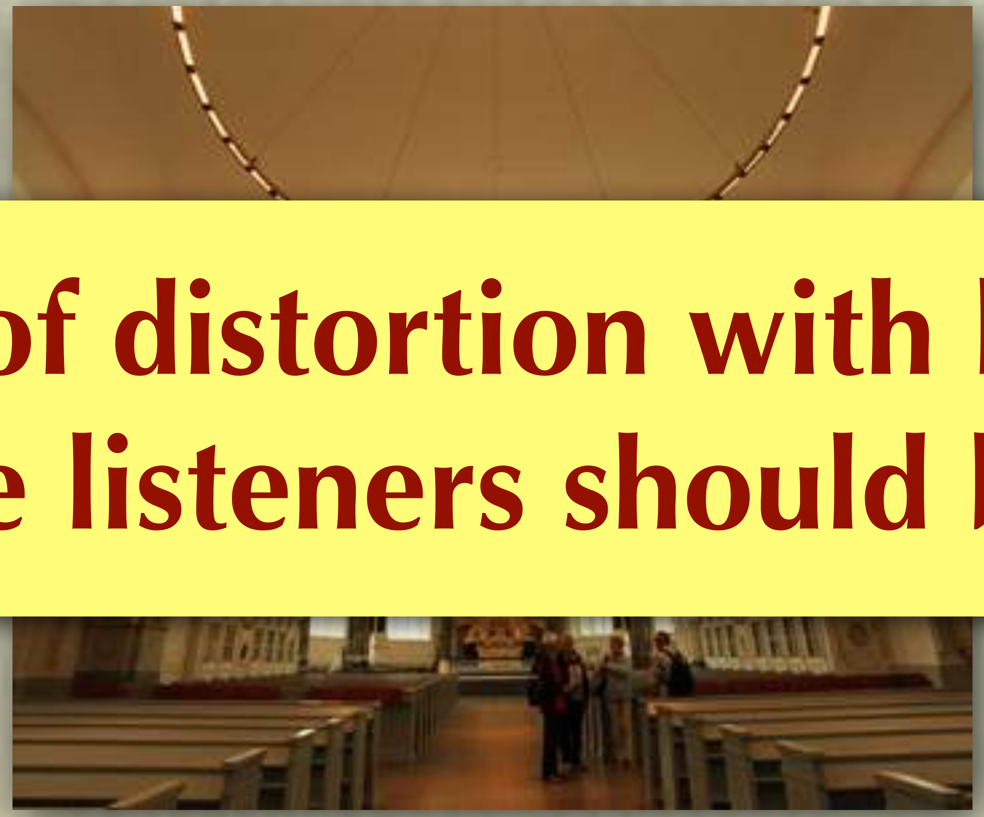
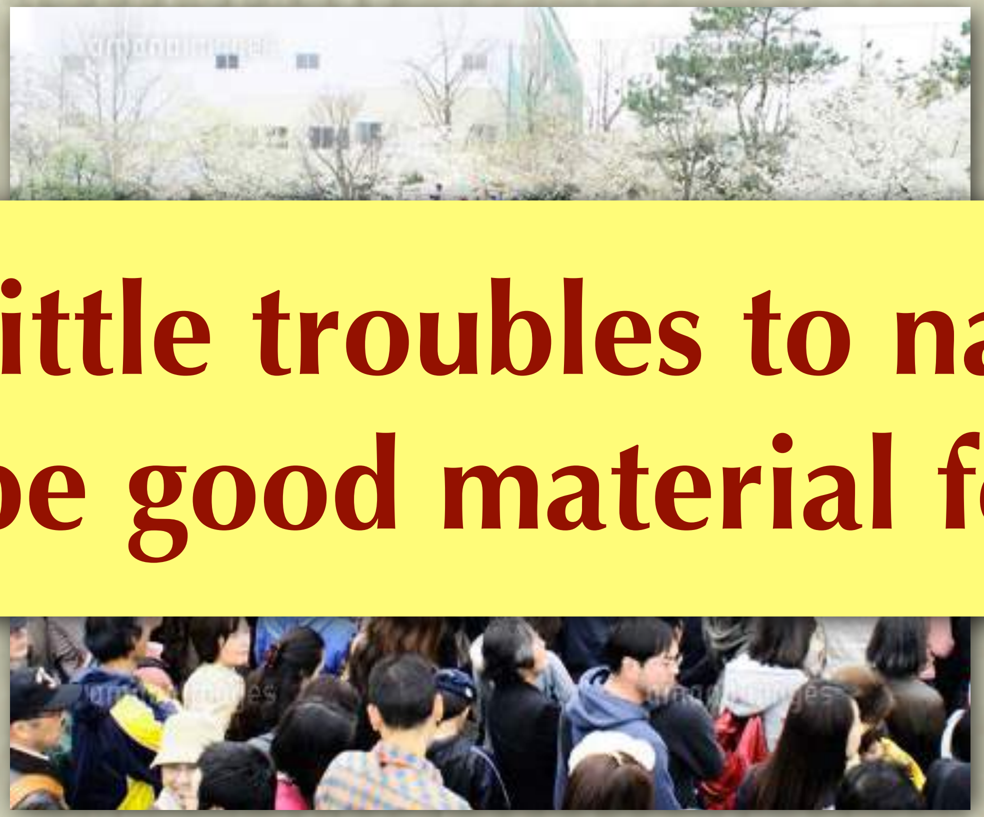
H. Zhang, et al., “Computer-aided high variability phonetic training to improve robustness of learners’ listening comprehension,” Proc. ICPhS, 2019

A. Guevara-Rukoz, et al, “Prototyping a web-based phonetic training game to improve /r/-/l/ identification by Japanese learners of English,” Proc. SLaTE 2019 (Best Paper Award)

Examples of speech conversion

Variously converted speech can be obtained easily.

- **Original** “February 14th is a day for people who are falling in love.”
- **VTL** VTL x 1.5 (giant), VTL / 1.5 (fairy)
- **Reverb** a big cathedral
- **Noise** babble noise (voice noise)
- **Channel** 2G mobile phone, air traffic control (ATC)
- **Combination with quantitative control of degree of distortion**
  - A small girl is praying in a cathedral, surrounded by chatty tourists and her pray is recorded and transmitted via. a 2G mobile phone network.

Specific types of distortion with little troubles to native listeners but big troubles to non-native listeners should be good material for robust listening training?

Very difficult EIKEN grade 2 listening test

4-choice questions after listening to monologues or dialogues

- Male ➙ giant pilot (ATC)
- Female ➙ fairy pilot (ATC)

Question: What is one thing the girl says?

1 She is not good at sports.
2 She will not go to college.
3 She needs more time to study.
4 She wants to practice basketball more.

Accuracy of Japanese college students and native speakers

TOEIC	original	G/F	ATC	G/F + ATC
400-600	58.3			
600-800	78.2			
800-990	81.5			
Native				

Very difficult EIKEN grade 2 listening test

4-choice questions after listening to monologues or dialogues

- Male → giant pilot (ATC)
- Female → fairy pilot (ATC)

Question: What is one thing the girl says?

1 She is not good at sports.
2 She will not go to college.
3 She needs more time to study.
4 She wants to practice basketball more.

Accuracy of Japanese college students and native speakers

TOEIC	original	G/F	ATC	G/F + ATC
400-600	58.3	**50.0**	**30.6**	**32.8**
600-800	78.2	**62.0**	**35.1**	**23.4**
800-990	81.5	**79.6**	**45.4**	**25.0**
Native				

Very difficult EIKEN grade 2 listening test

4-choice questions after listening to monologues or dialogues

- Male → giant pilot (ATC)
- Female → fairy pilot (ATC)

Question: What is one thing the girl says?

1 She is not good at sports.
2 She will not go to college.
3 She needs more time to study.
4 She wants to practice basketball more.

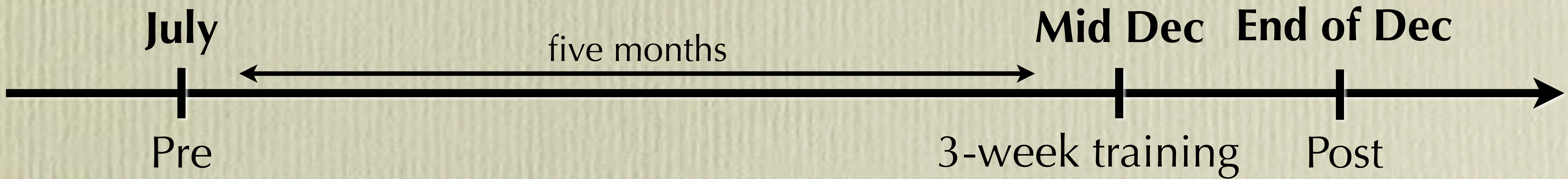
Accuracy of Japanese college students and native speakers

TOEIC	original	G/F	ATC	G/F + ATC
400-600	58.3	**50.0**	**30.6**	**32.8**
600-800	78.2	**62.0**	**35.1**	**23.4**
800-990	81.5	**79.6**	**45.4**	**25.0**
Native	**100**	**100**	**100**	**93.6**

Pre-tests → special drills with ATC → post-tests

Procedure of the experiments

- Pre: EIKEN G2 listening tests with original, **GF**, **ATC**, and **GF+ATC** samples.
- Training: Listening drills with varying degrees of **ATC distortions only**
- Post: EIKEN G2 listening tests (= Pre)



Pre-tests → special drills with ATC → post-tests

Procedure of the experiments

- Pre: EIKEN G2 listening tests with original, **GF**, **ATC**, and **GF+ATC** samples.
- Training: Listening drills with varying degrees of **ATC distortions only**
- Post: EIKEN G2 listening tests (= Pre)

Effects of technically-enhanced HVPT

- Accuracies of Pre and Post (A: dialogues, B: monologues)

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	66.7	48.3	25.0	41.7
	600–800	32	77.3	65.6	38.3	25.8
	800–990	8	84.4	84.4	43.8	21.9
B	400–600	15	50.0	43.3	28.3	23.3
	600–800	32	65.6	48.4	39.1	30.5
	800–990	8	78.1	62.5	37.5	28.1

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	70.0	66.7	26.7	35.0
	600–800	32	73.4	73.4	40.6	32.8
	800–990	8	96.9	96.9	75.0	40.6
B	400–600	15	66.7	48.3	38.3	23.3
	600–800	32	61.7	51.6	42.2	35.2
	800–990	8	87.5	84.4	62.5	31.3

Pre-tests → special drills with ATC → post-tests

Procedure of the experiments

- Pre: EIKEN G2 listening tests with original, **GF**, **ATC**, and **GF+ATC** samples.
- Training: Listening drills with varying degrees of **ATC distortions only**
- Post: EIKEN G2 listening tests (= Pre)

Effects of technically-enhanced HVPT

- Error reduction rates from Pre to Post (A: dialogues, B: monologues)

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	9.9	35.6	2.3	-11.5
	600–800	32	-17.2	22.7	3.7	9.4
	800–990	8	**80.1**	**80.1**	**55.5**	23.9
B	400–600	15	33.4	8.8	13.9	0
	600–800	32	-11.3	6.2	5.1	6.8
	800–990	8	**42.9**	**58.4**	**40.0**	4.5

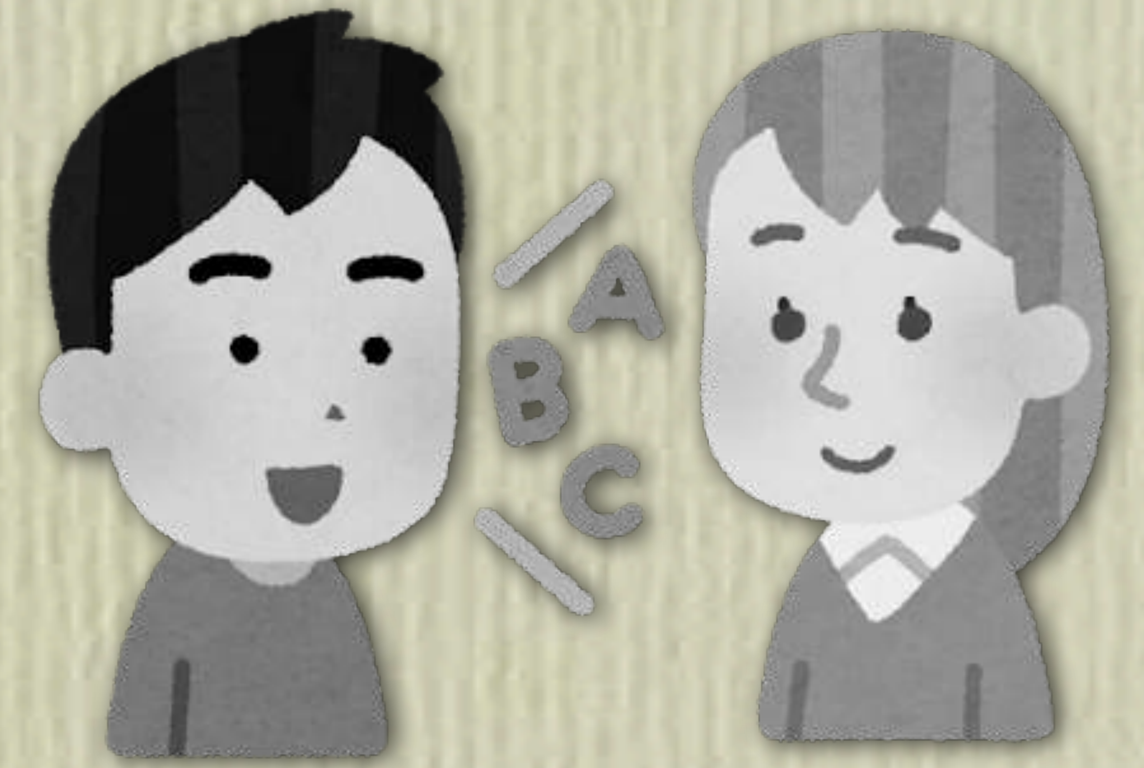
- Error Reduction Rate
  - Accuracy: 70% → 85%
  - ERR = (30-15)/30 = 50%

- In advanced learners, HVPT with ATC is very effective and ERR is larger than 40%
- Further, listening robustness was transferred effectively to other types of stimuli.
- Proposed HVPT is effective but ATC distortions seem to be too difficult for non-advanced learners.

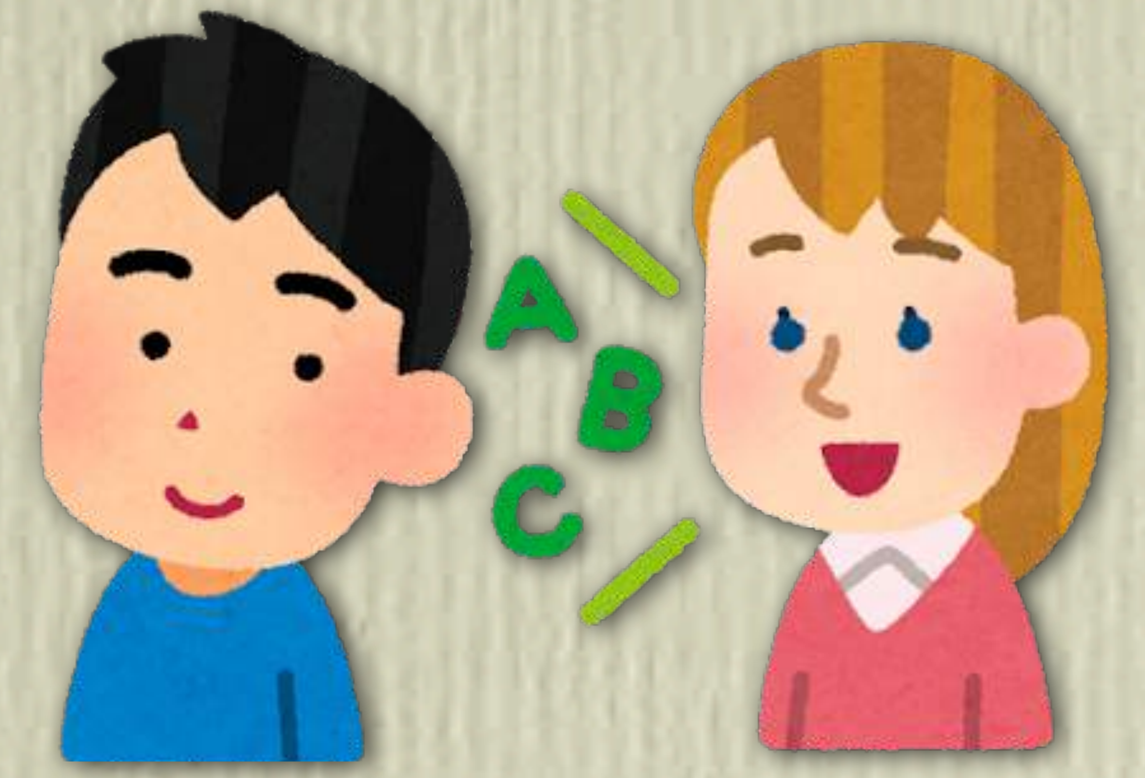
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

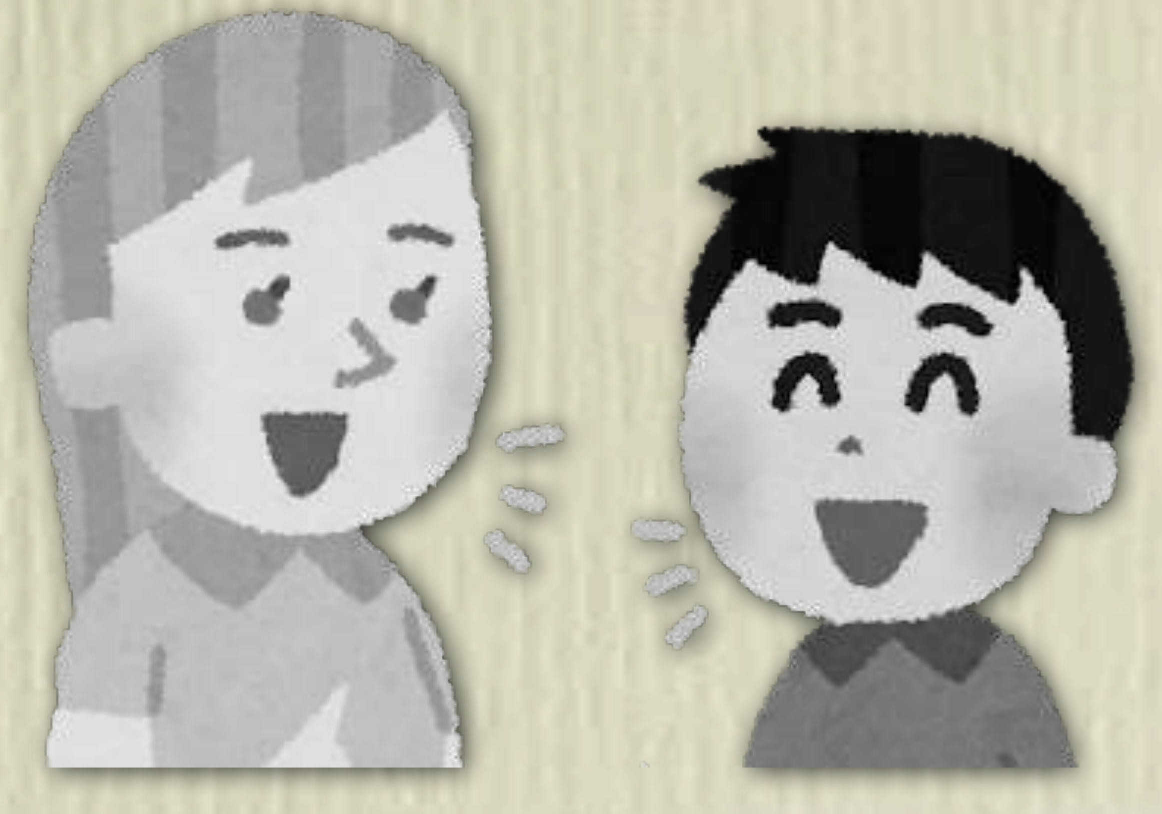
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



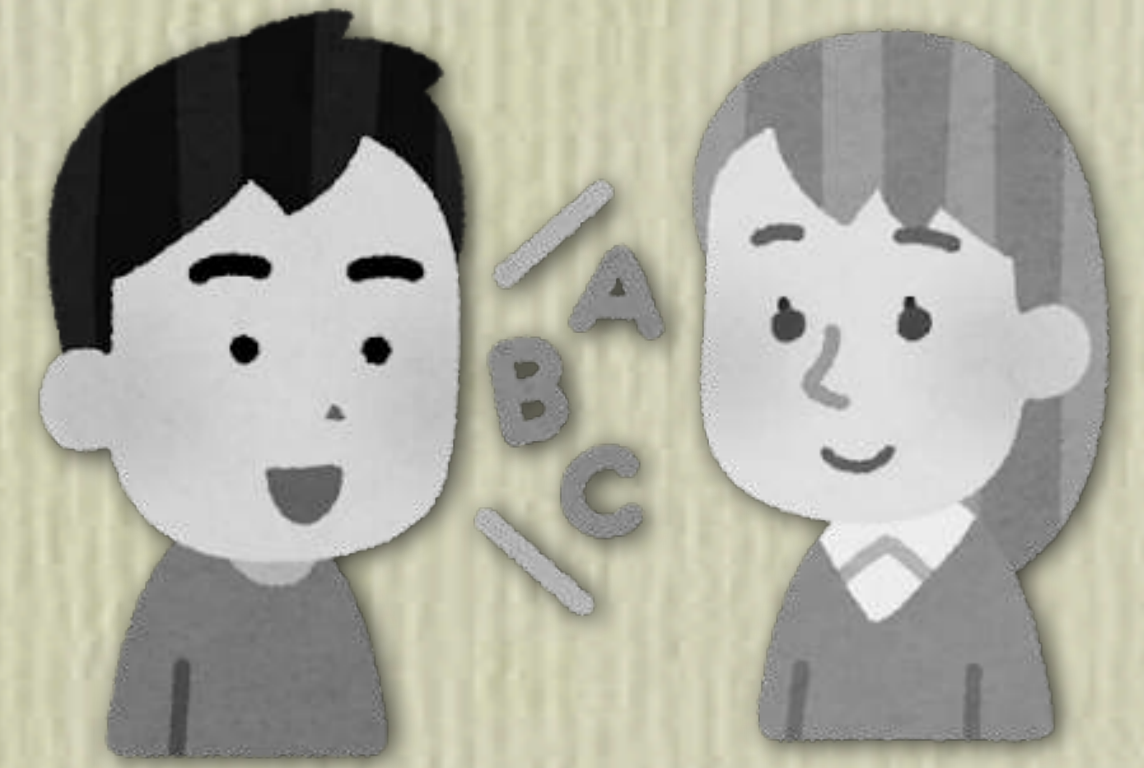
with speech recognition technologies

with new speech technologies being developed in our new project

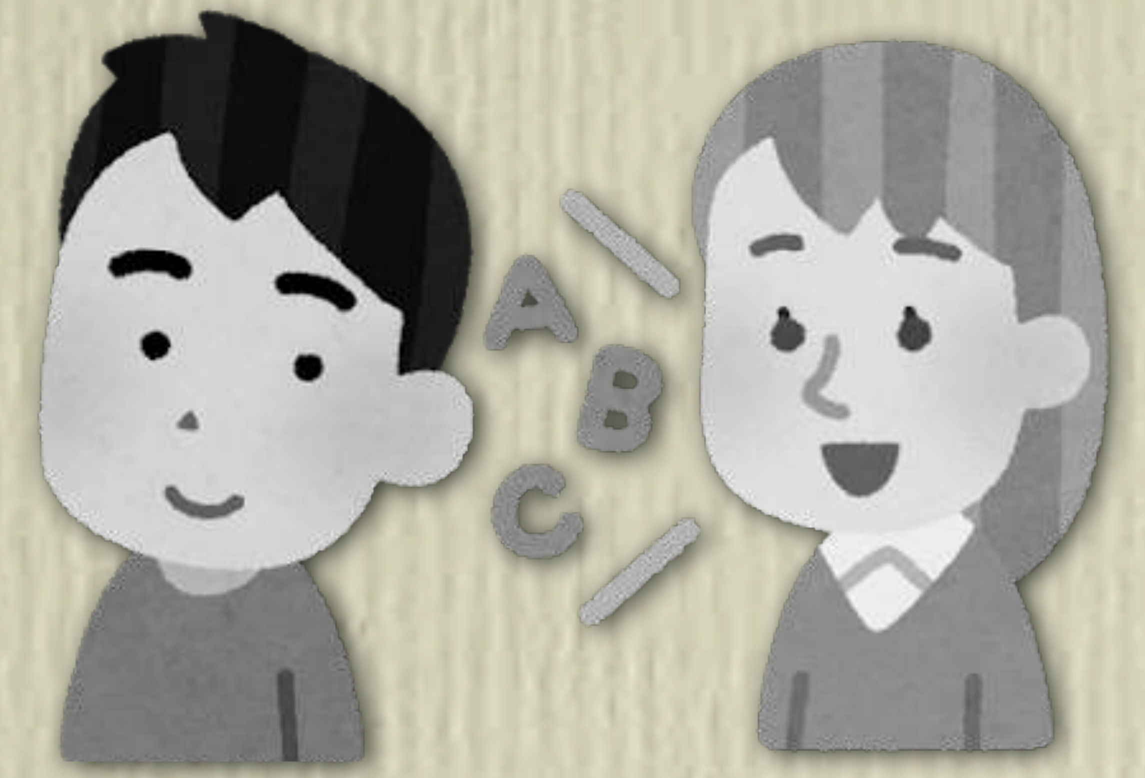
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

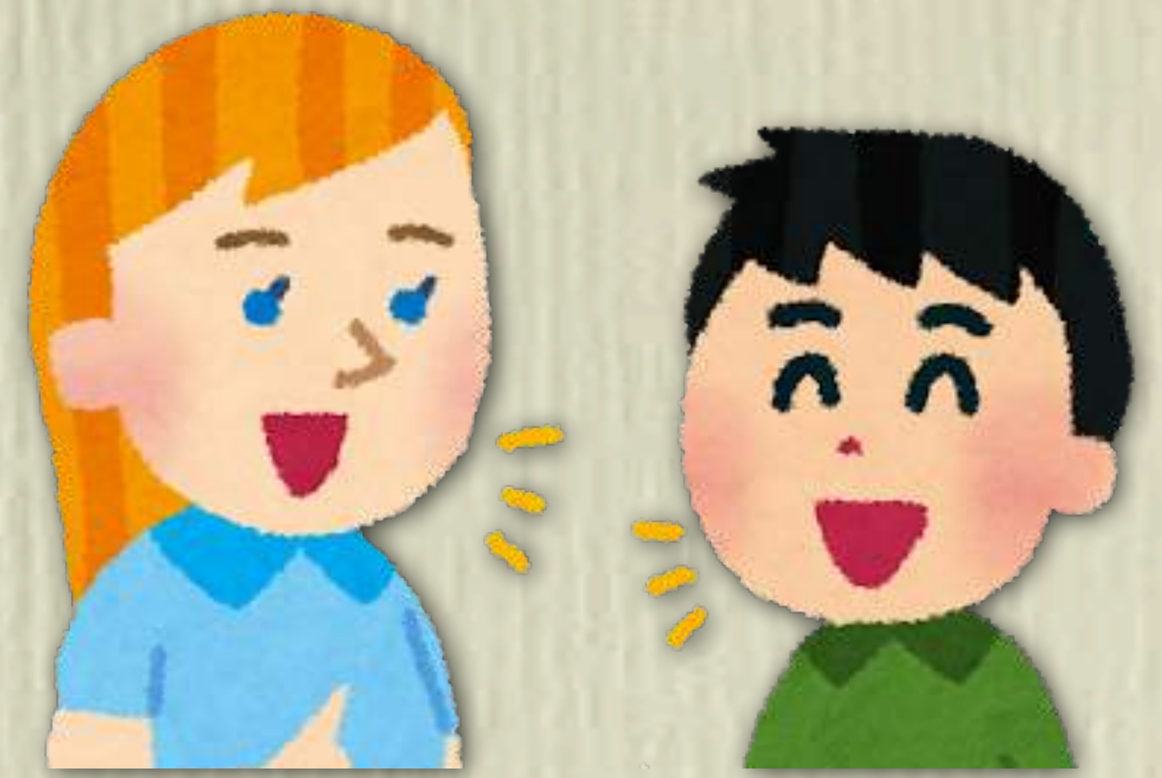
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies

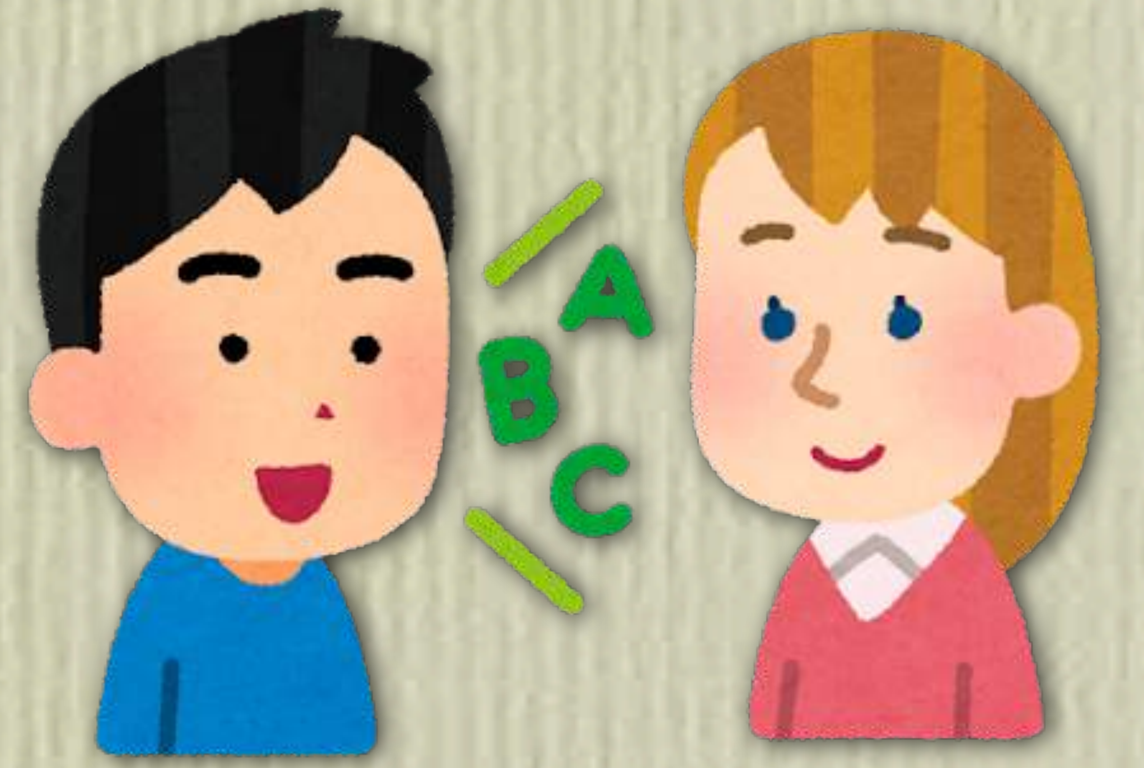


with speech recognition technologies

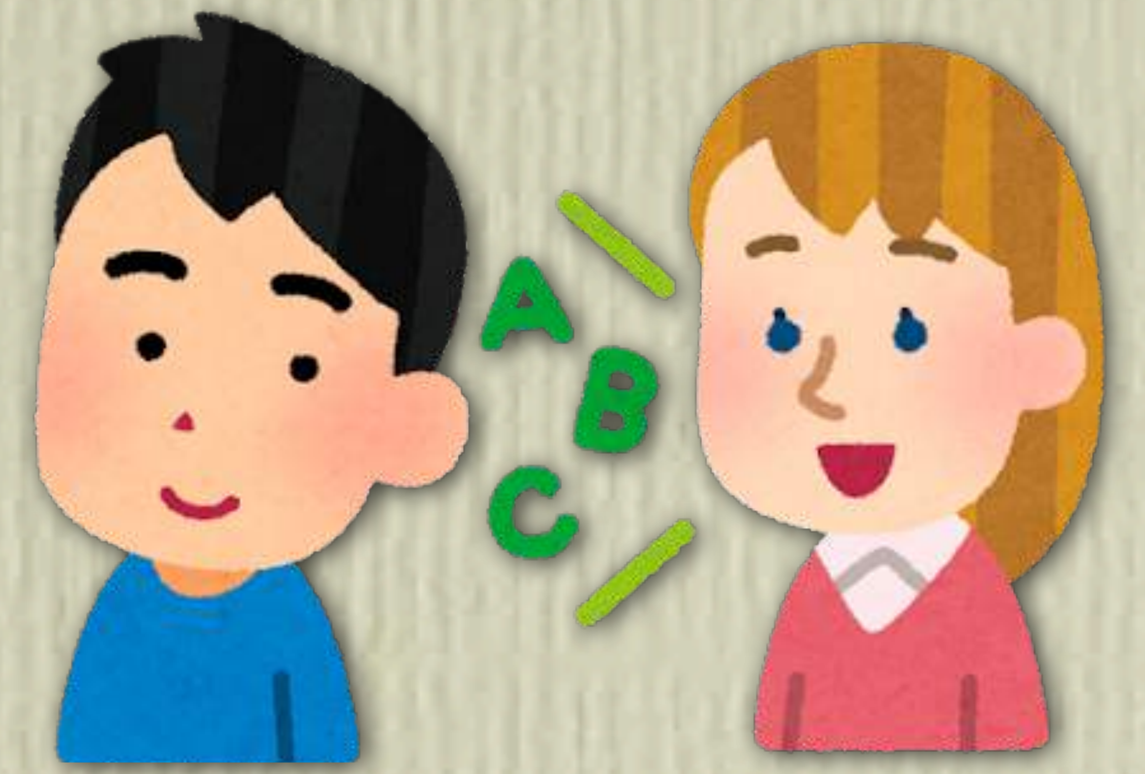
with new speech technologies being developed in our new project

Conversation is a multi-task speech activity.

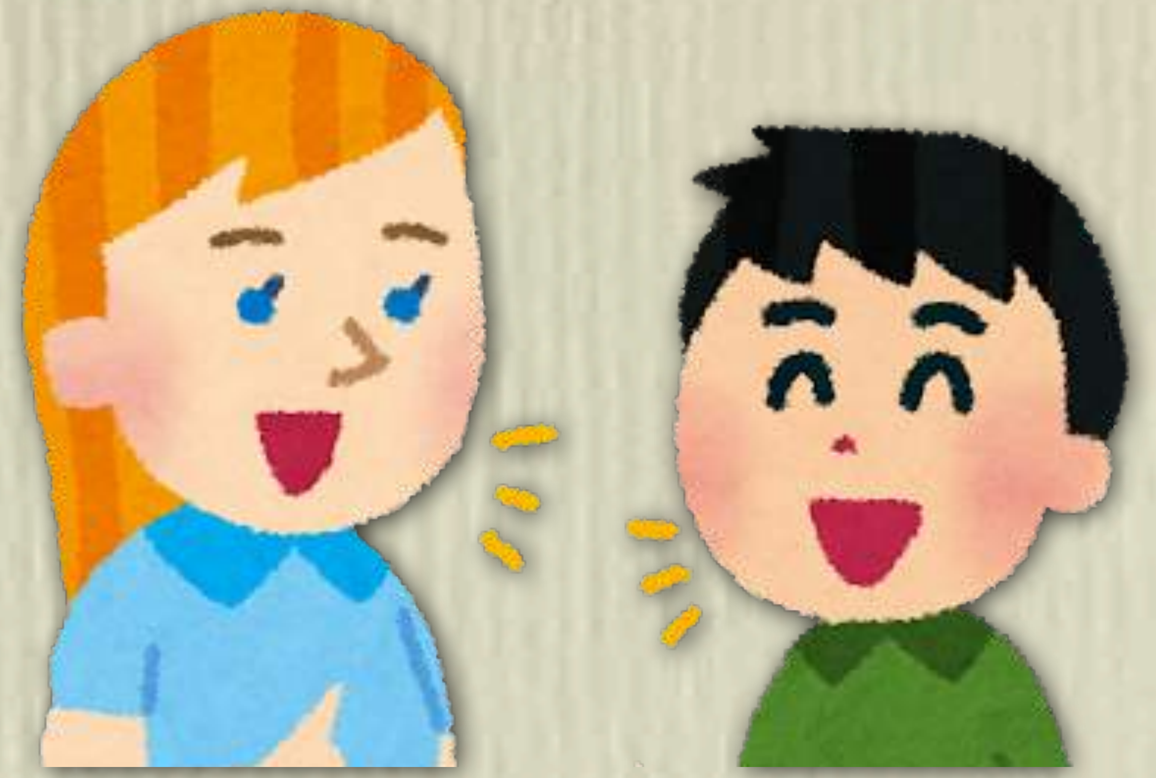
- Listening, understanding, and speaking running almost together



+

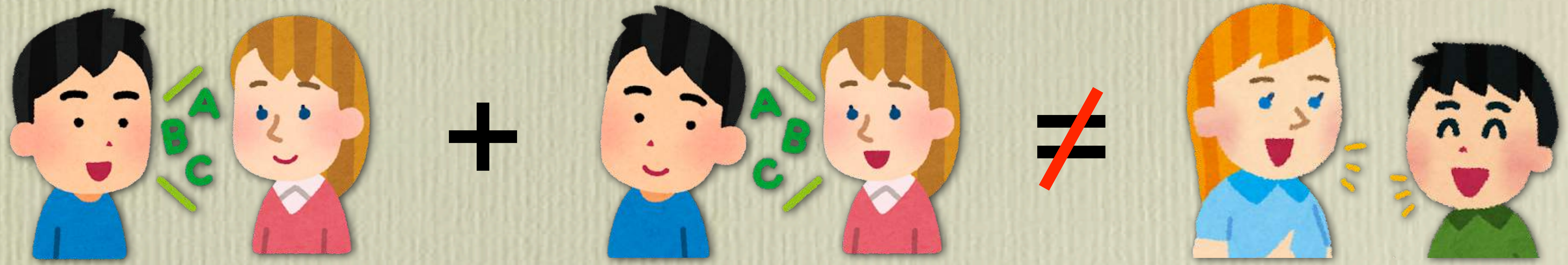


$\neq$

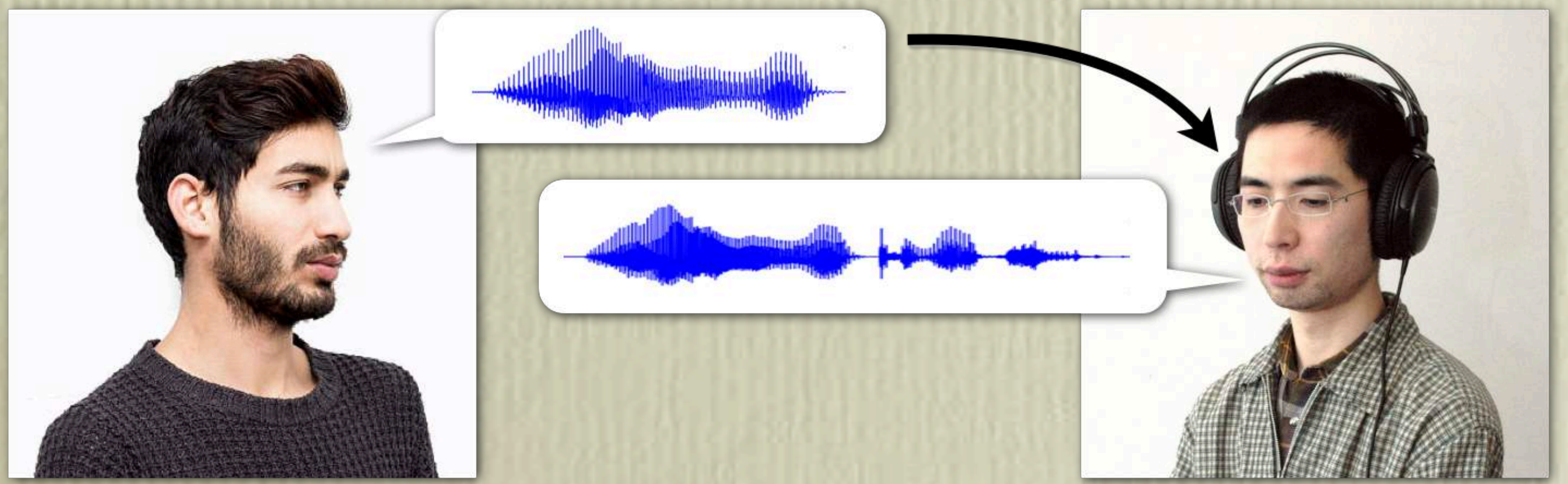


Conversation is a multi-task speech activity.

- **Listening, understanding, and speaking running almost together**



- **Shadowing is a multi-task speech training.**
  - A special form of listen-and-repeat practice, with as short delay as possible

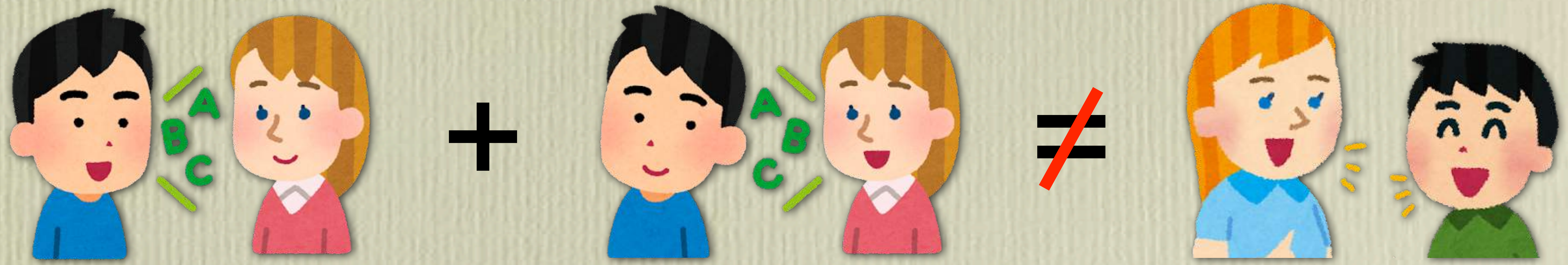


native

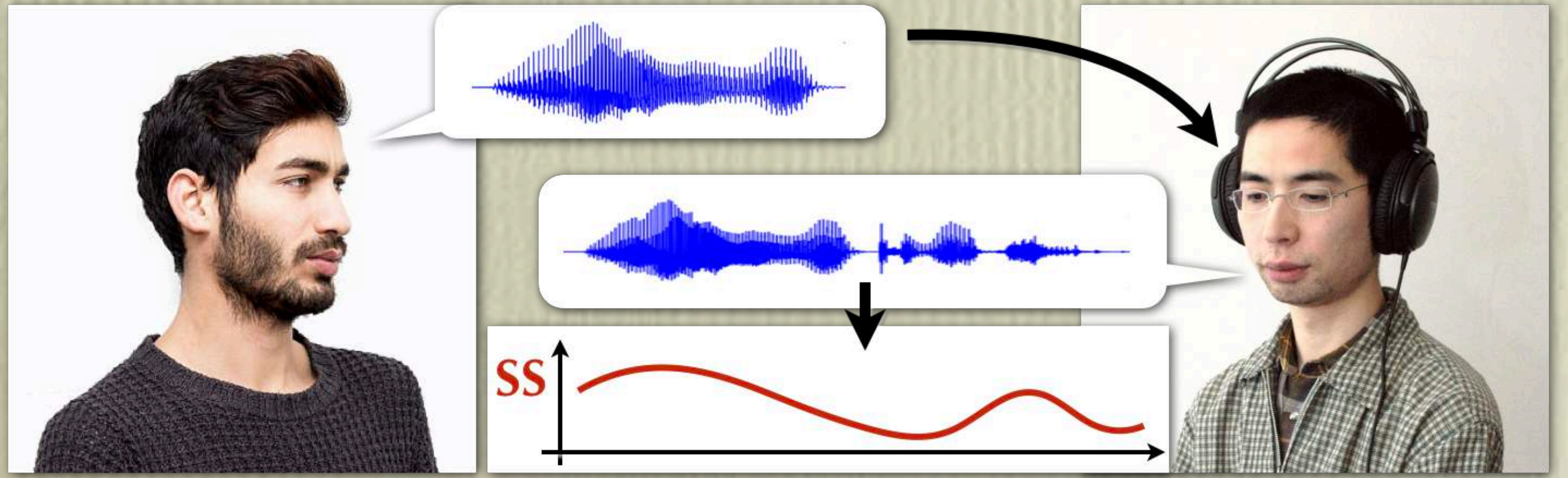
learner

Conversation is a multi-task speech activity.

- **Listening, understanding, and speaking running almost together**



- **Shadowing is a multi-task speech training.**
  - A special form of listen-and-repeat practice, with as short delay as possible



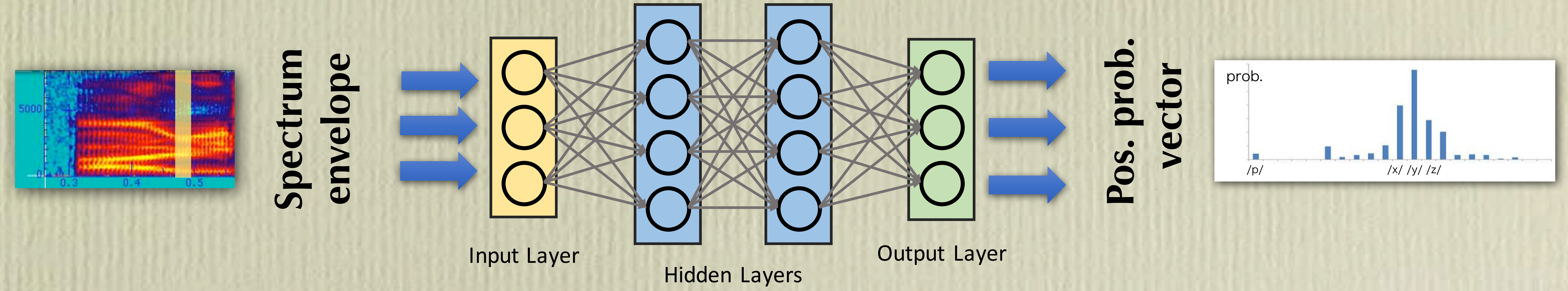
native ss=smoothness of shadow learner

with ASR

Spectrogram is converted to posteriogram

Phoneme posterior probabilities calculated by DNN

- A front-end module of current ASR systems.



DNN processing can be viewed as strong abstraction.

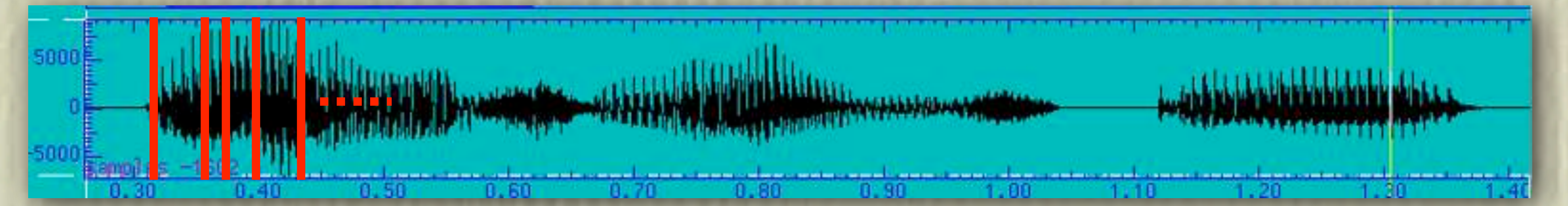
- Spectrogram is acoustic representation, including extra-linguistic features.
- Posteriogram is phonetic/phonemic representation, suppressing those features.

Two methods of DNN-based assessment of shadowing utterances

- DNN-GOP and DNN-DTW

DNN-based calculation of GOP

- **GOP = Goodness Of Pronunciation = phoneme-based posteriors**



+phonemes intended
by the model speaker

time

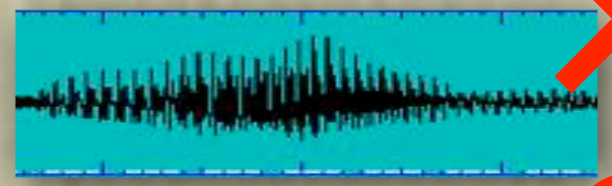
Frame	Phoneme
1	a
2	a
3	u
…	…
1232	sil

phoneme

time

Frame	sil	a	i	u	…
1	0.01	$\boxed{0.8}$	0.1	0.02	…
2	0.01	$\boxed{0.7}$	0.1	0.1	…
3	0.01	0.5	0	$\boxed{0.4}$	…
…	…	…	…	…	…
1232	$\boxed{0.9}$	0	0.01	0	…

M



text (phonemes)

L

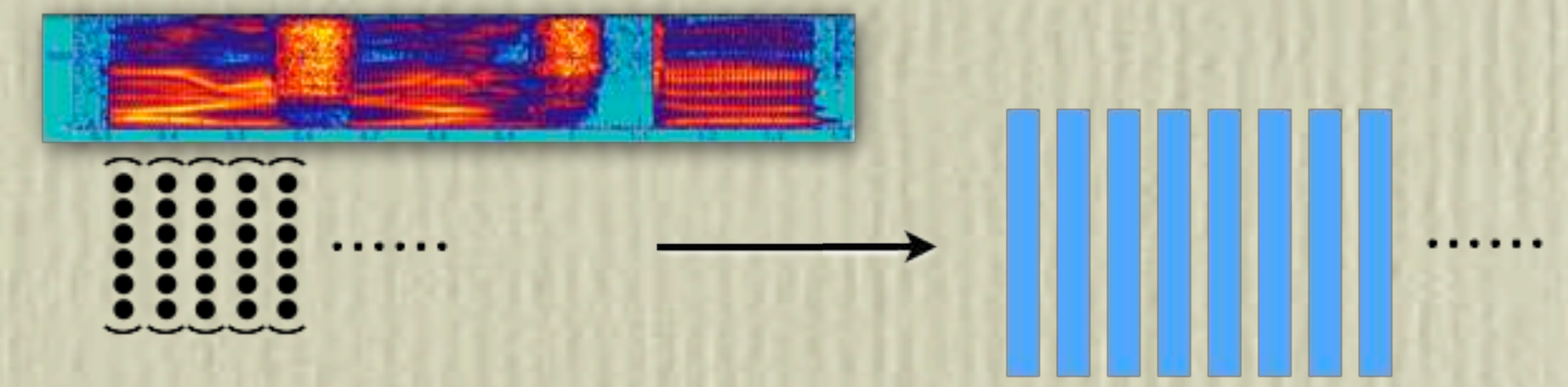
$$\mathrm{GOP} = \frac{0.8 + 0.7 + 0.4 + \cdots + 0.9}{1232} = 0.63$$

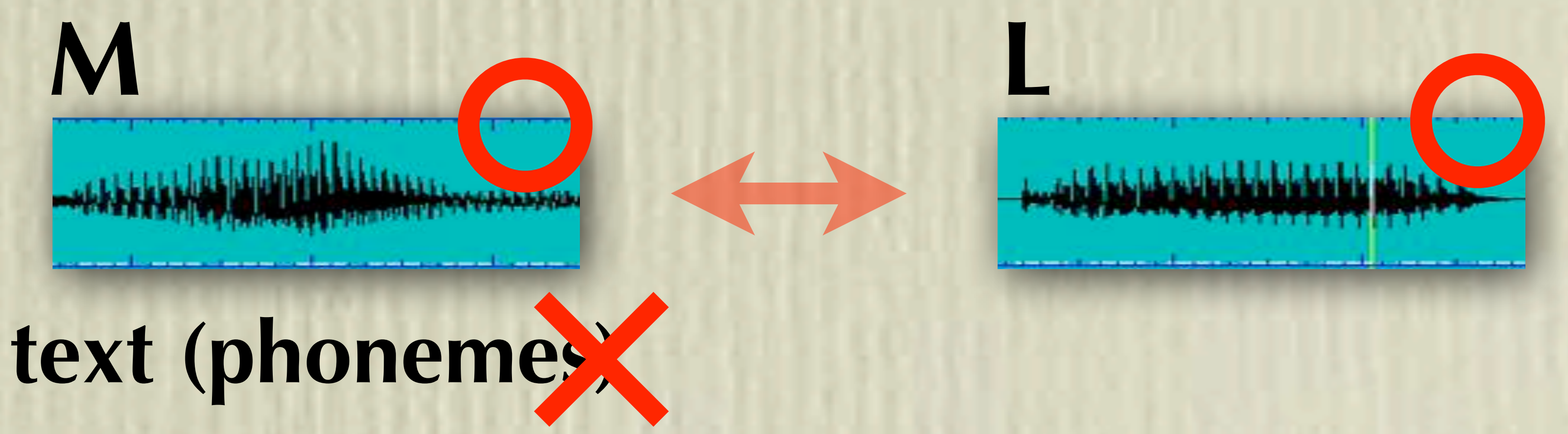
DNN-GOP

Another method for utterance comparison

The two utterances are compared directly.

- Dynamic Time Warping (DTW)
  - Alignment of two sequences of different length
- The two utterances are converted into prob. vector sequences.
  - Spectrum vectors are sensitive to age, gender, etc.





- DTW-based comparison between the two

DNN-DTW

Correlations bet. human scores and machine scores

Sentence-based and speaker-based rating scores

- Sentence-based scores are averaged to obtain speaker-based scores.

Regression model to predict human scores

- Variants of DNN-GOP and some other features are used for regression.

Table 2. Feature-based correlations with teachers' scores

features	P	S	C	P+S+C
bGOP [16]	0.74	0.83	0.71	0.83
pGOP	0.79	0.84	0.78	0.88
vGOP	0.70	0.83	0.70	0.81
cGOP	0.79	0.82	0.78	0.87
v1GOP	0.63	0.78	0.64	0.75
v2GOP	0.42	0.41	0.43	0.46
v0GOP	0.71	0.75	0.78	0.78
DNN-DTW	-0.66	-0.84	-0.69	-0.80
RS	-0.34	-0.21	-0.29	-0.30
WRR	0.79	0.81	0.71	0.84

Table 3. Model-based correlations in a speaker level

models	P	S	C	P+S+C
bGOP [16]	0.74	0.83	0.71	0.83
Lasso	0.84	0.89	0.76	0.90
SVR	0.85	0.89	0.83	0.89
Random Forest	0.77	0.84	0.79	0.86
inter-rater	0.77	0.69	0.86	0.87

Table 4. Model-based correlations in a sentence level

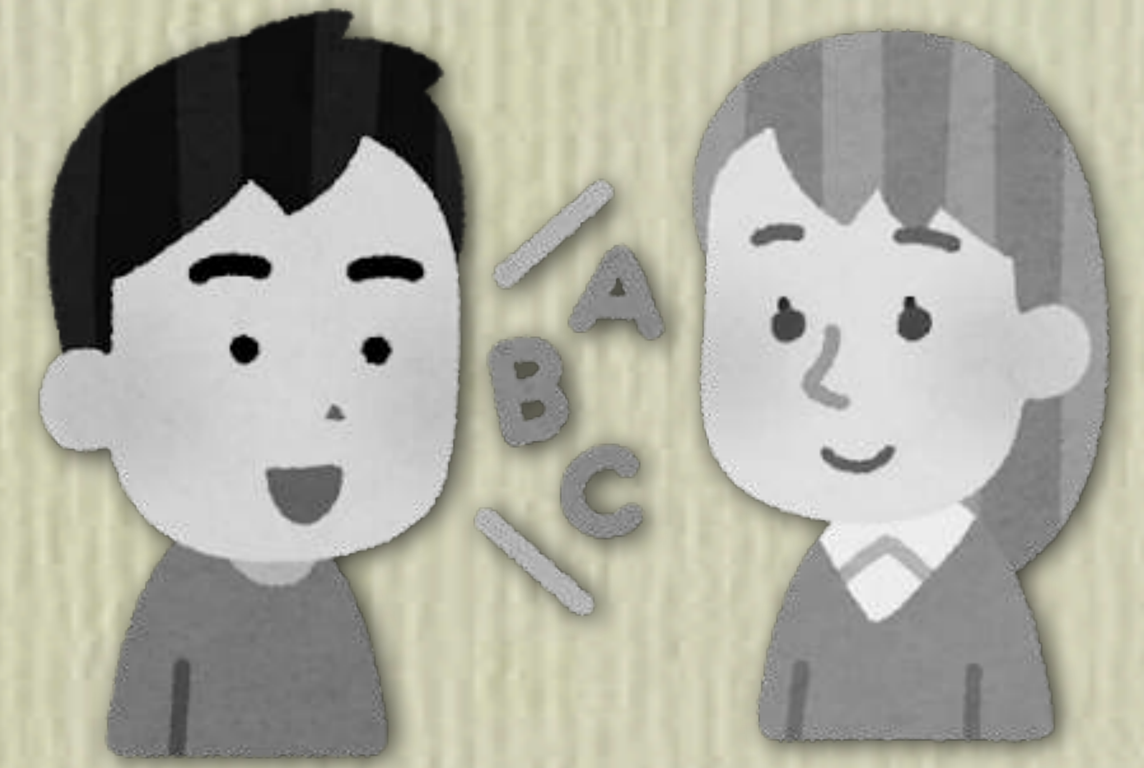
models	P	S	C	P+S+C
Lasso	0.68	0.73	0.65	0.77
SVR	0.70	0.73	0.68	0.78
Random Forest	0.67	0.68	0.61	0.74
inter-rater	0.58	0.54	0.74	0.75

S. Kabashima, et al., "DNN-based scoring of language learners' proficiency using learners' shadowings and native listeners' responsive shadowings," Proc. Spoken Language Technology, 2018

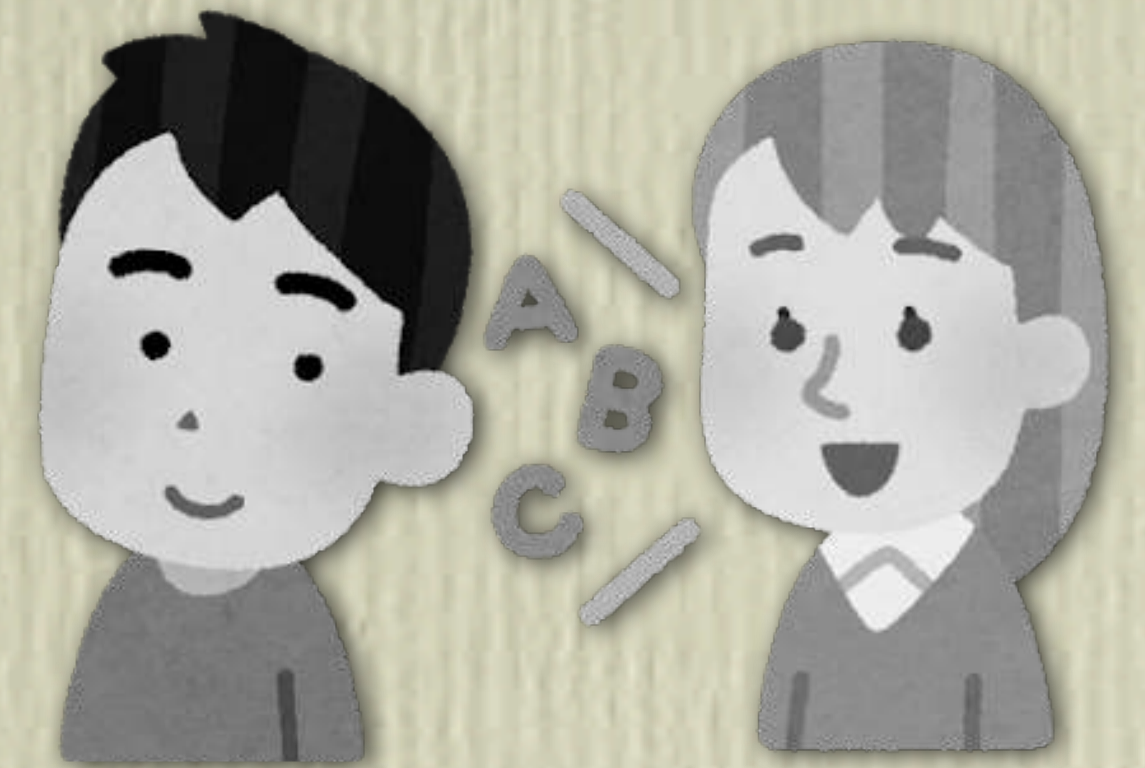
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

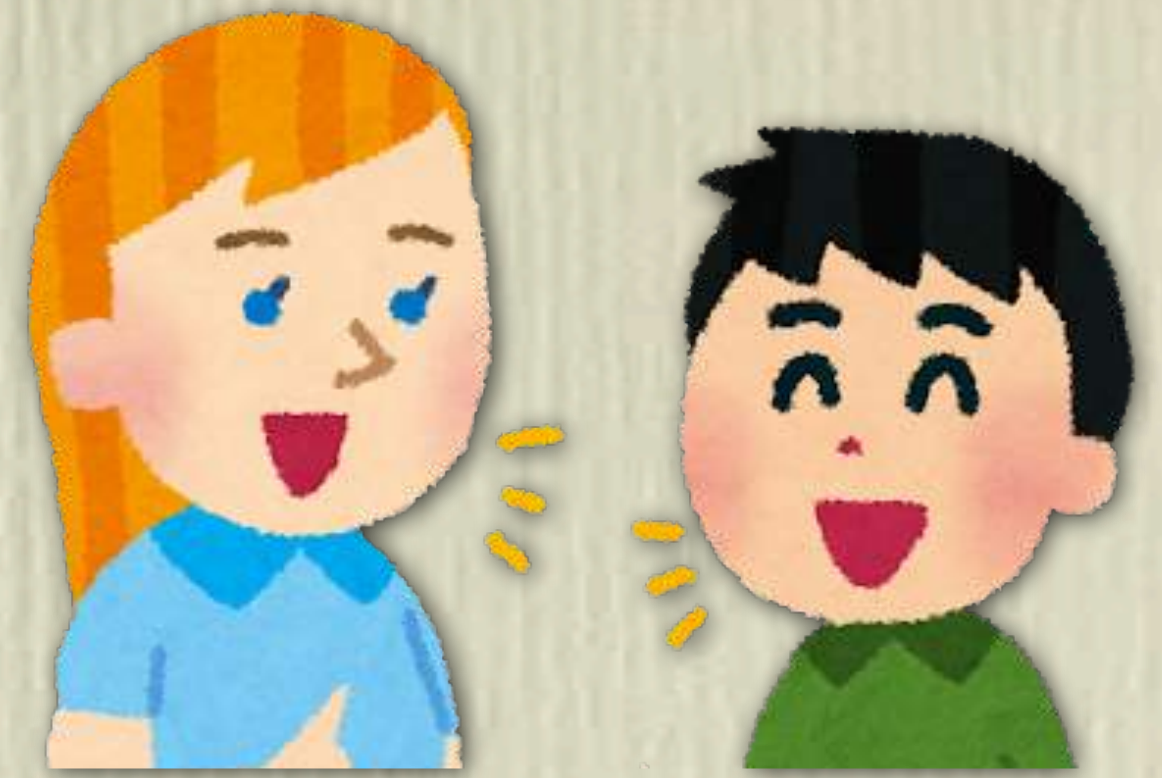
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



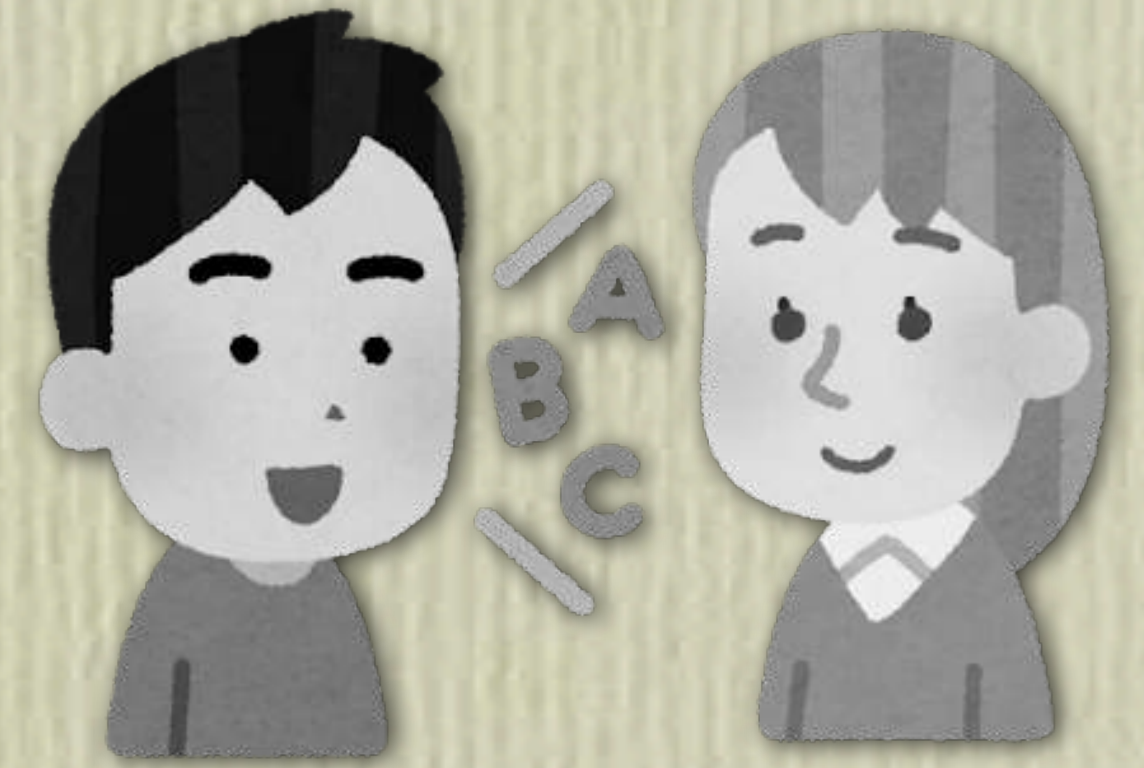
with speech recognition technologies ✔

with new speech technologies being developed in our new project

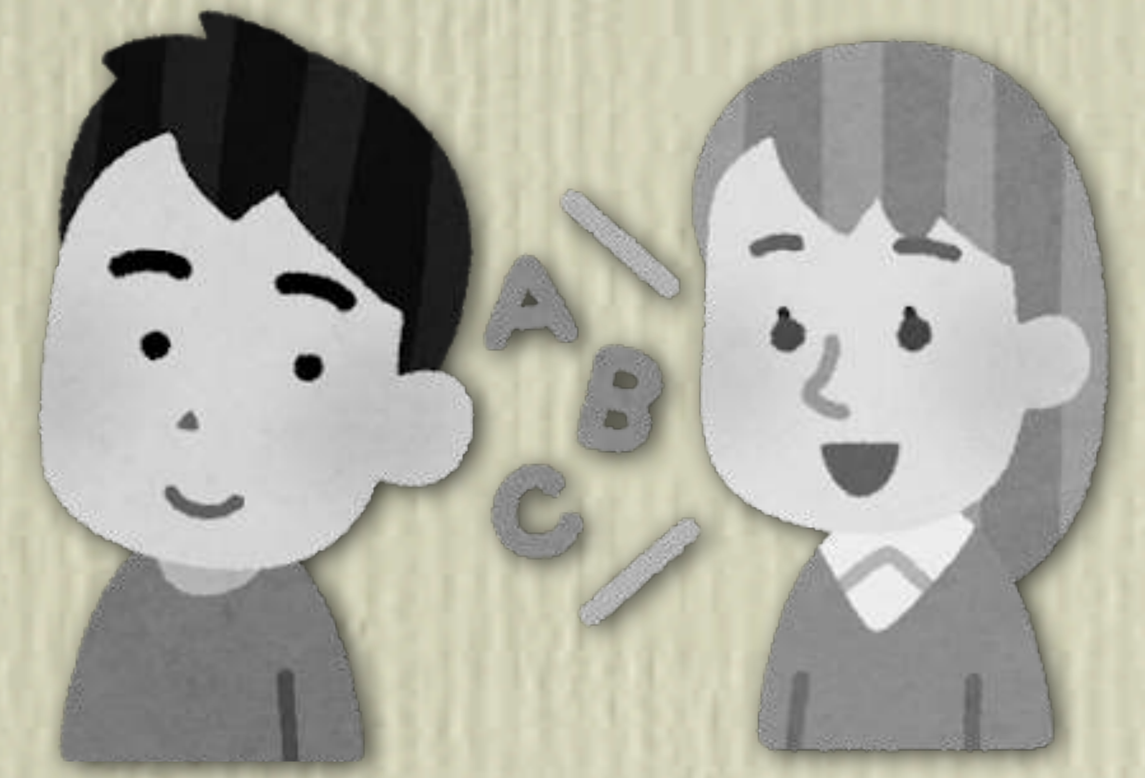
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

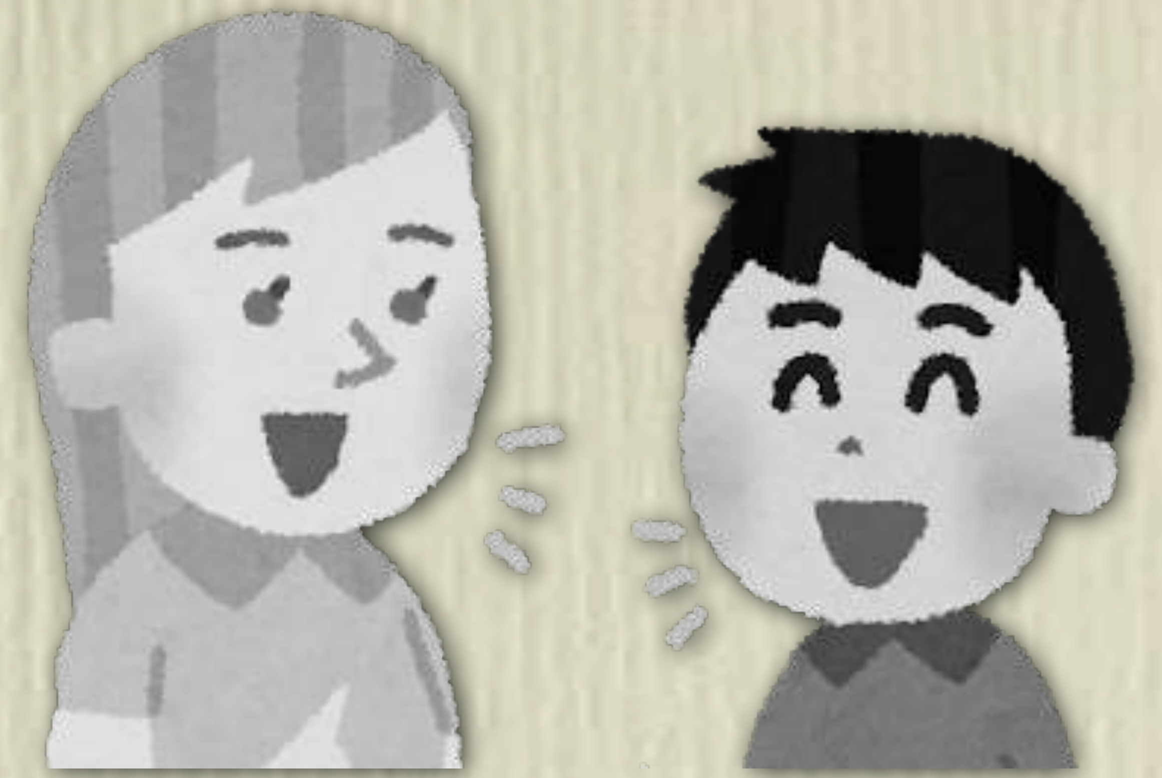
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies

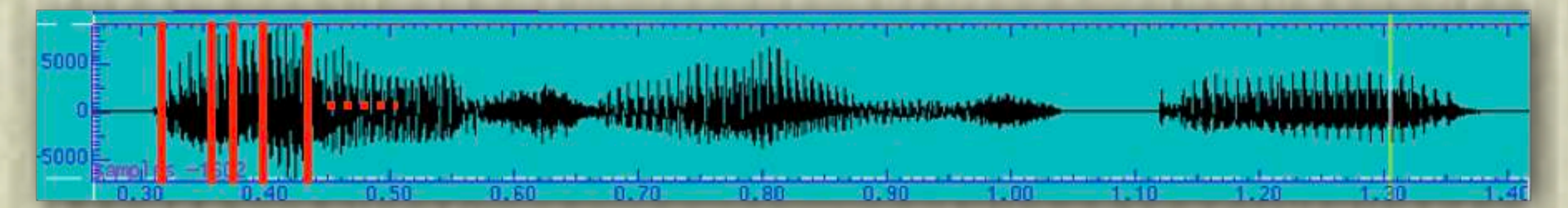


with speech recognition technologies

with new speech technologies being developed in our new project ✓

DNN-GOP and DNN-DTW

- **DNN-GOP = comparison bet. an L2 utterance and native models**
- **DNN-DTW = comparison bet. an L2 utterance and its native version**



+phonemes intended by the model speaker

time ↓

Frame	Phoneme
1	a
2	a
3	u
...	...
1232	sil

phoneme →

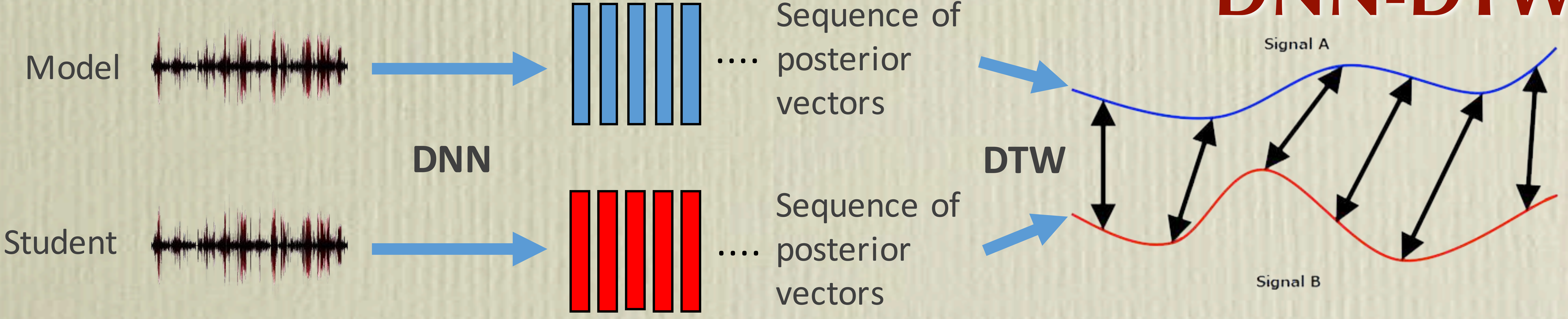
time ↓

Frame	sil	a	i	u	...
1	0.01	0.8	0.1	0.02	...
2	0.01	0.7	0.1	0.1	...
3	0.01	0.5	0	0.4	...
...	...	...	...	...	...
1232	0.9	0	0.01	0	...

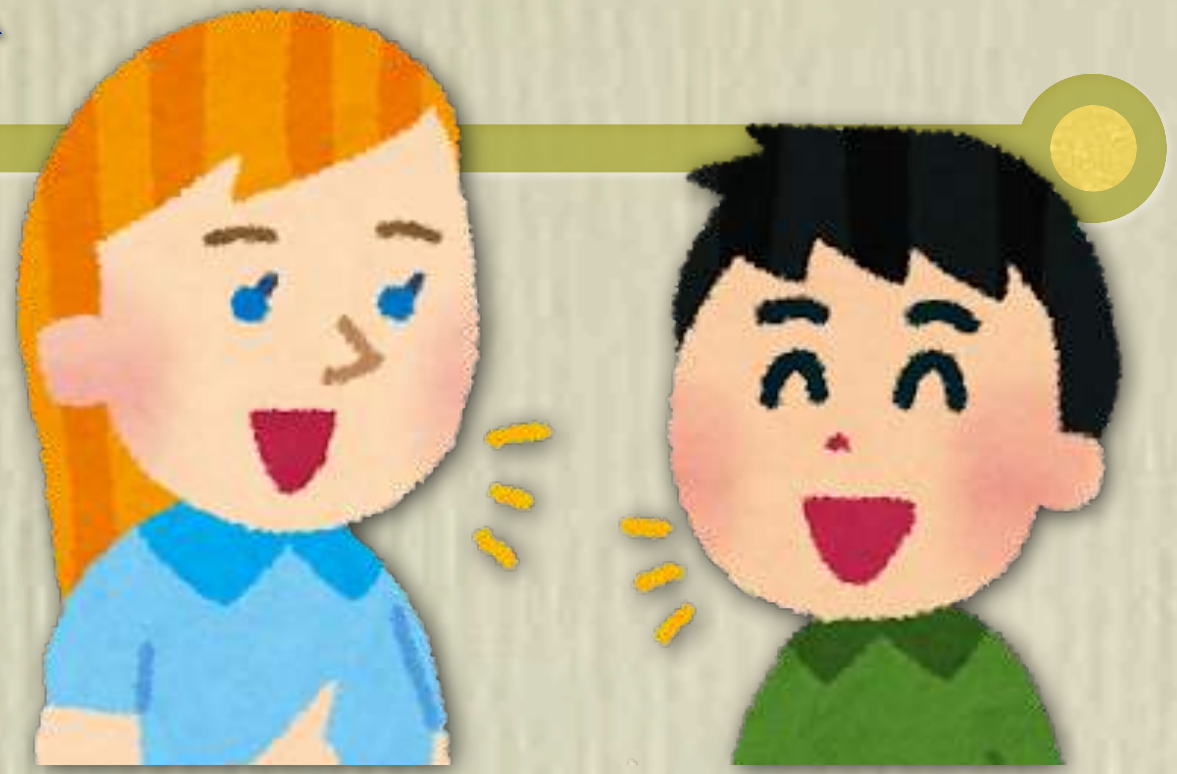
Native-likeness

DNN-GOP

DNN-DTW



Accented pronunciations are OK



- **if they are *intelligible or comprehensible enough*.**
  - What is the definition of intelligible/comprehensible enough prons?
- **Interesting experimental facts**
  - AE and JE were presented to and transcribed by American listeners with no exposure to JE.
  - Some extreme samples of transcriptions
    - "The misquote was retracted with an apology."

- # the misquote was retracted with an apology
- # the misquote was retracted with an apology
- # the misquote was retracted with an apology #
- # the misquote was retracted with an apology #
- the misquote quote was retracted with an apology
- the misquote was [S>] attr(acted)- [
- the misquote was attra(cted)- was retracted with an apology
- the misquote was attract was retracted with an apology
- the misquote was attracted with an apology
- the misquote was attracted with an apology

- i don't know
- sammy's coat was instructed
- constructed
- distracted @
- was instructed with an apology
- @ by an apology
- something @ without apology
- @ was something
- instructed with an apology
- an apology

    - Utterances of a learner are more intelligible to him/her than native utterances

N. Minematsu, et al., “Measurement of objective intelligibility of Japanese accented English using ER h Read by Japanese) database,” Proc. INTERSPEECH, 1481-1484, 2011

What kind of technologies are needed for learners?

"How are my utterances perceived by listeners?"

- "The misquote was retracted with an apology."

• # the misquote was retracted with an apology • # the misquote was retracted with an apology • # the misquote was retracted with an apology # • # the misquote was retracted with an apology # • the misquote quote was retracted with an apology • the misquote was [S>] attr(acted)- [• the misquote was attra(cted)- was retracted with an apology • the misquote was attract was retracted with an apology	• i don't know • sammy's coat was instructed • constructed • distracted @ • was instructed with an apology • @ by an apology • something @ without apology • @ was something

- Utterances of a learner are more intelligible to him/her than native utterances.

$$\underset{w}{\operatorname{argmax}} P_l(w|o)$$

= prediction of what listeners perceived.

$$\underset{w}{\operatorname{argmax}} P_s(w|o)$$

= prediction of what the speaker meant.

N. Minematsu, et al., "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) database," Proc. INTERSPEECH, 1481-1484, 2011

Online observation of listeners' behaviors

Measurement of listening efforts or cognitive load

- Electroencephalogram (EEG) for listening efforts (Song+'18)
- Pupillometry for cognitive load (Govender+'18)

Online observation of listeners' behaviors

Measurement of listening efforts or cognitive load

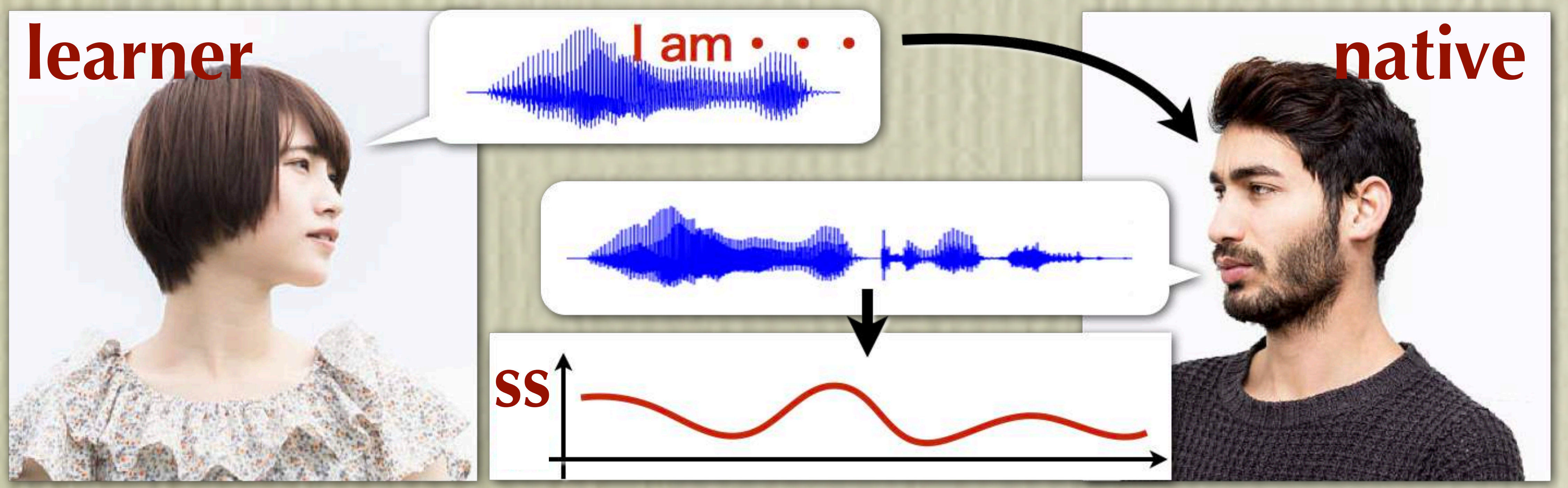
- Electroencephalogram (EEG) for listening efforts (Song+'18)
- Pupillometry for cognitive load (Govender+'18)

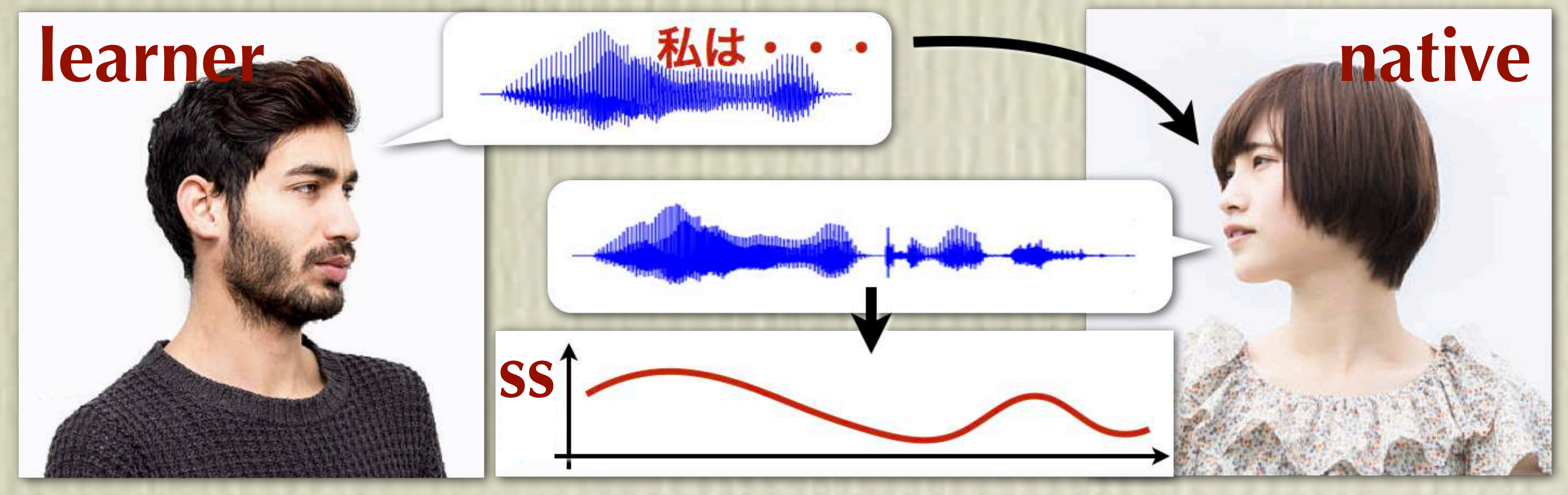
native
listener

non-native
speech

Native listeners' shadowing of learners' utterances

- Shadowing = almost simultaneous reproduction of what a speaker said.
  - Smooth shadowing = easy understanding = low listening efforts / low cognitive load
- DNN-based ASR frontend is used to calculate *shadowability* quantitatively.
  - Listeners' shadowings showed higher correlation to comprehensibility than learners' utterances.

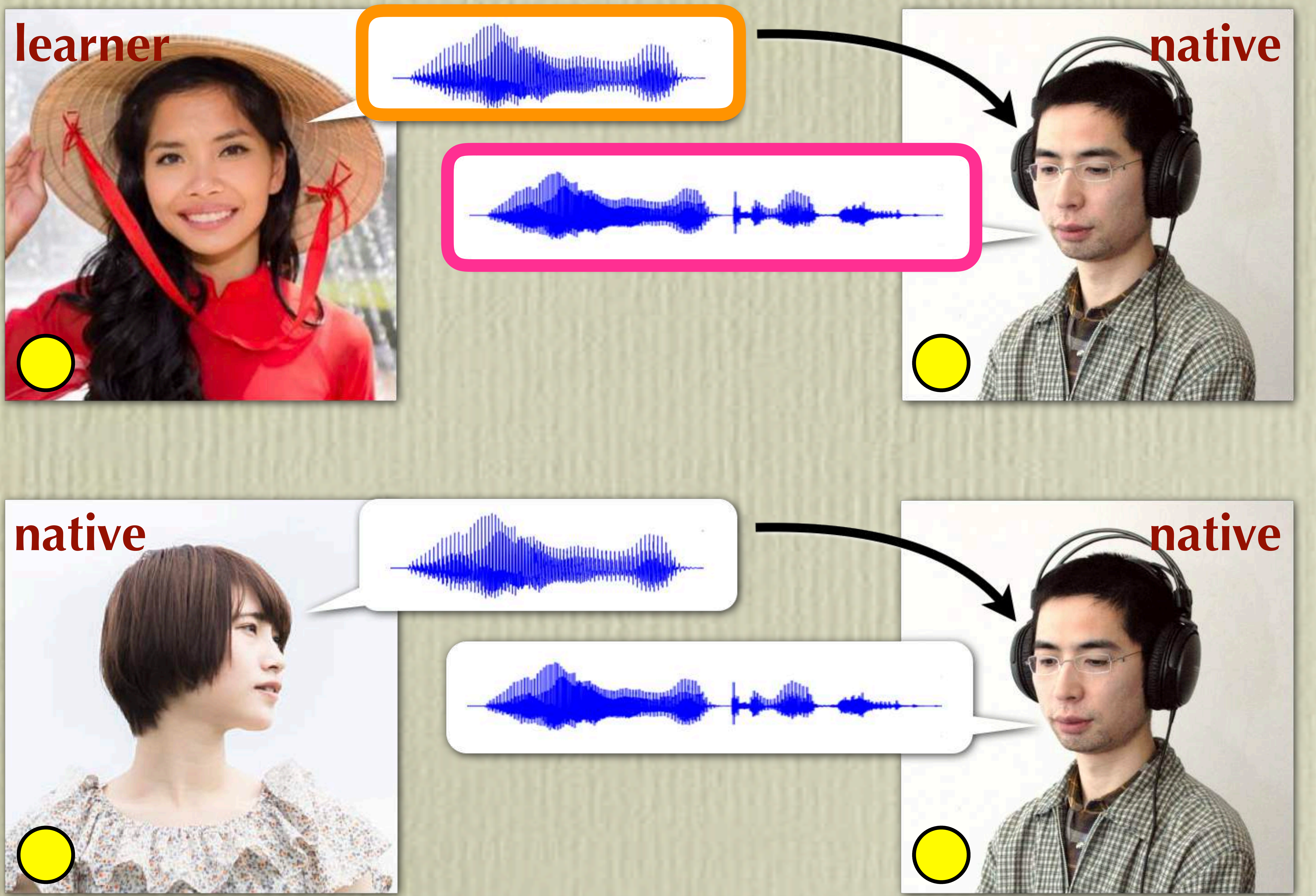


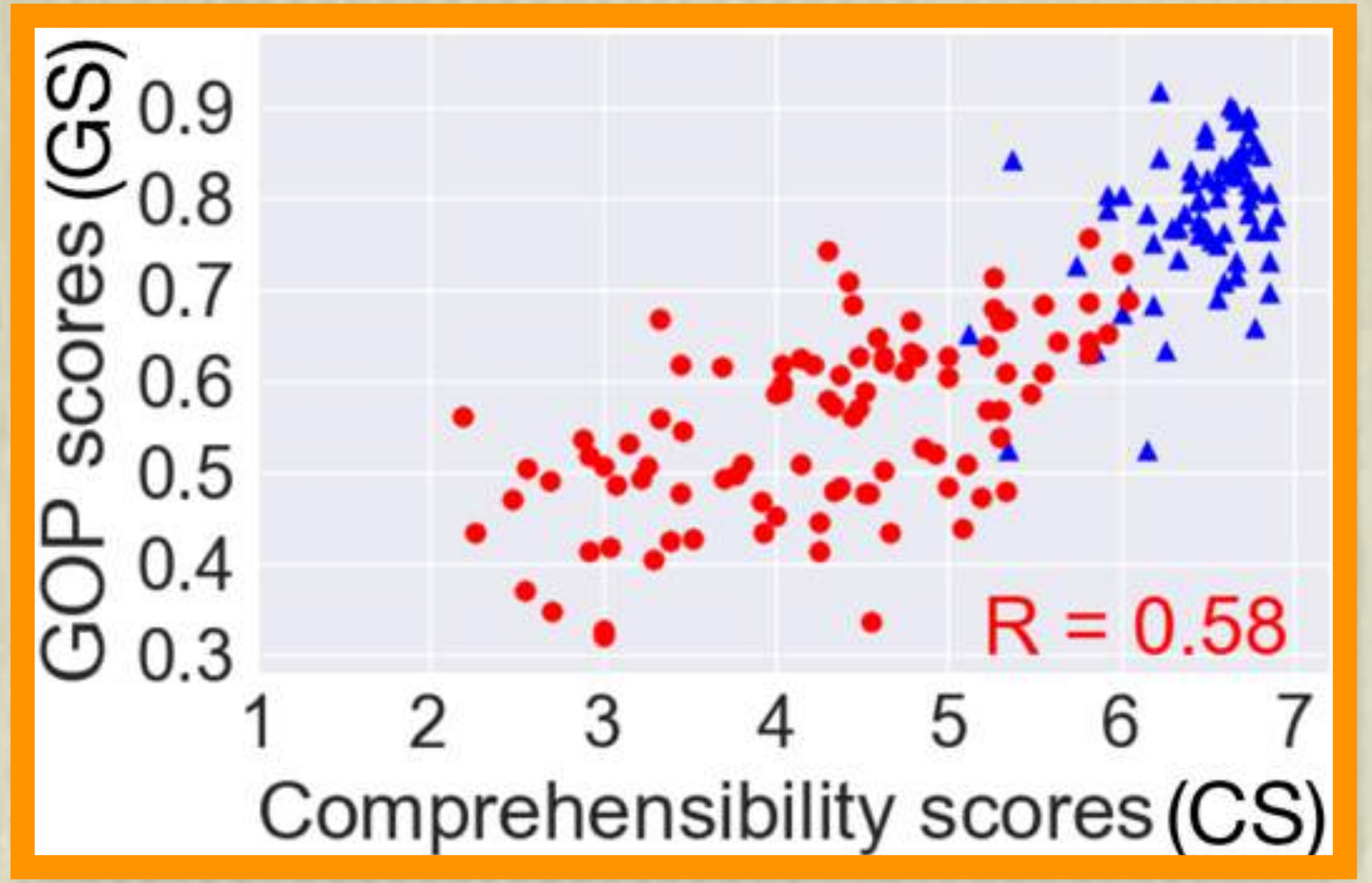


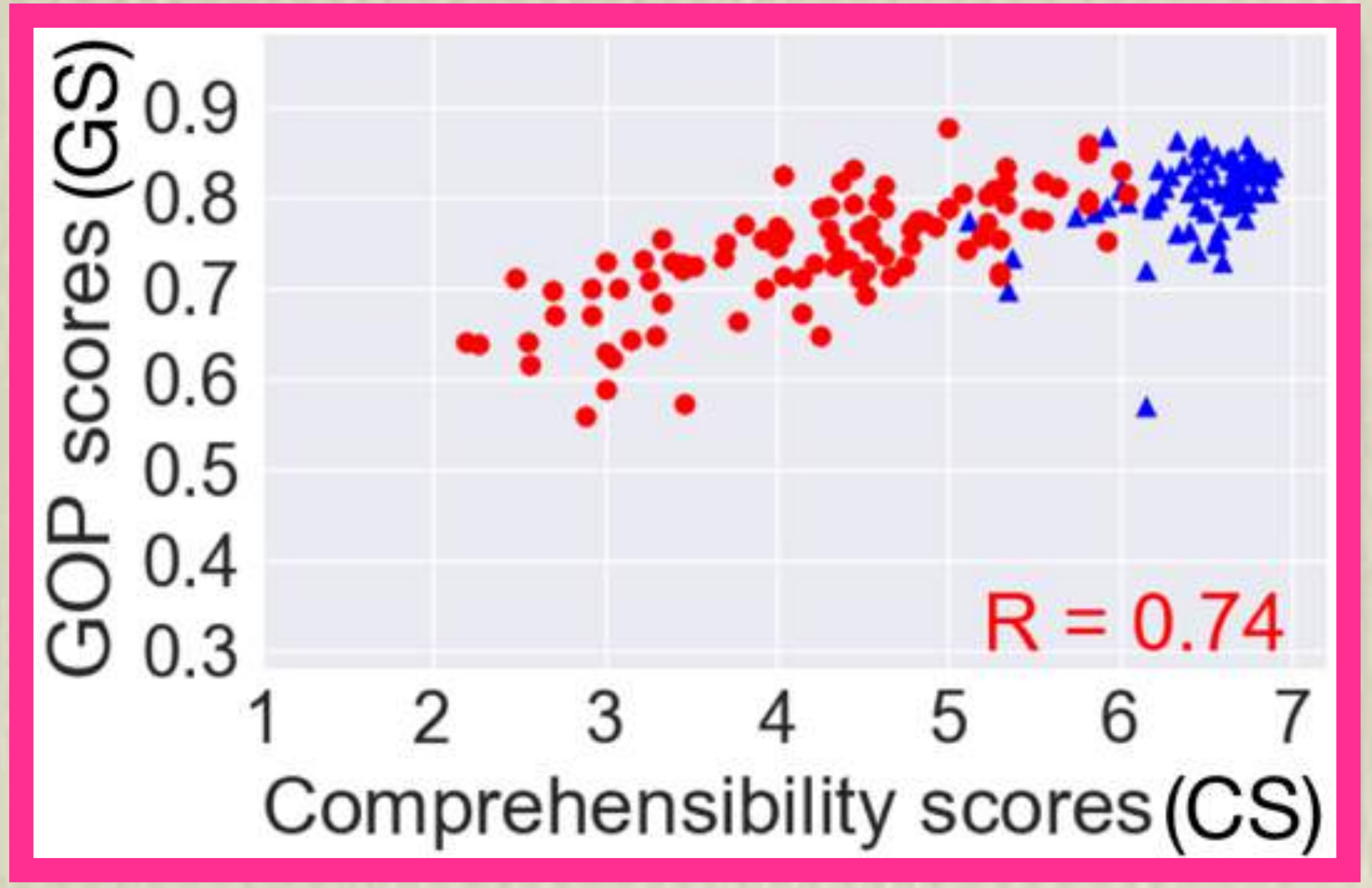
Y. Inoue et al., "A study of objective measurement of comprehensibility through native speakers' responsive shadowing of learners' utterances," Proc. INTERSPEECH, 1651-1655, 2018

Natives' shadowings of non-native and native speech

Japanese speakers shadow Vietnamese and native Japanese (VJ + NJ).





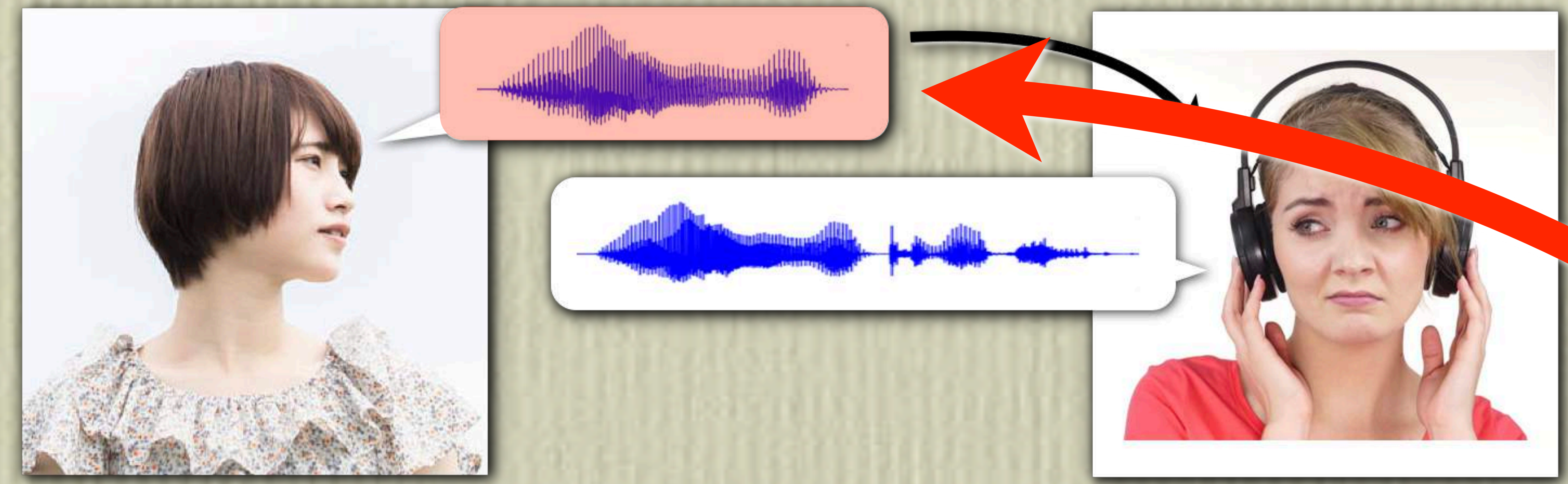


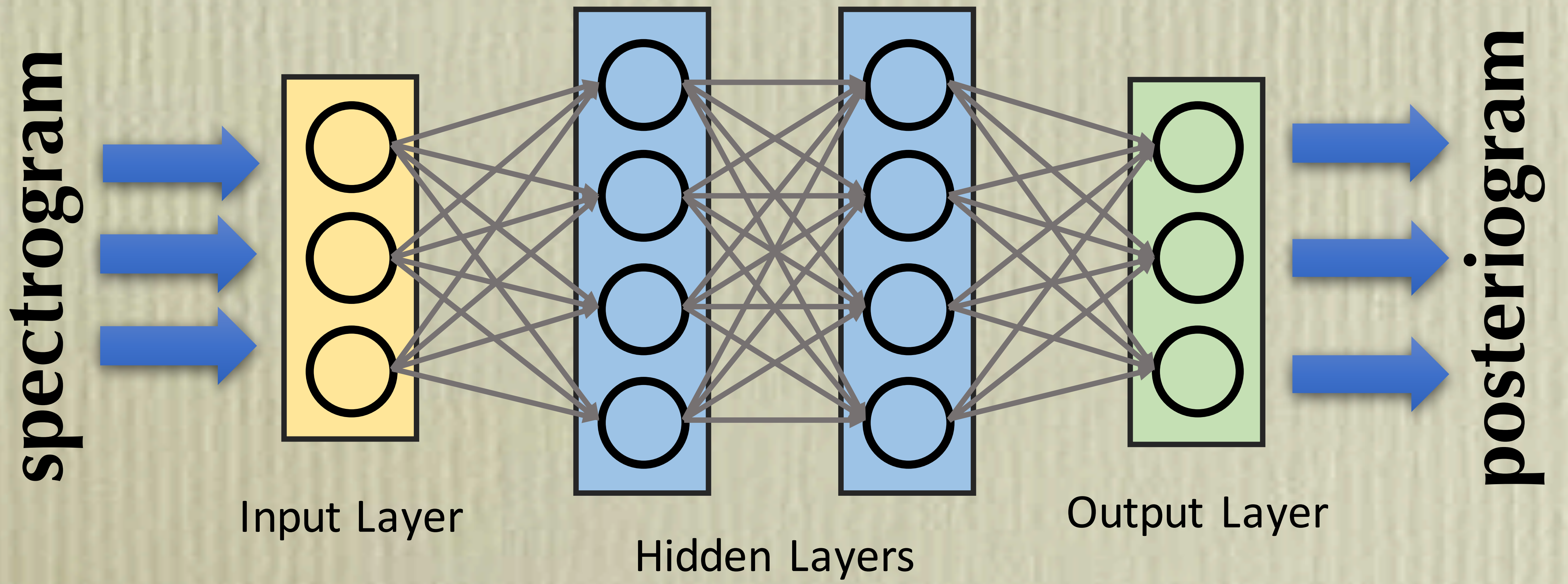
Y. Inoue et al., "A study of objective measurement of comprehensibility through native speakers' responsive shadowing of learners' utterances," Proc. INTERSPEECH, 1651-1655, 2018

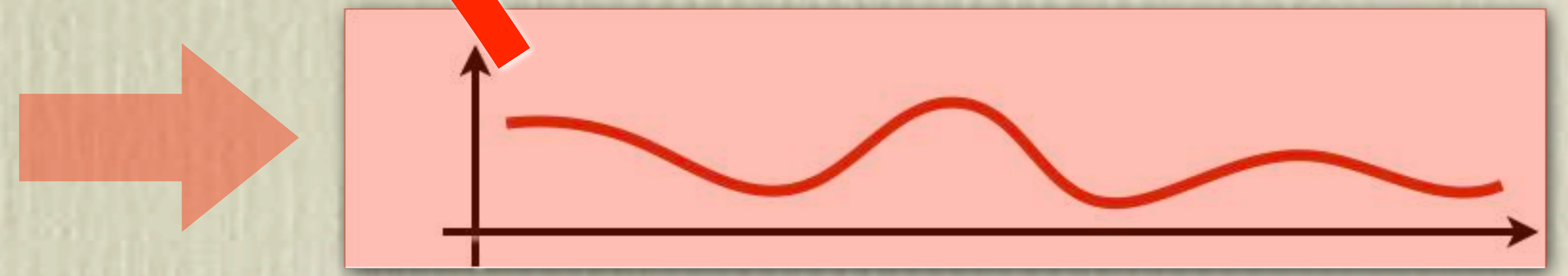
Natives' shadowings as spoken annotations

Native speakers' shadowings can be viewed as spoken annotations.

- Native shadowings + DNN-ASR front end = sequential data of shadowability
- Shadowability sequences can characterize listeners' *dynamic* behaviors of listening.





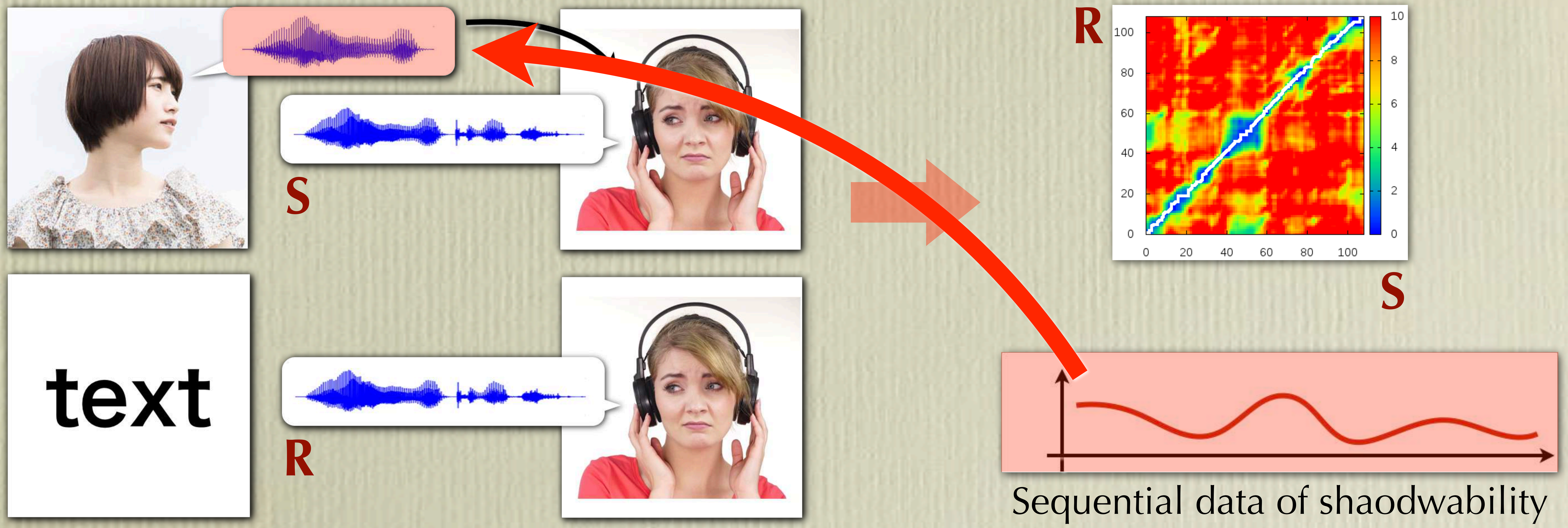


Sequential data of shaodwability

Natives' shadowings as spoken annotations

Native listeners are asked to read after shadowing.

- Shadowing = the least prepared speech, reading = the most prepared speech
- DTW between shadowing and reading gives us more reliable annotations than DNN.



Z. Lin, et. al., "Native listeners' shadowing of non-native utterances as spoken annotation representing comprehensibility of the utterances," Proc. SLaTE, 2019

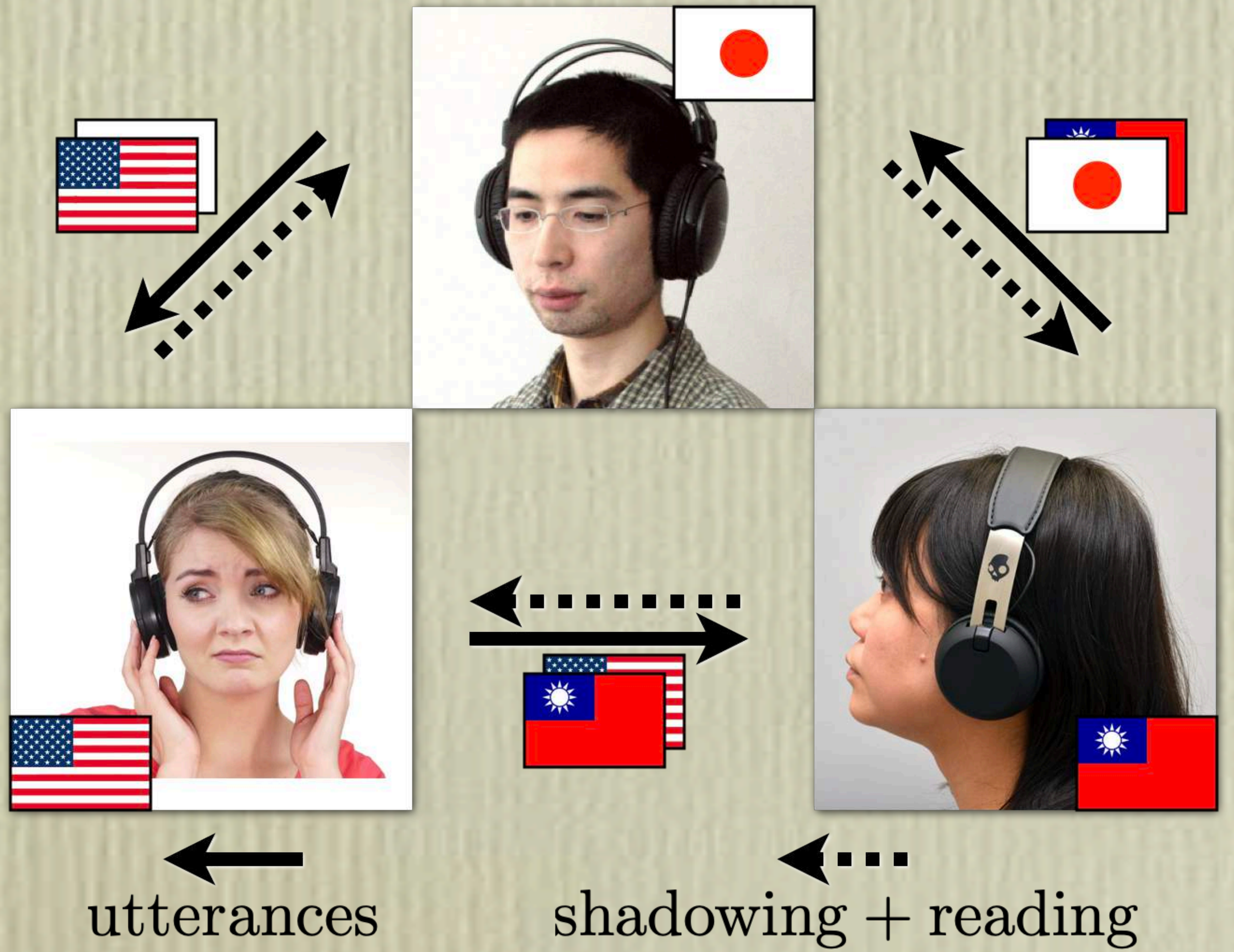
Inter-learner shadowing (ILS) to develop a virtual shadower

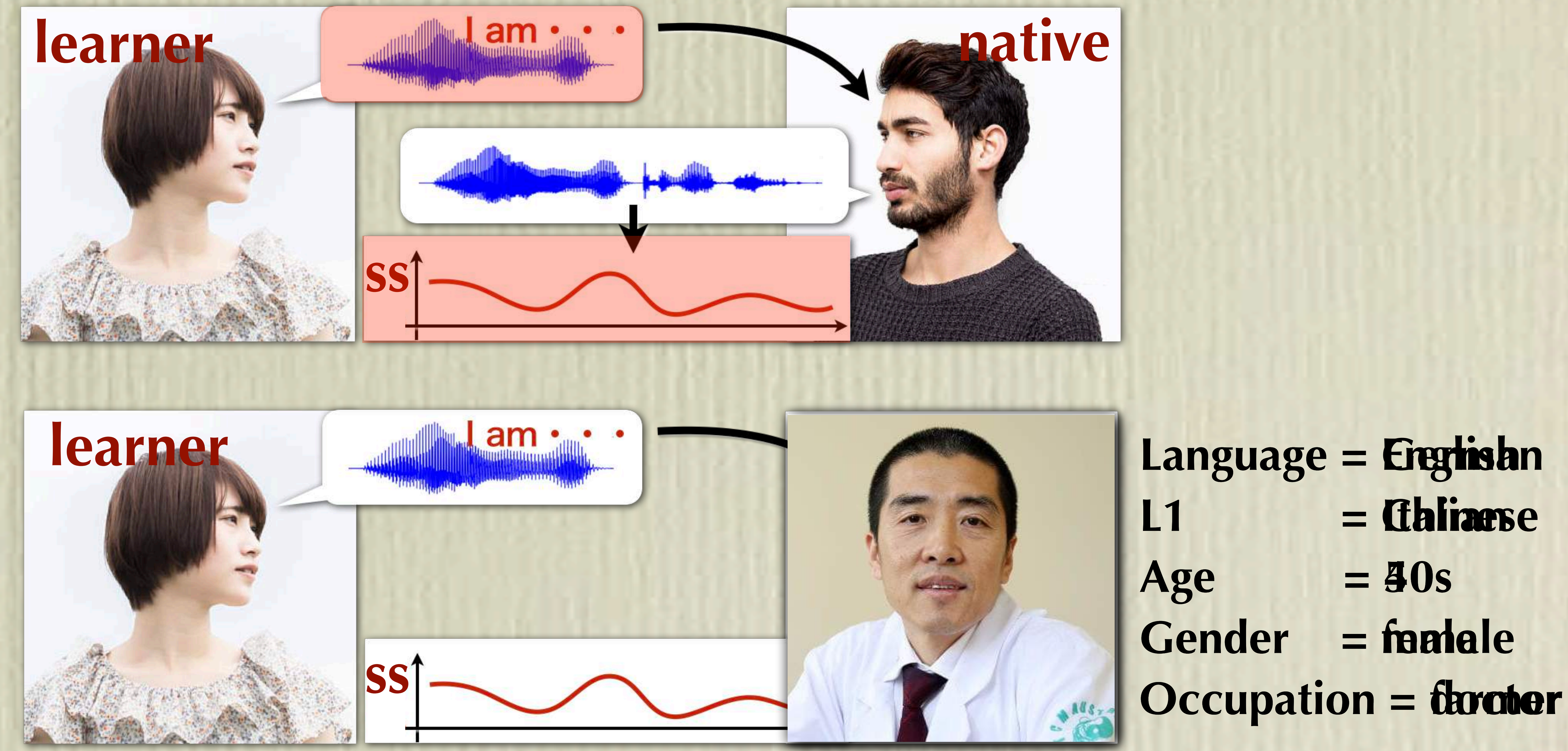
Every learner can be shadowed by shadowing other learners.

- Inter-supportive framework among all the language learners irrespective of languages.

Toward development of a virtual shadower

- *Language-independent* virtual shadower which can simulate various listener profiles.





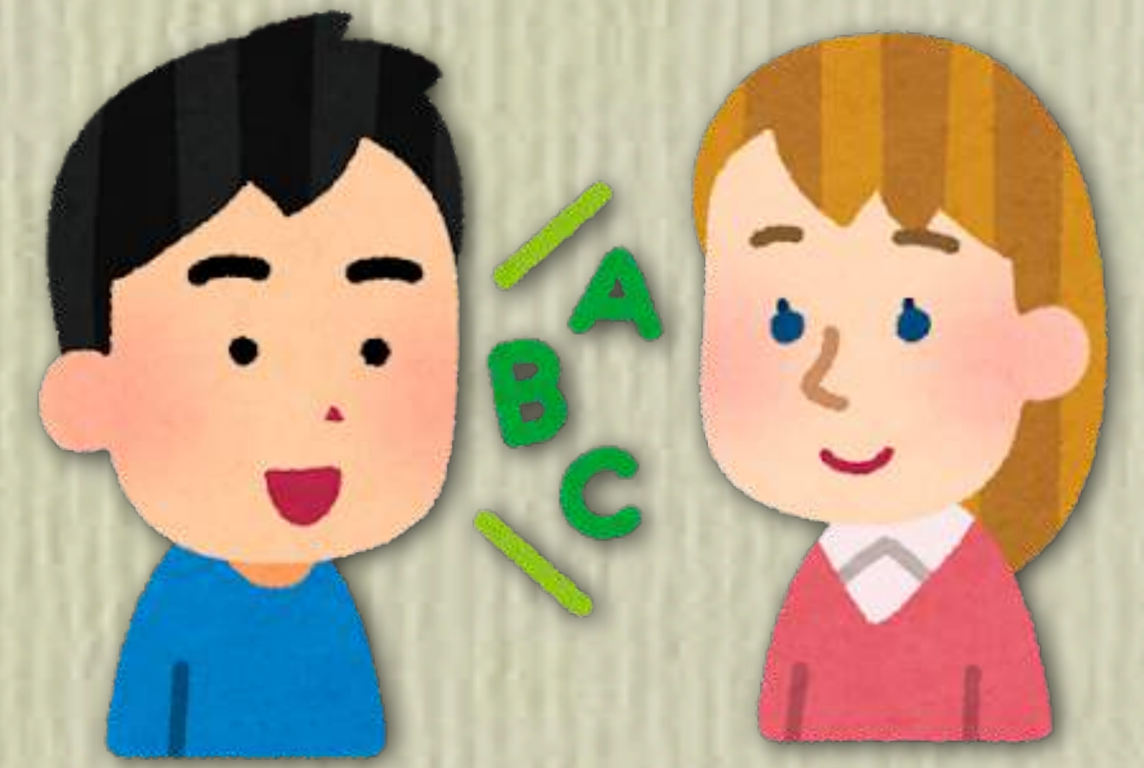
N. Minematsu, et. al., "Inter-learner shadowing with speech technologies enables automatic and objective measurement of comprehensibility of learners' utterances," Proc. AAAL, 2019

World-wide OJAD tutorial workshops

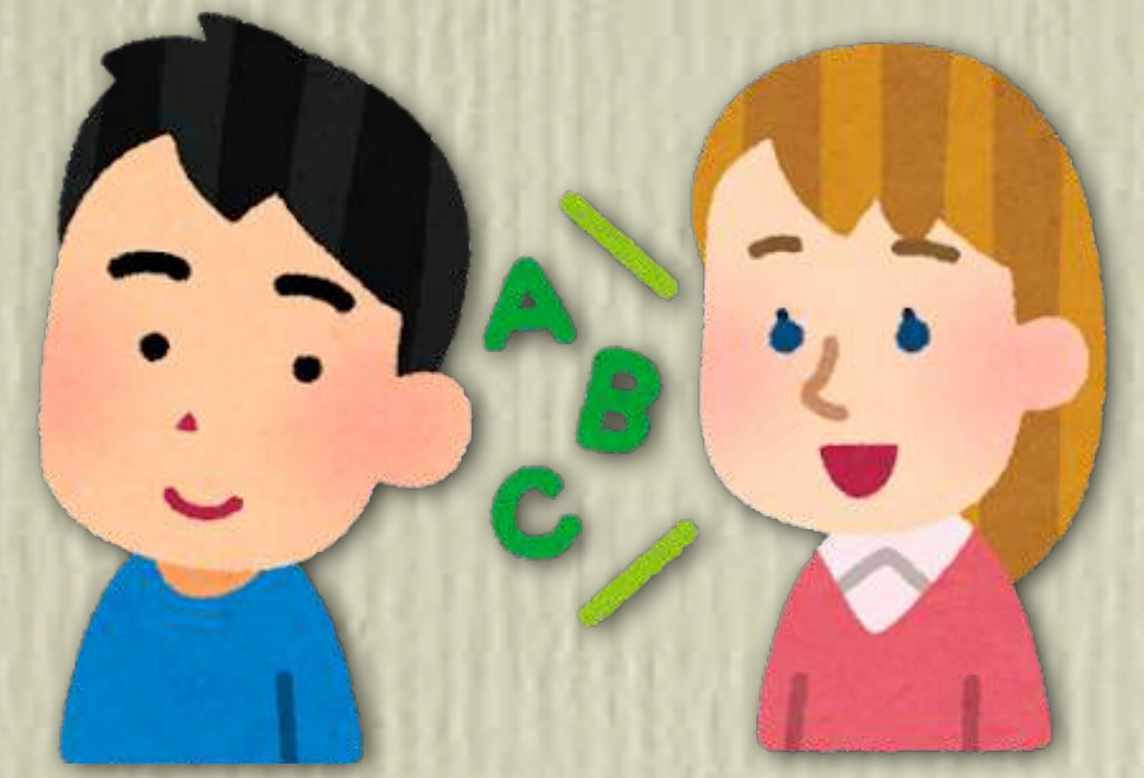
Outline of the presentation

CALL for speaking (reading aloud), listening, conversation, and more

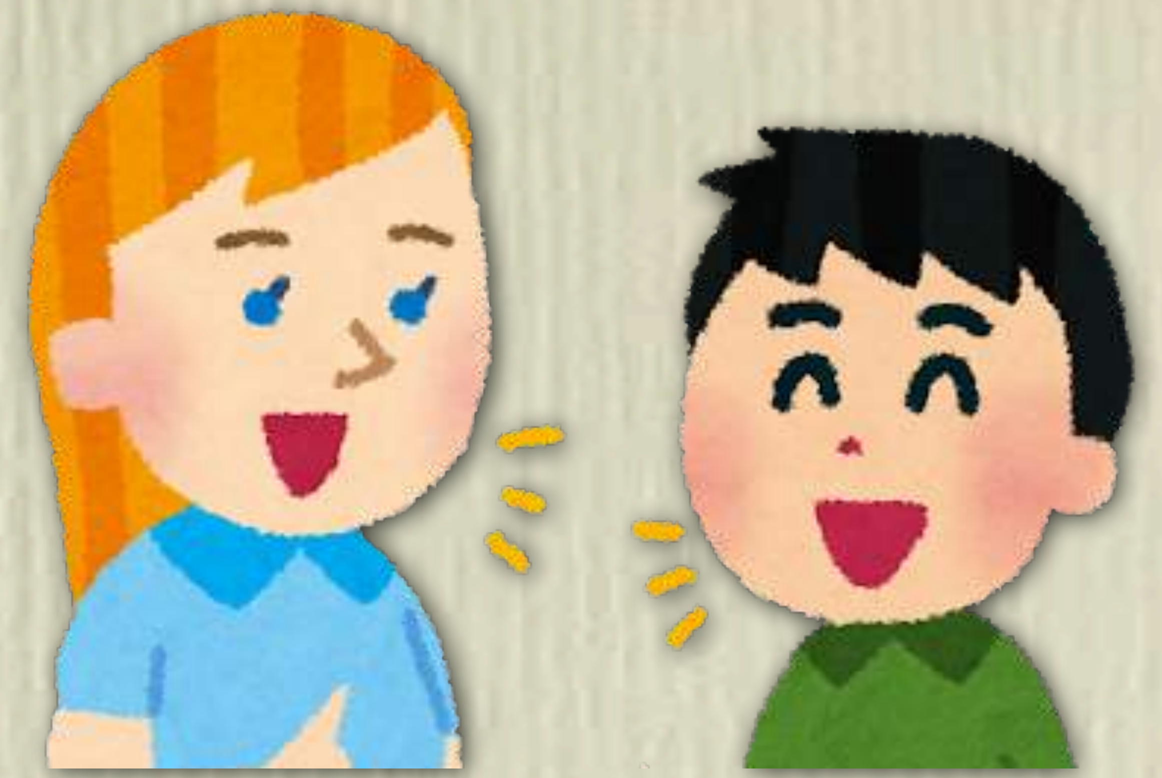
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



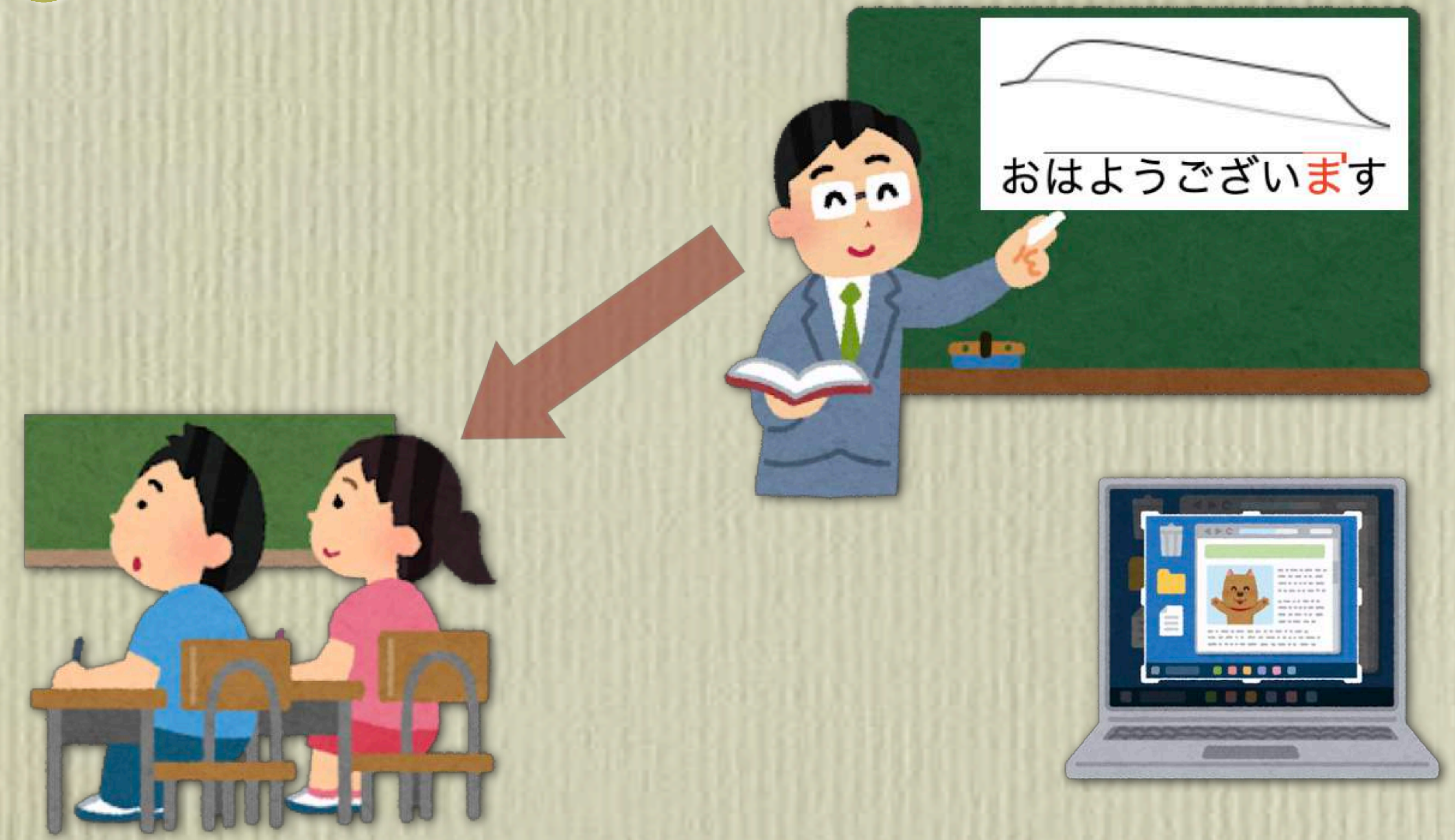
with speech analysis technologies



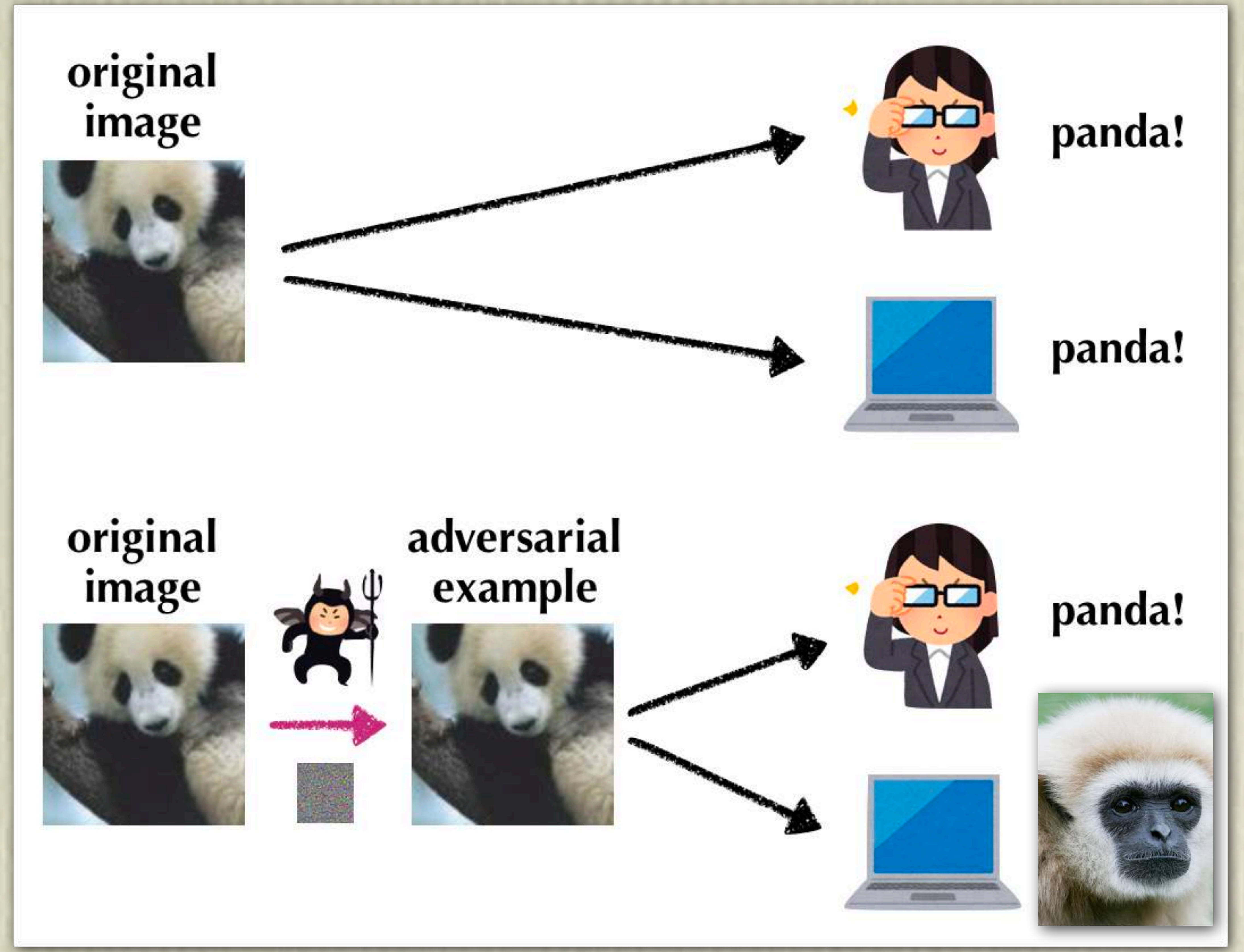
with speech recognition technologies

with new speech technologies being developed in our new project

Conclusions with two illustrations



Listening drills with 



Conclusions with two illustrations

$$\operatorname*{argmax}_{s} P_l(s|o) \quad \operatorname*{argmax}_{u} P_l(u|w)$$

Thank you, boys and girls!!

Thank you, any questions?