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

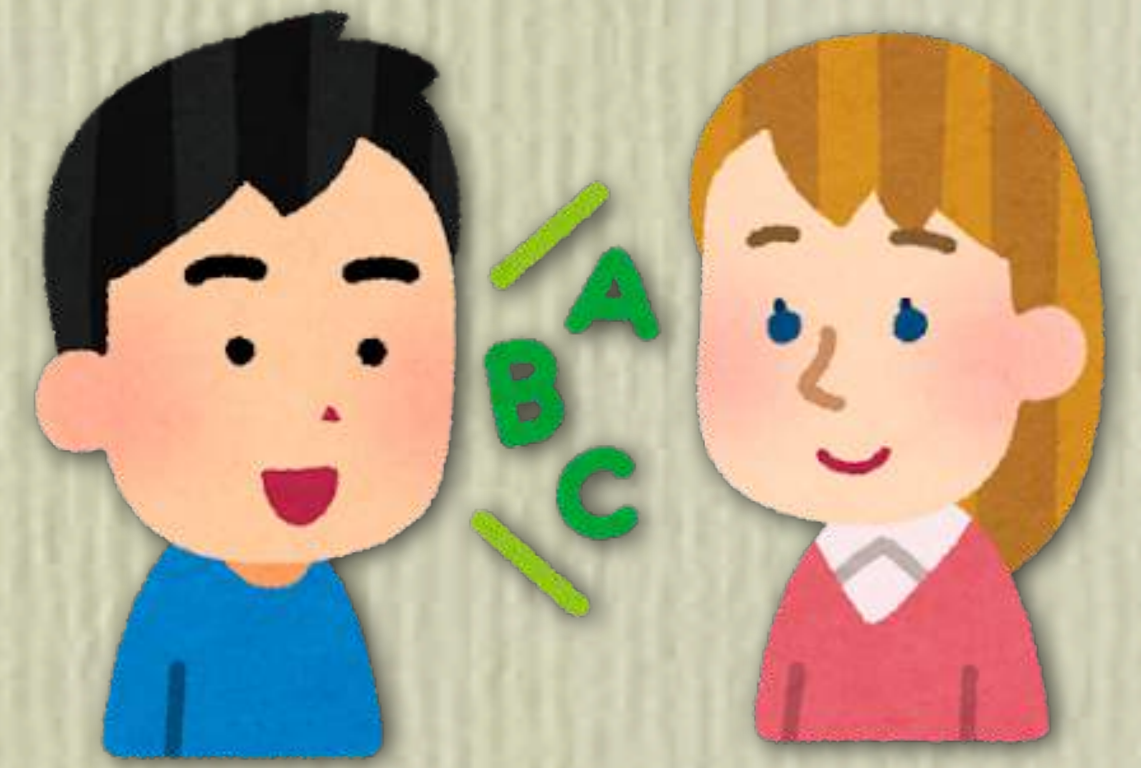
Biography of Nobuaki MINEMATSU

Nobuaki MINEMATSU earned the doctor of Engineering in 1995 from UTokyo and since 2012, he has been a professor there. From 2002 to 2003, he was a visiting researcher at KTH, Sweden. He has a wide interest in speech communication covering speech science and speech engineering, especially he has an expert knowledge on Computer-Aided Language Learning (CALL). When he was a high-school student, he wanted to be a teacher of English, and when he was a university student, he was an amateur actor on English stages. He has published more than 450 journal and conference papers and received paper awards from RISP, JSAI, ICIST, O-COCOSDA, IEICE and an encouragement award from PSJ. He gave tutorial and invited talks on CALL at conferences such as APSIPA2011, INTERSPEECH2012, O-COCOSDA2014, and CASTEL/J2017. He was a distinguished lecturer of APSIPA from 2015 to 2016. He served as secretary of Speech Prosody 2004 and INTERSPEECH2010, co-organizer of SLaTE2010, and program chair of O-COCOSDA2018. He is the general chair of Speech Prosody 2020.

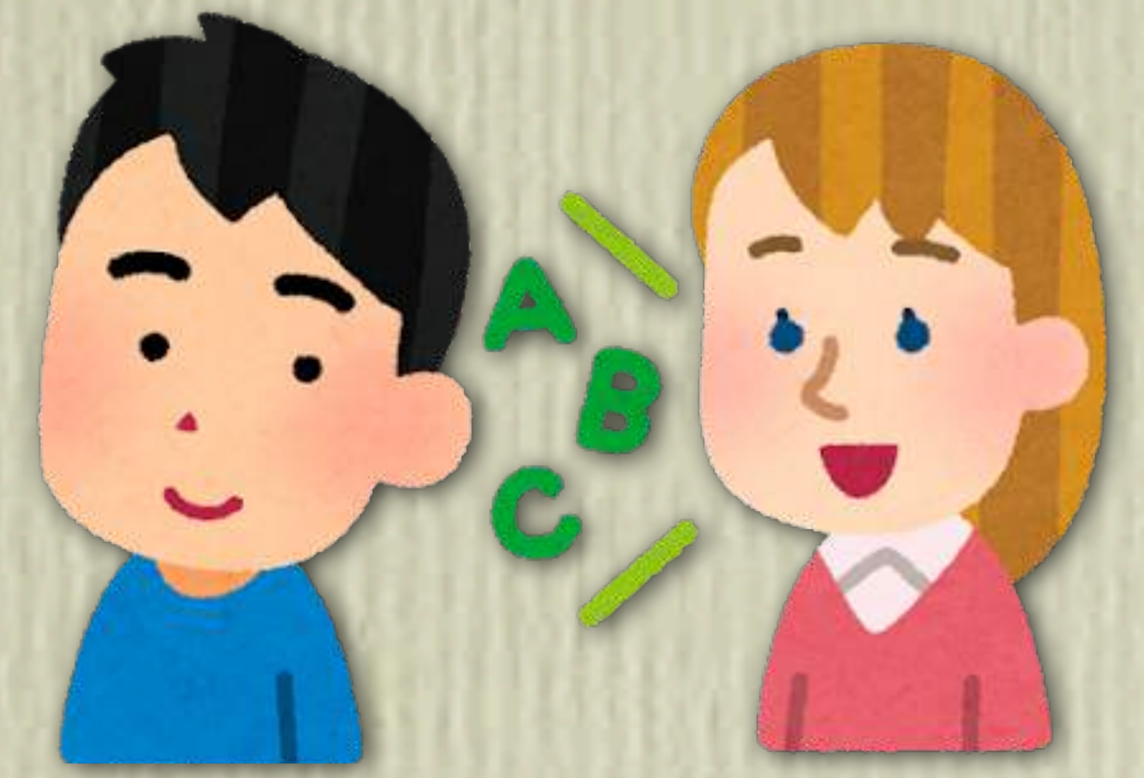
Outline of the presentation

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

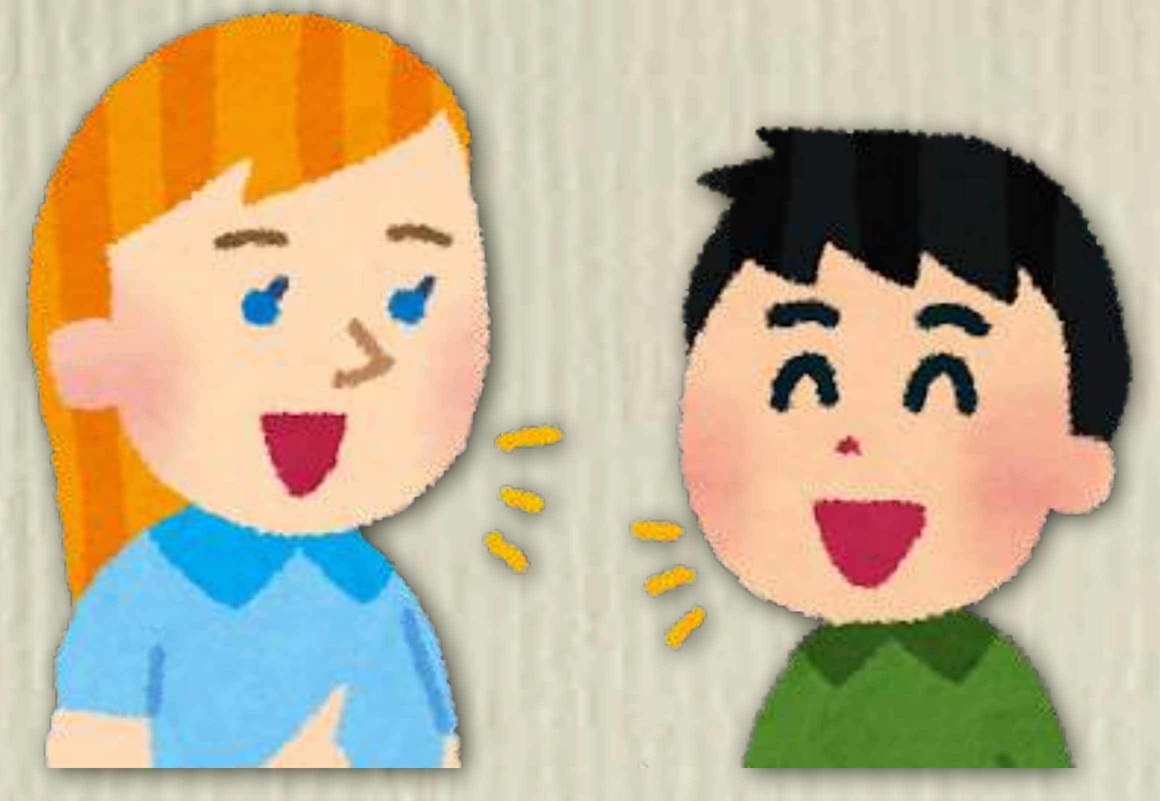
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



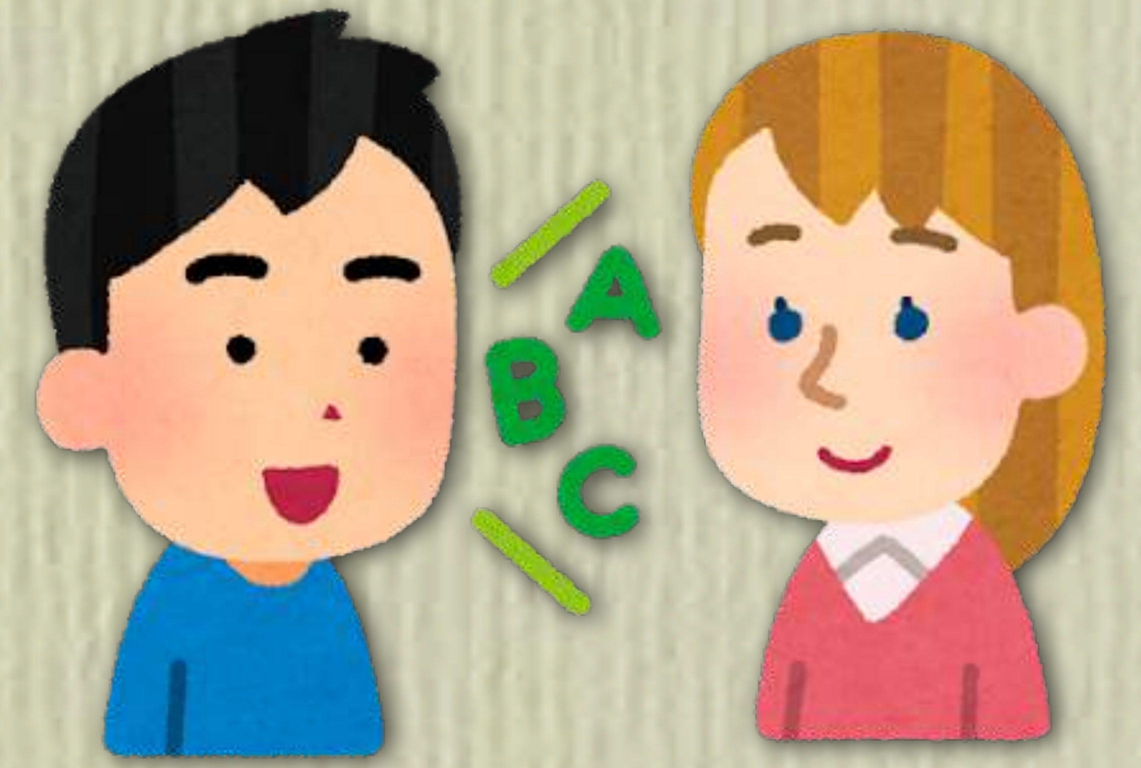
with speech recognition technologies

with new speech technologies being developed in our new project

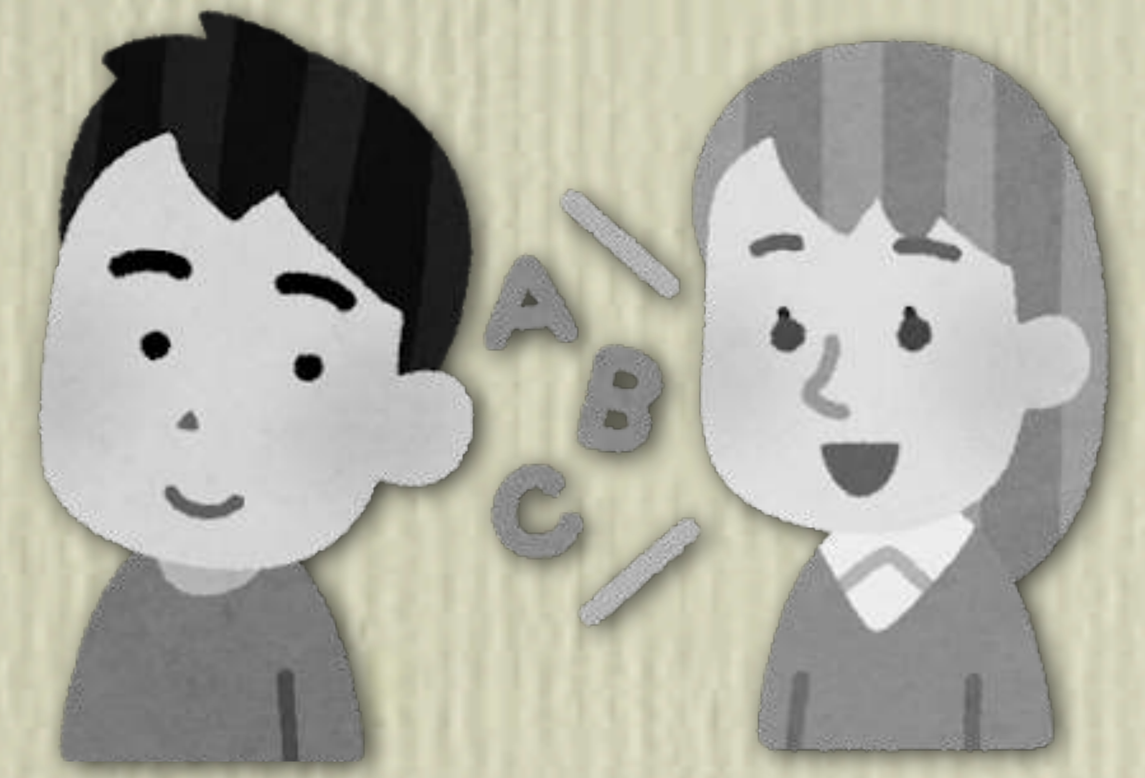
Outline of the presentation

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

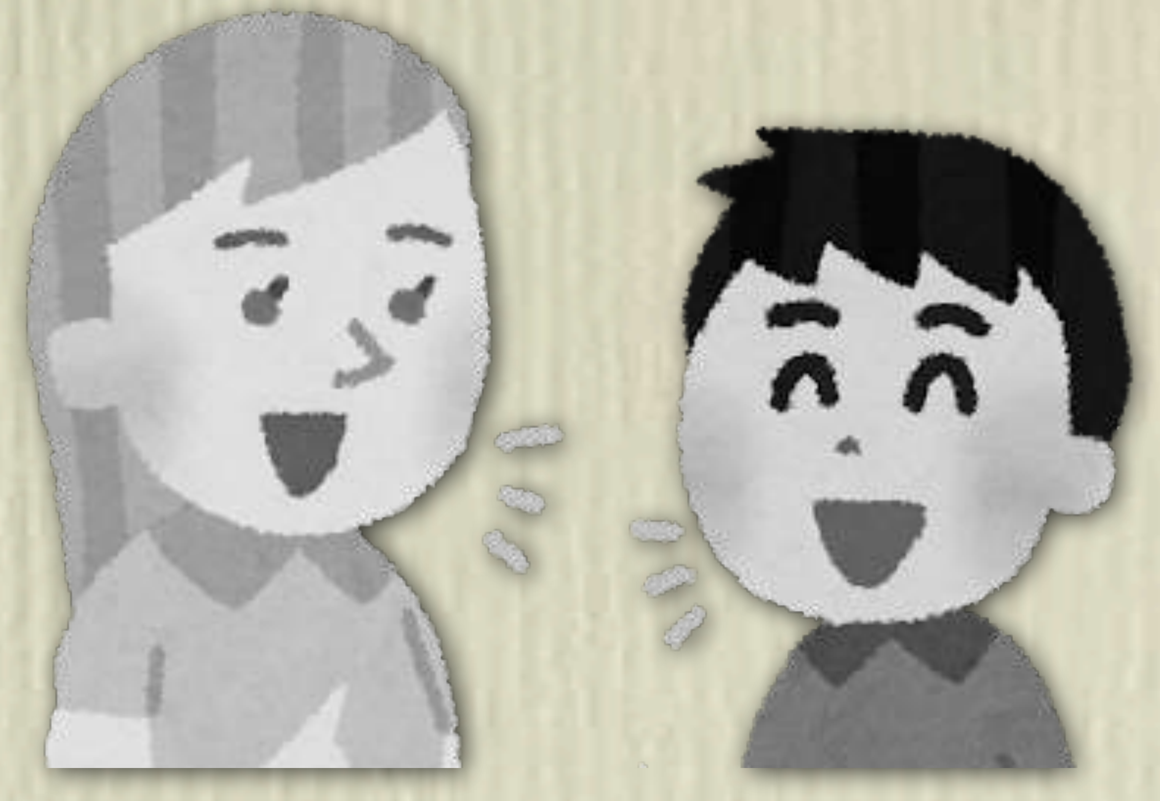
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies ✔



with speech analysis technologies



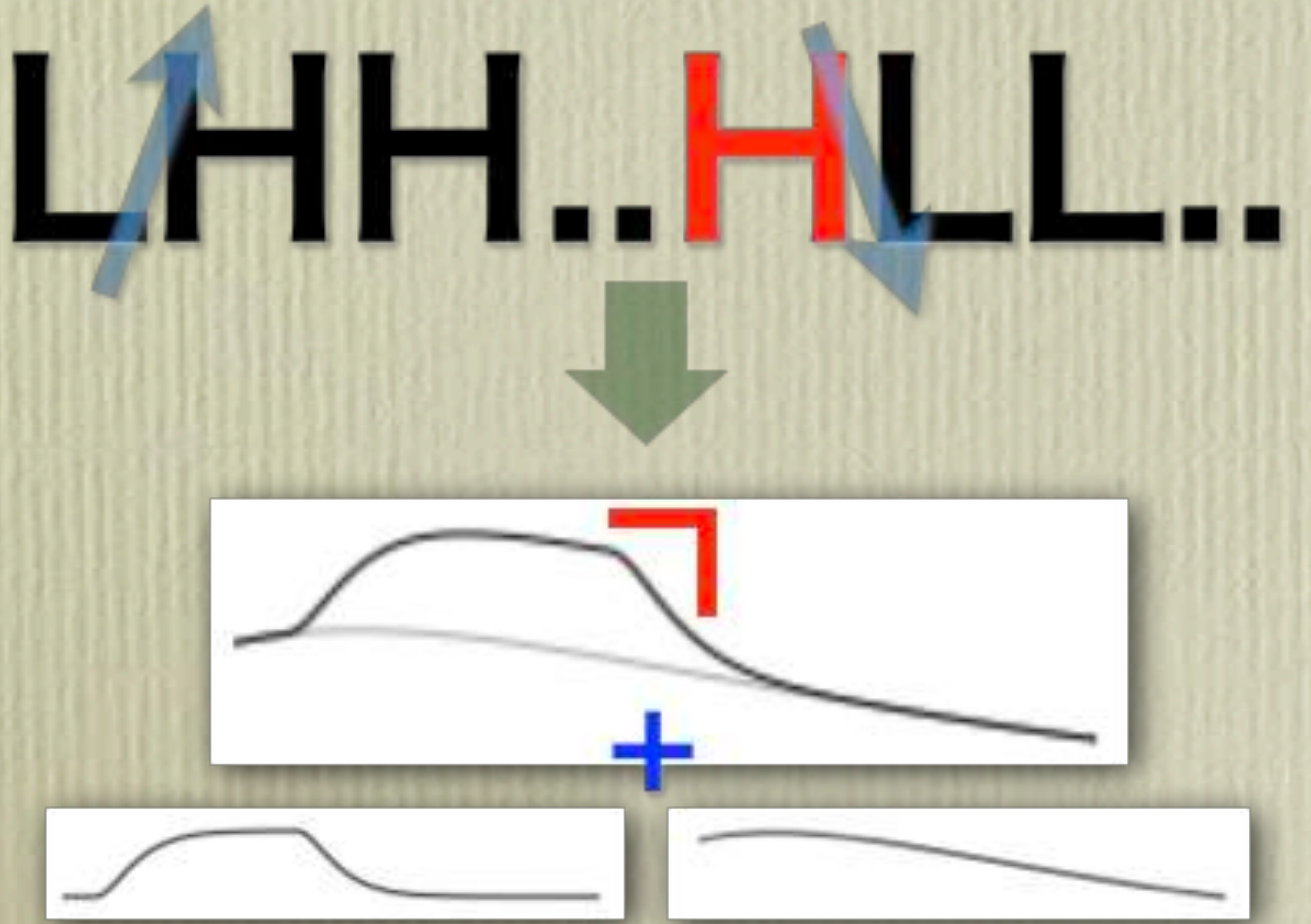
with speech recognition technologies

with new speech technologies being developed in our new project

Japanese prosody — lexical level —

The princeple of word accent control of Tokyo Japanese

- From the 1st mora to the 2nd mora, the pitch level generally goes up (L ➙ H).
- The pitch level goes down somewhere in the word, never goes up again.
- Word accents are classified based on the falling position of pitch (➙ accent type)
  - Accent nucleus = the previous mora (syllable) before the pitch downfall.



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	

Notorious word accent sandhi (change) in Japanese

It is true that each word has its own accent type.

- However, it's also true that accent control is done on a phrase level when speaking.
- That means that lexical accent changes so often depending on context when speaking.

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	

Lexical accent control of Japanese is
SOOOO MYSTERIOUS!!

A fact that is not rare.

An email from a Canadian user of our system, OJAD

OJAD

Dear OJAD,

I just wanted to say THANK YOU for creating your wonderful website!

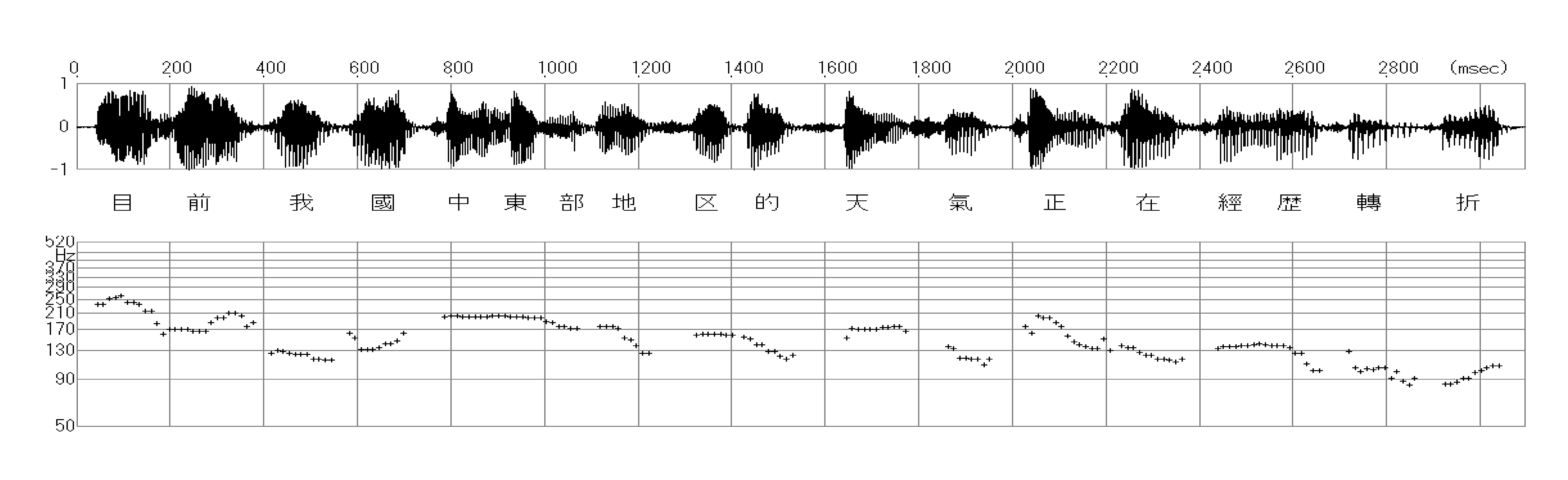
I live in Canada and have been studying Japanese for 10 years - 4 of which were at university. I also worked in Japan for two years. And I had NEVER heard of Pitch Accent - I knew from my Japanese friends that there were different ways to say "hashi" and "ame", but I couldn't hear the difference, and I didn't know that pitch accent actually is a feature that ALL Japanese words have. I finally came across the topic of pitch accent when I started searching for pronunciation books while in Japan, and found a Japanese book that talked about 高低アクセント. I was intrigued and started looking online for more information.

Now that I'm back home and continuing my self-study of Japanese, I have discovered pitch accent, and it makes SOOOO much more sense to me! I always wondered how I could sound "more Japanese" and get rid of my painfully obvious foreign accent. And I knew that Japanese people sounded different than me, but I didn't know what it was or how to train myself to copy it. Thanks to finding online resources like your website, I am enjoying learning about Japanese and am now able to hear pitch changes in Japanese. I also feel like my pronunciation is improving. And of all the resources I've found online, your website is definitely my favorite! *^v^*

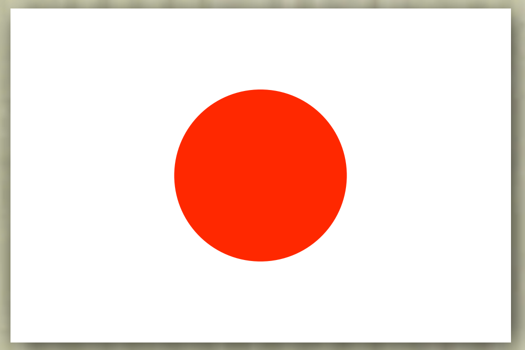
Japanese prosody — phrase and sentence level —

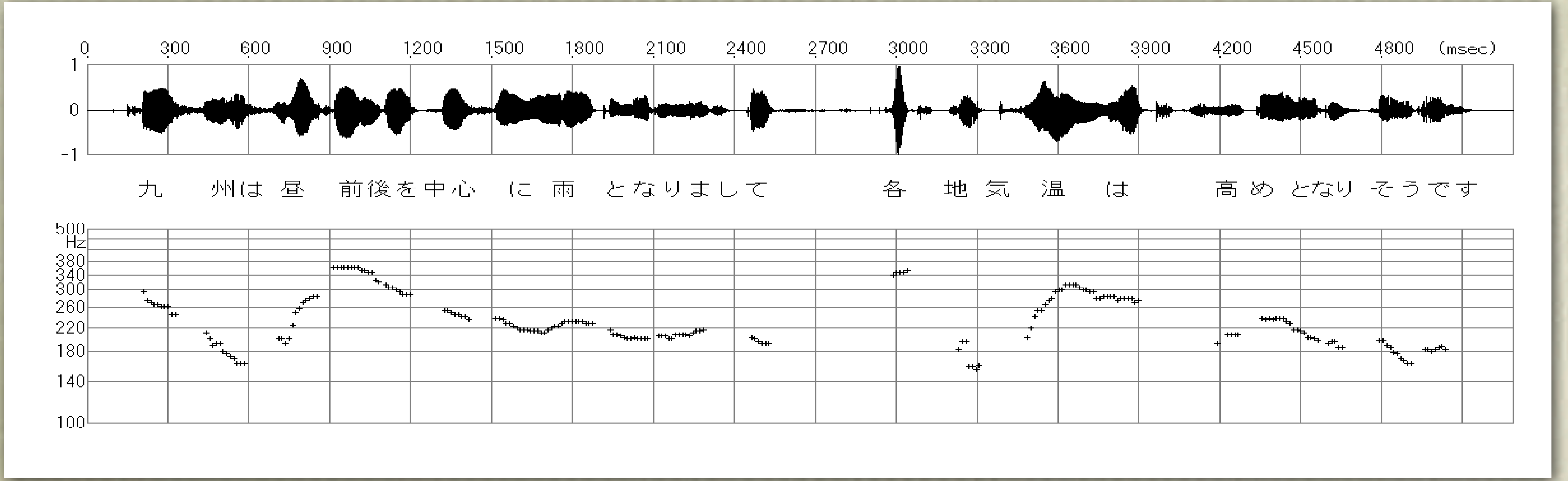
An interesting example of comparison between Chinese and Japanese

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



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

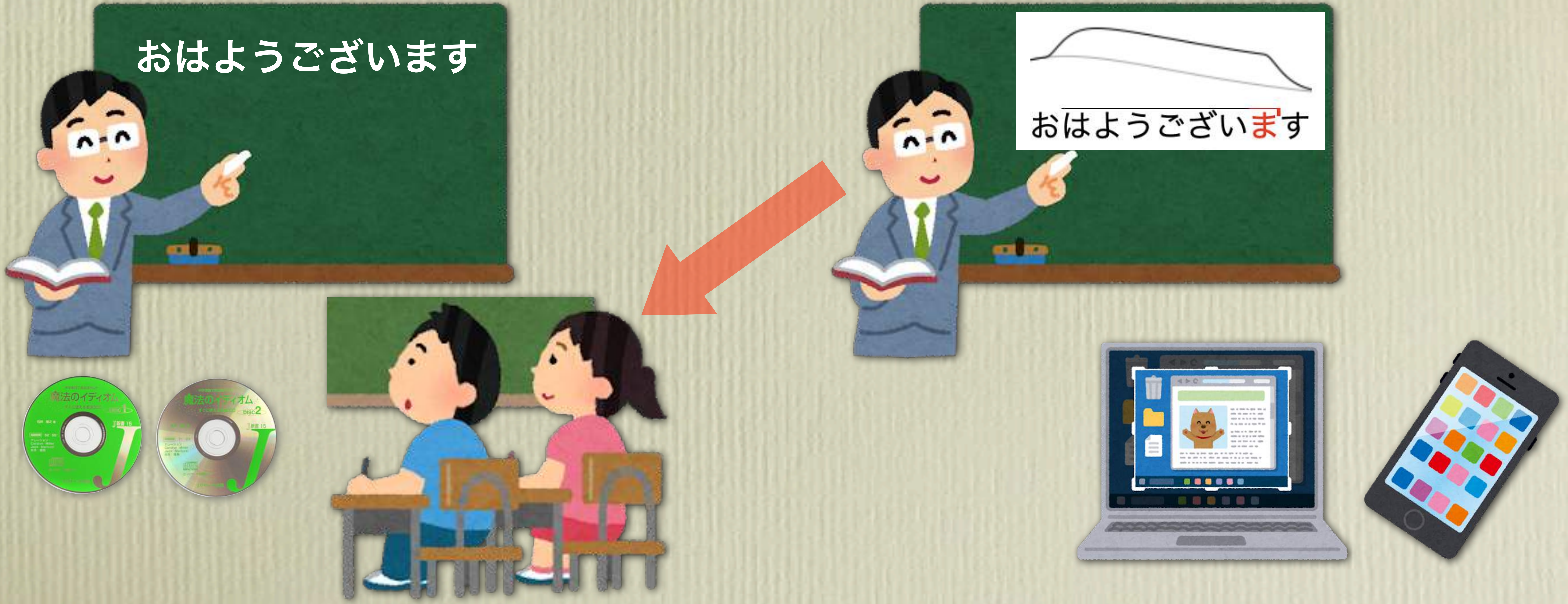




Two kinds of language teachers

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.

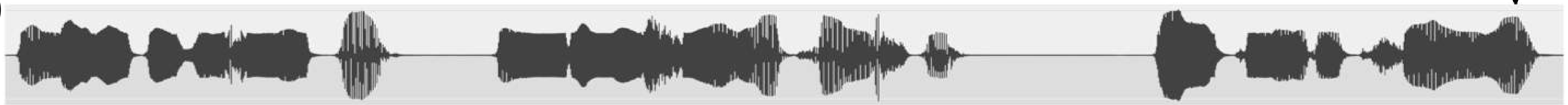
Visualization of prosodic control for speaking in Tokyo Japanese.

1) 日本語はとっても難しいけど，アニメが好きだから，頑張ります。

2) にほんごはとってもむずかしいけど，あにめがすきだから，がんばります。

3) にほんごわ/とっても/むずかし'ーけど_あ'にめが/す%き'だから_がんばりま'す%.

5)



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

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

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."

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.”

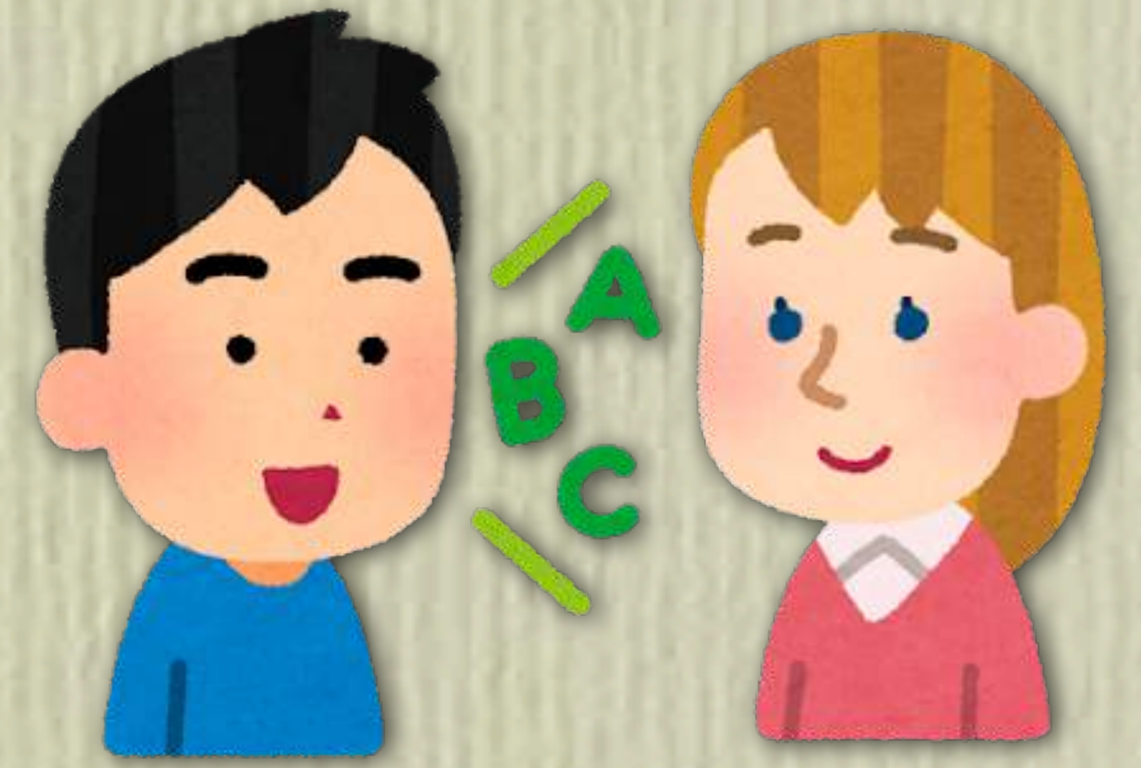
1.5-min promotion video for Suzuki-kun of OJAD

- **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."

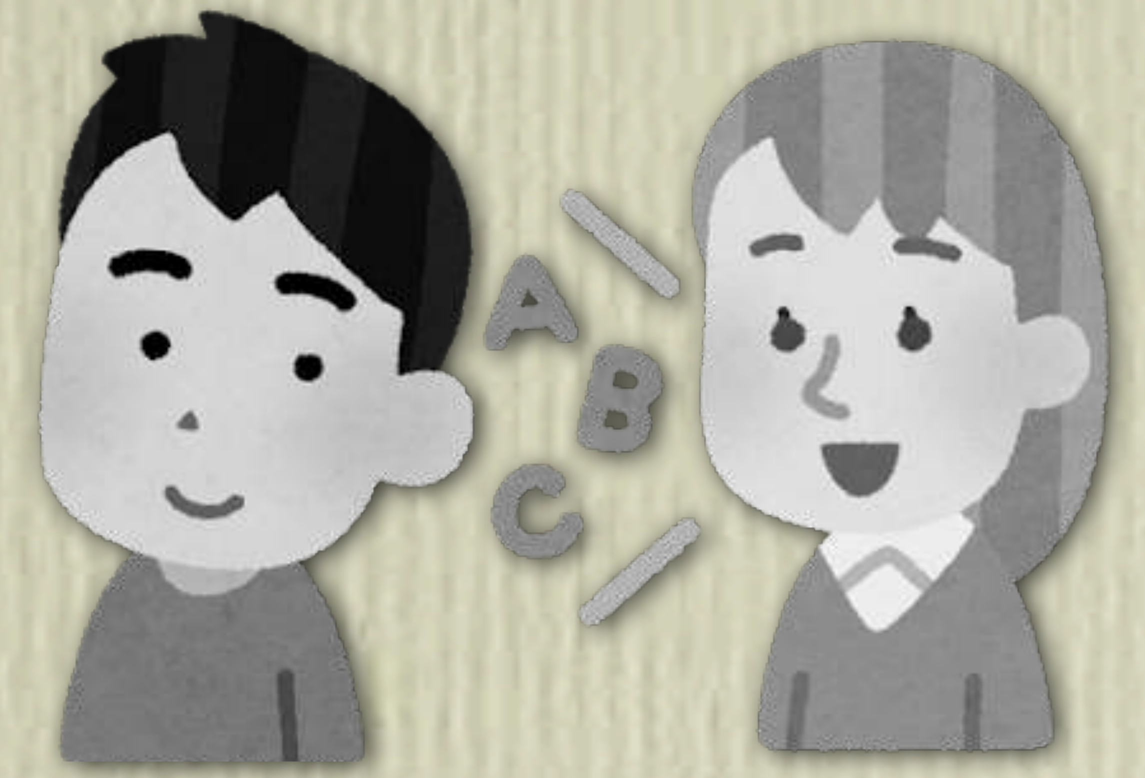
Outline of the presentation

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

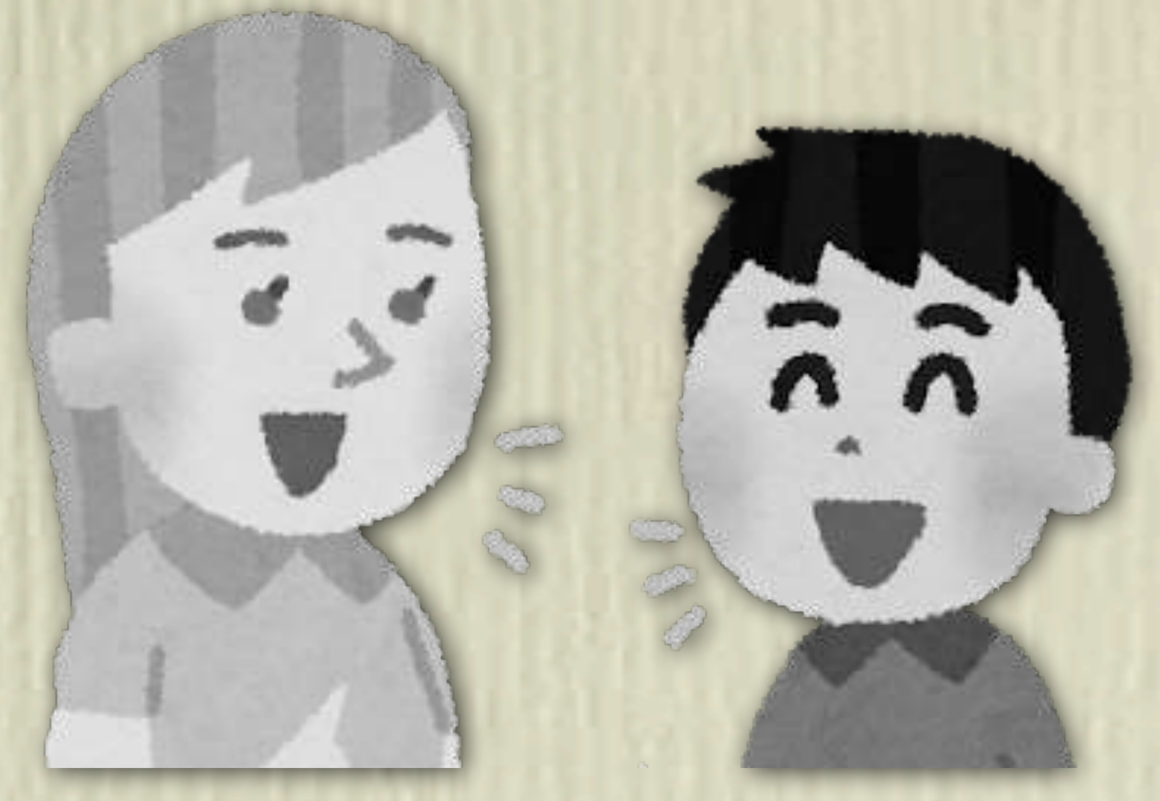
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



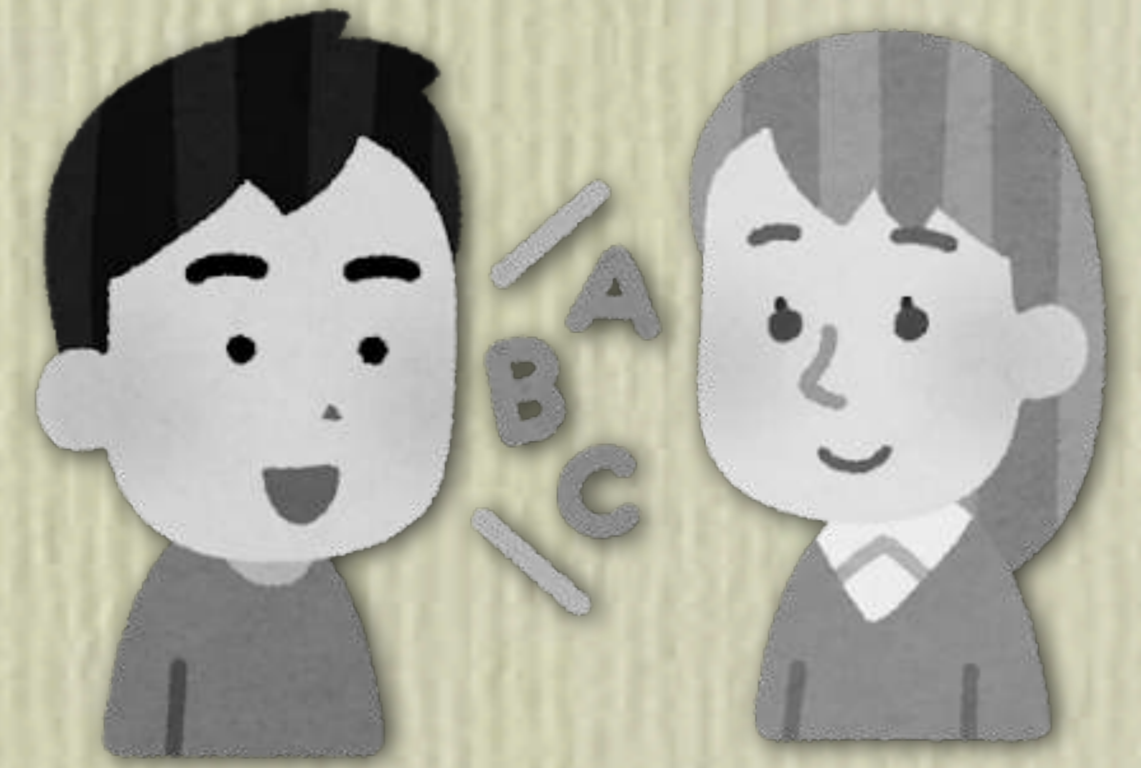
with speech recognition technologies

with new speech technologies being developed in our new project

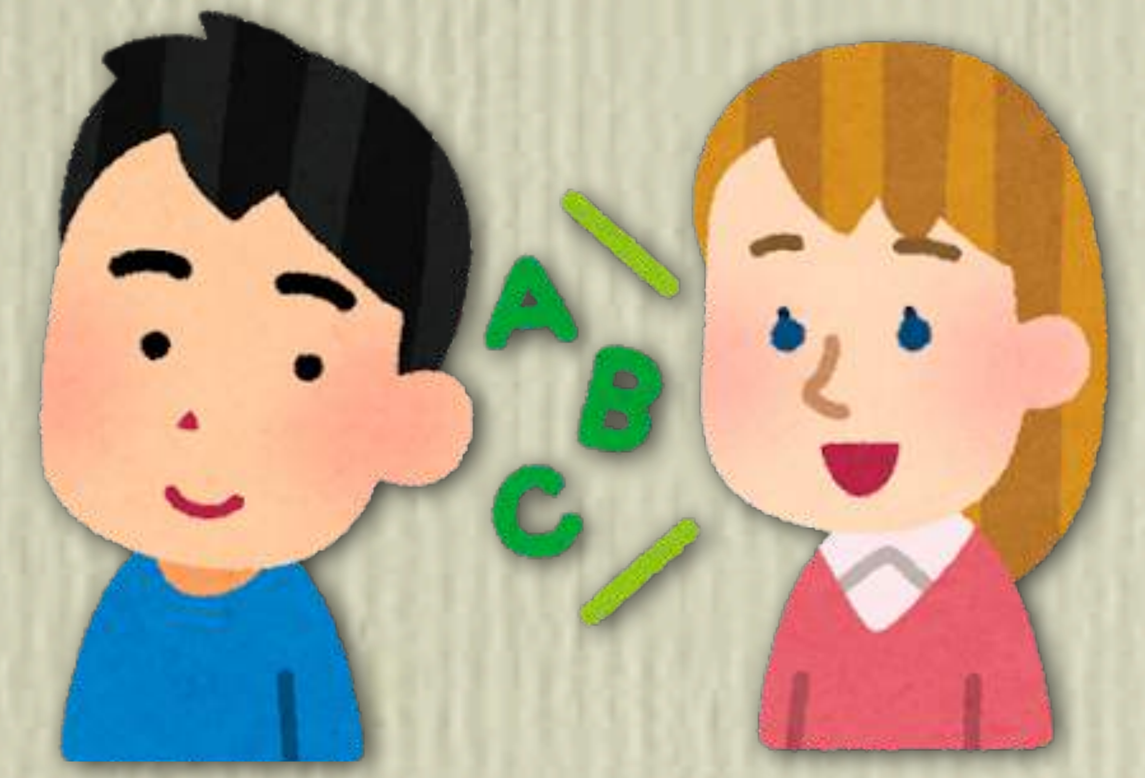
Outline of the presentation

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

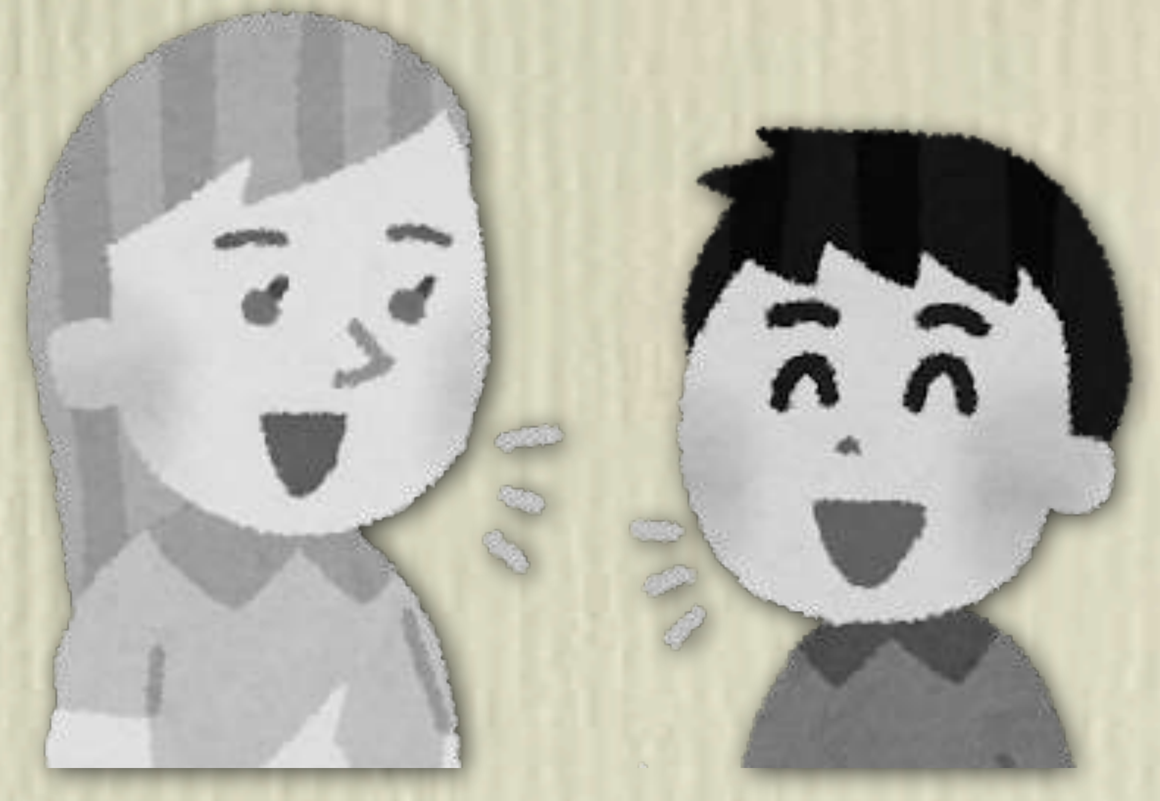
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies ✔



with speech recognition technologies

with new speech technologies being developed in our new project

Real-world conditions of learners' listening

Acoustic conditions of general listening material for learners

- Monologues and dialogues in a clean (no noise) condition.
- Background noises and telephone speeches are sometimes included.
- But noise level or acoustic distortion involved is generally very mild.

Real-world conditions that learners will be faced with

Announcements in a train or bus	background noises
Utterances from very tall or small speakers	age and gender (vocal tract length)
Animation voices designed with a voice changer	characters
Speeches heard in a very big hall	reverberation (echo)
Utterances transmitted through a radio channel	channel distortion

Necessity of robust listening

- Acoustics and speech quality can be easily degraded with various factors.
- Natives can listen but learners may have severe difficulty in listening to those utterances.

A honest confession from a young Japanese pilot

desperate efforts needed for listening to air traffic controllers

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

goo.gl/7YiavD

A training method for robust listening

High Variability Phonetic Training (HVPT)

- Listening training or exercises using speech samples with high variability
  - Speakers, speaking style, gender, age, accents, background noises, etc
  - Often used by language teachers with good knowledge of phonetics
- Many papers or reports of the effectiveness of HVPT
  - Lively+1993, Masuda+2012, Wong+2014, Hwang+2015,
- Teachers often collect various audio samples *manually*.

Technically-enhanced HVPT

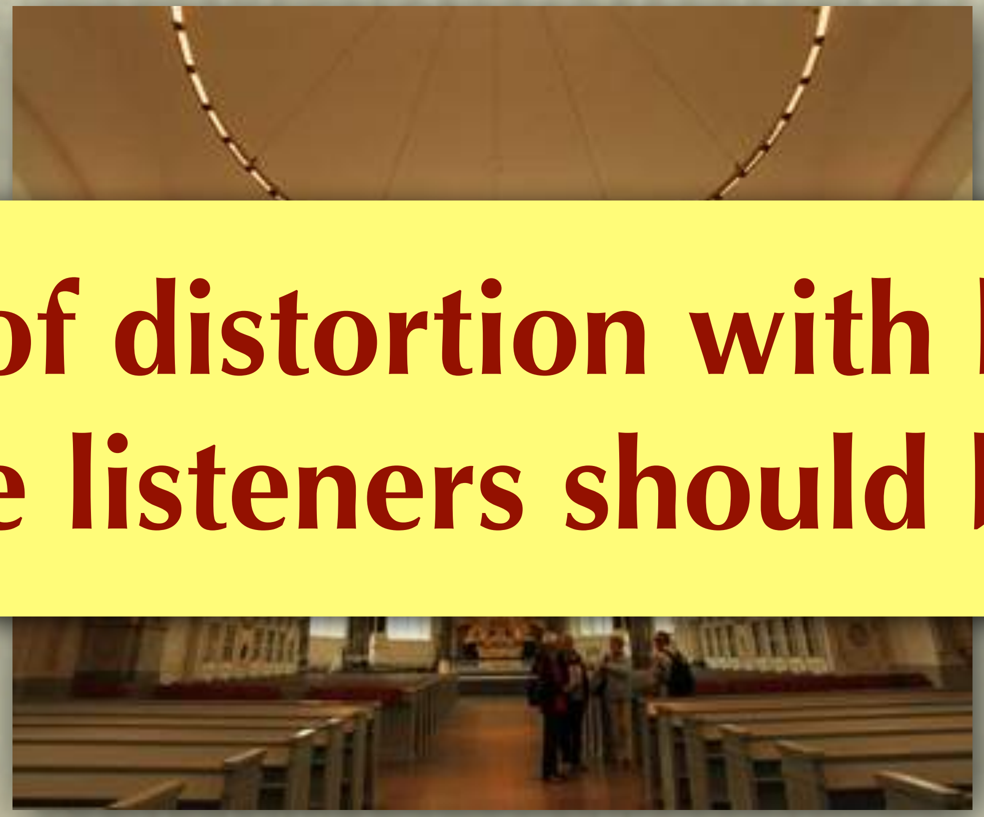
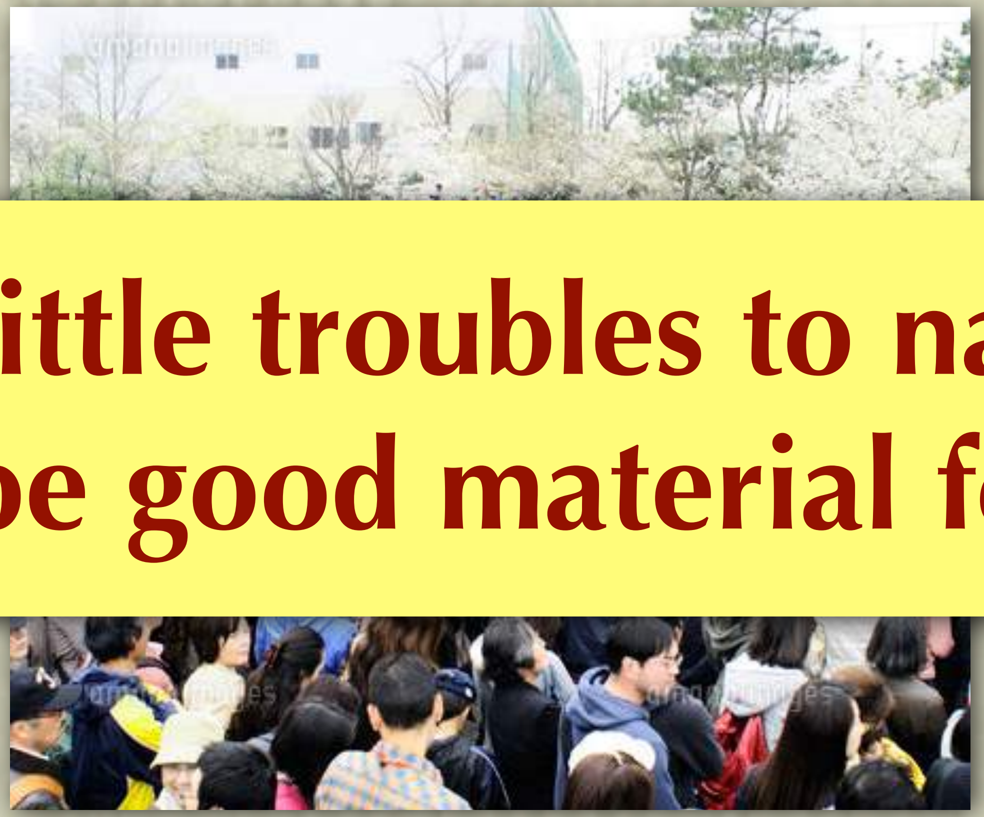
- Speech analysis-resynthesis technologies
  - can convert a single utterance into acoustically various versions with its message unchanged.
- Usability or validity of training with artificially converted audio samples
  - Not only for dictation tasks but also for comprehension tasks

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

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 harsh EIKEN grade 2 listening test

4-choice questions after listening to monologues or dialogues

- Male → giant or giant pilot (ATC)
- Female → fairy or 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 harsh EIKEN grade 2 listening test

- **4-choice questions after listening to monologues or dialogues**
  - Male → giant or giant pilot (ATC)
  - Female → fairy or 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 harsh EIKEN grade 2 listening test

- **4-choice questions after listening to monologues or dialogues**
  - Male → giant or giant pilot (ATC)
  - Female → fairy or 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.

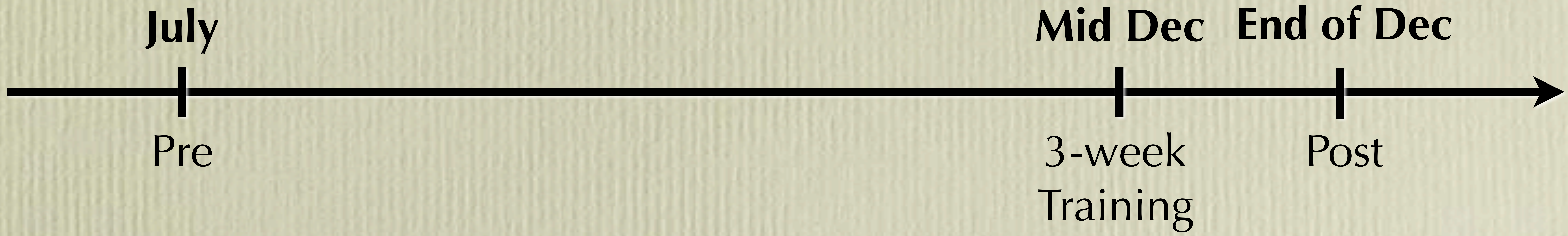
- **Accuracy of Japanese college students and native speakers**

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**

Harsh listening exam → harsh listening drills

Hash listening training drills were developed using ATC distortions.

- Pre: Harsh listening test of EIKEN grade 2
- Training: Listening drills with varying degrees of **ATC distortions only**
- Post: Harsh listening test of EIKEN grade 2 (= Pre)



1. **Listening robustness is improved against ATC distortion?**
2. **Listening robustness is transferred to other kinds of distortion?**

Pre → Drill → Post

18-day listening drills of different levels of ATC

音声ファイル

・音声ファイルをダウンロードすると、雑音レベルに合わせてlevel0〜level3というフォルダが生成され、その下に音声ファイルが保存されます。各音声に対する質問は問題セットをクリックして下さい。
・雑音レベルを３段階用意していますが、level1でも難しすぎる場合はlevel1の難度を落としますので、その旨連絡ください（level0.2, level0.5 のサンプルを上に掲載しています）。 連絡先はこちら。
・音声ファイル（mp3）は10個ずつまとめて，zip ファイルとして提供しています。 スマホの場合，zip ファイルを解凍できないかもしれません。その場合は，以下のツールをダウンロードして zip ファイルを解凍して下さい。 iPhone : iZip Android: file manager

問題セット	音声ファイル（0 : 雑音なし〜3 : 雑音MAX）			
DAY01（11/27）	0	1	2	3
DAY02（11/28）	0	1	2	3
DAY03（11/29）	0	1	2	3
DAY04（11/30）	0	1	2	3
DAY05（12/1）	0	1	2	3

July

Mid Dec

End of Dec

Pre → Drill → Post

Accuracy of pre-test and post-test

- A half of the pre test examinees (55 students) undertook the post test.

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

- Error reduction rate
  - Accuracy: 70% → 85 %
  - ERD = (30-15)/30 = 50 %
  - A: monologue, B: dialogue

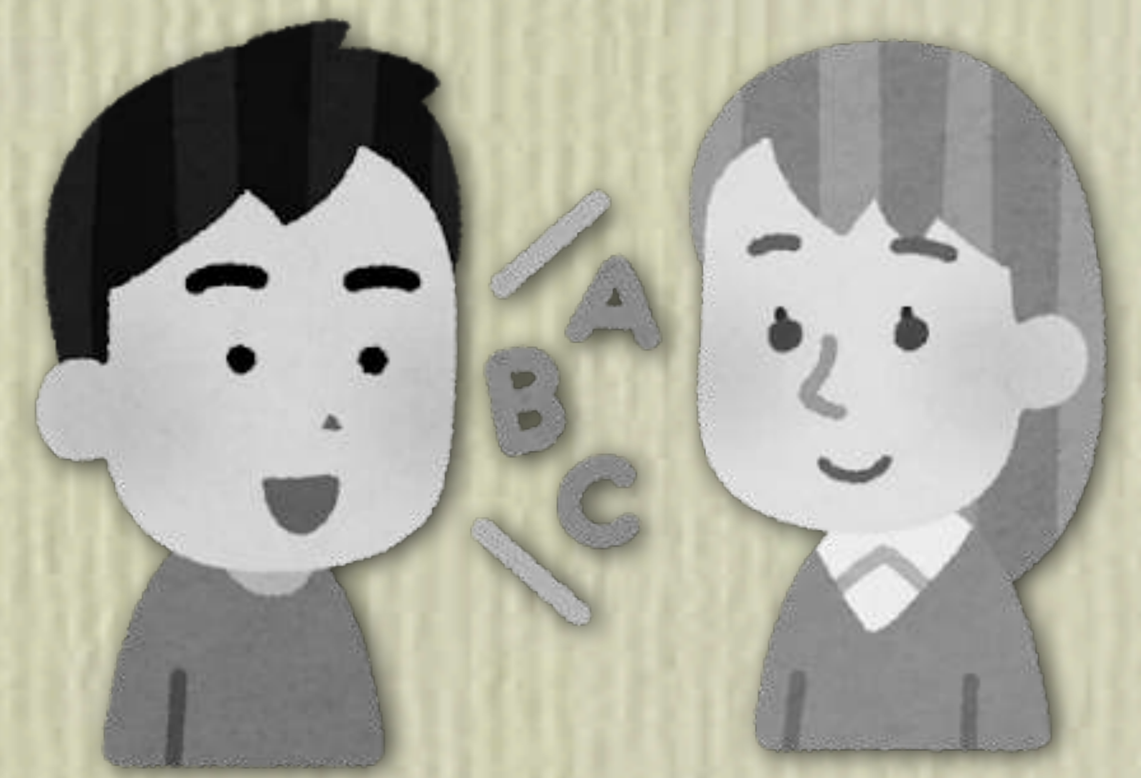
Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	9.9	35.6	2.3	-11.5
	600–800	32	-17.2	22.7	3.7	9.4
	800–990	8	**80.1**	**80.1**	**55.5**	23.9
B	400–600	15	33.4	8.8	13.9	0
	600–800	32	-11.3	6.2	5.1	6.8
	800–990	8	**42.9**	**58.4**	**40.0**	4.5

July → **Mid Dec** → **End of Dec**

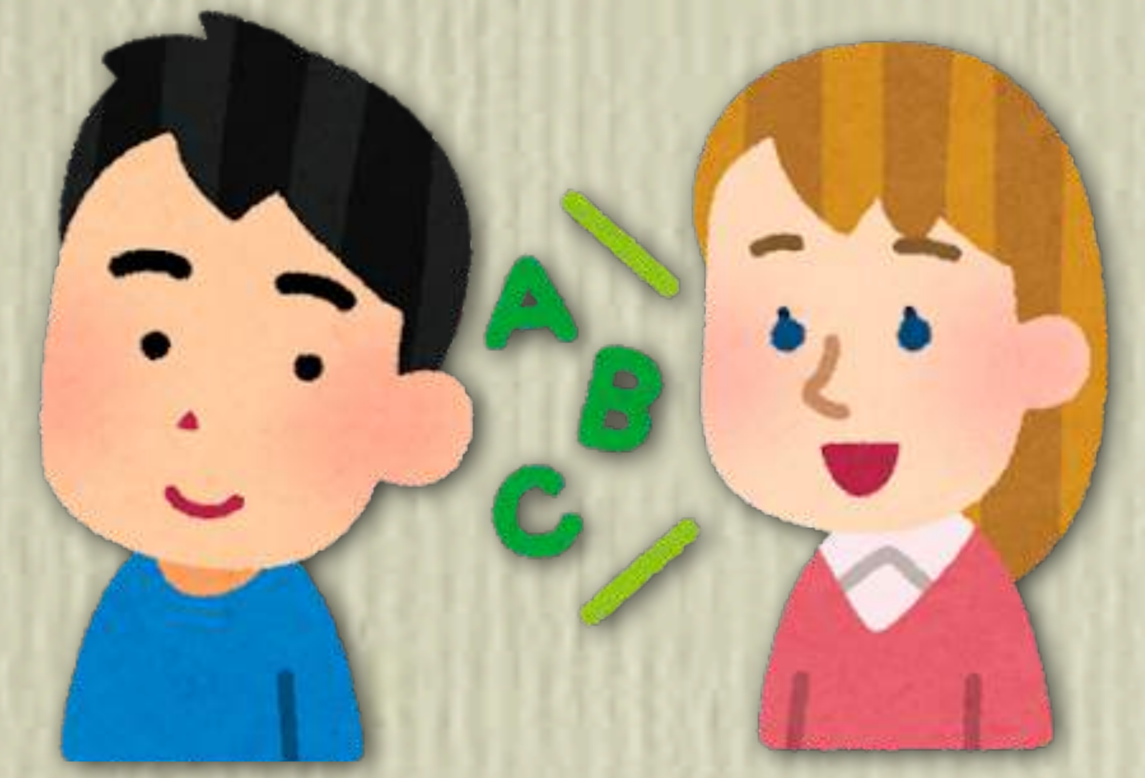
Outline of the presentation

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

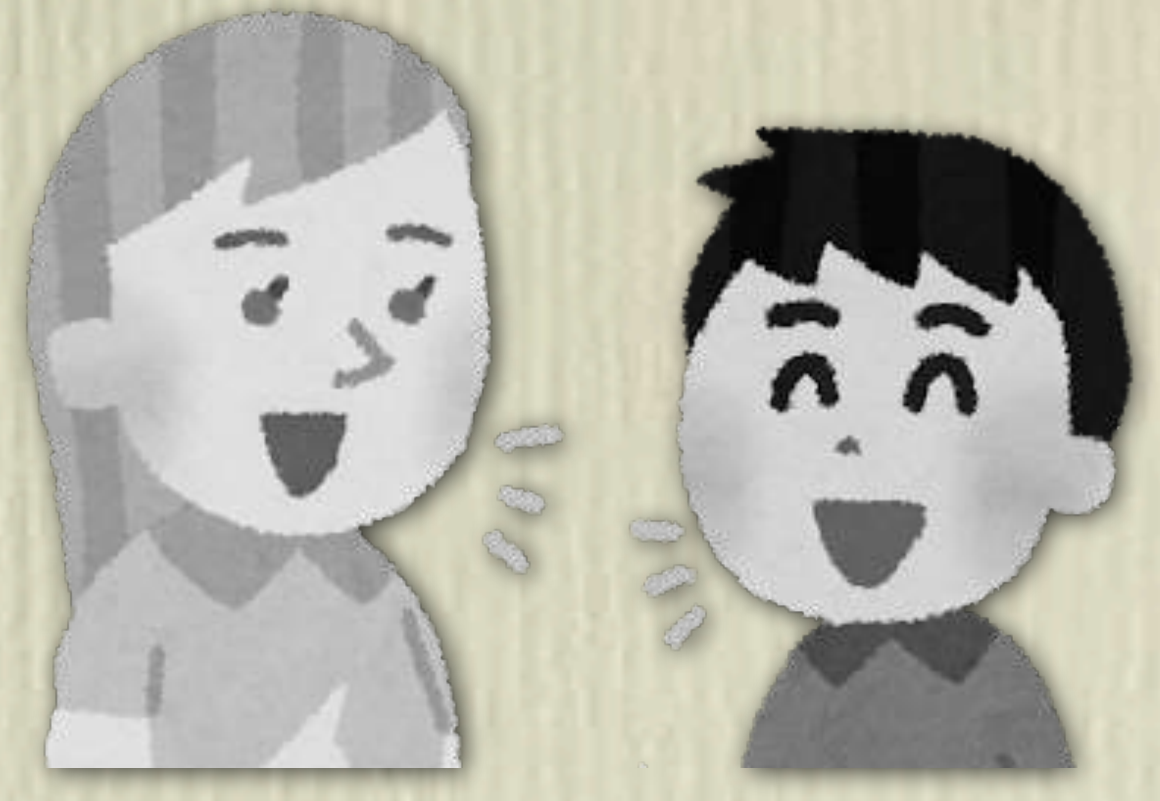
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



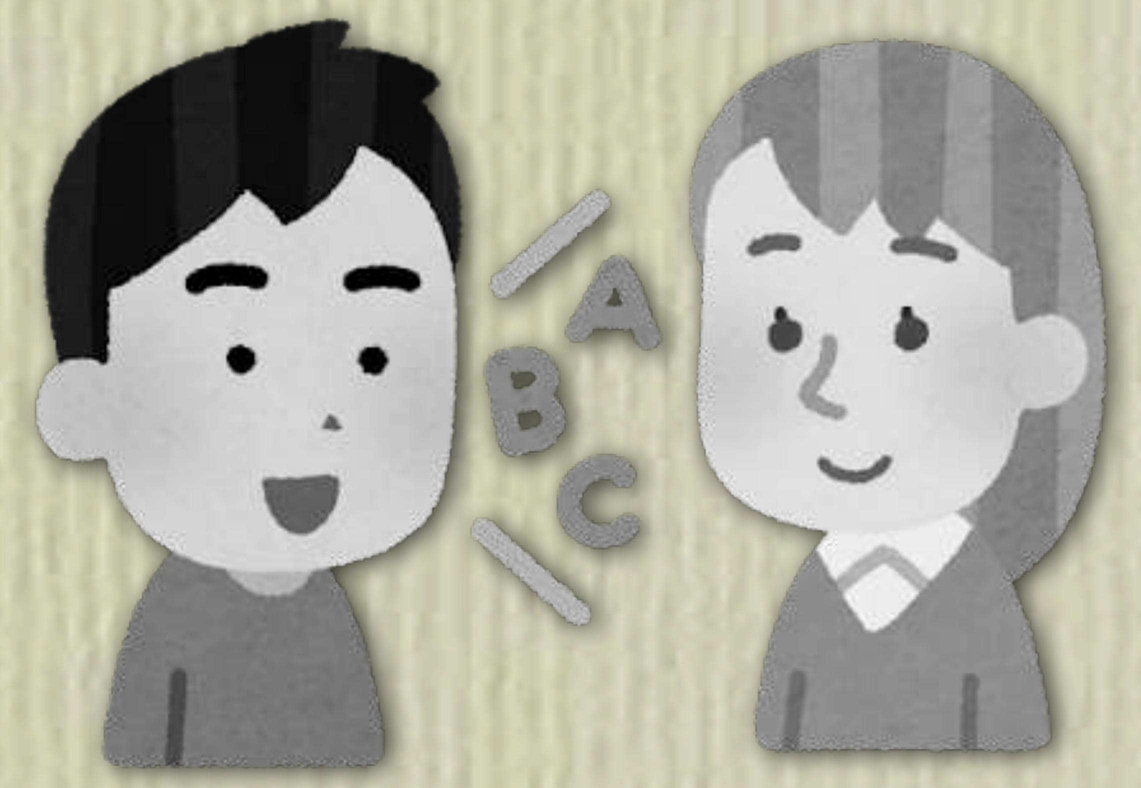
with speech recognition technologies

with new speech technologies being developed in our new project

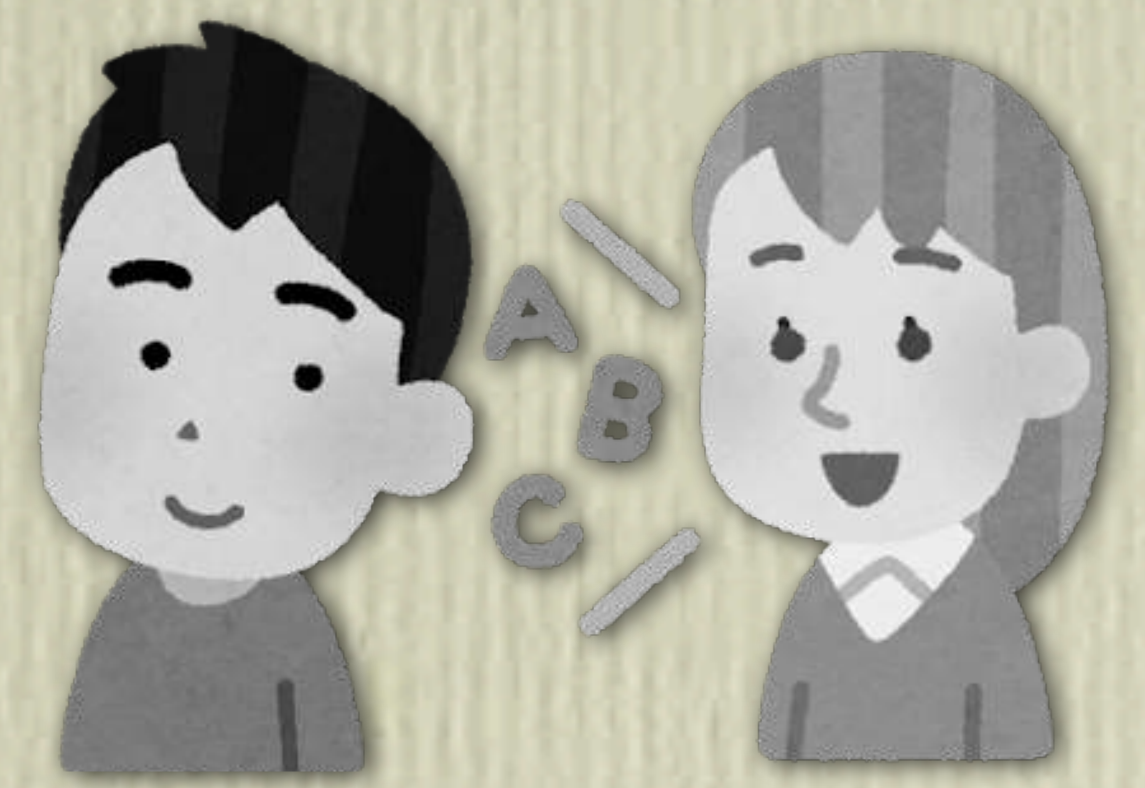
Outline of the presentation

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

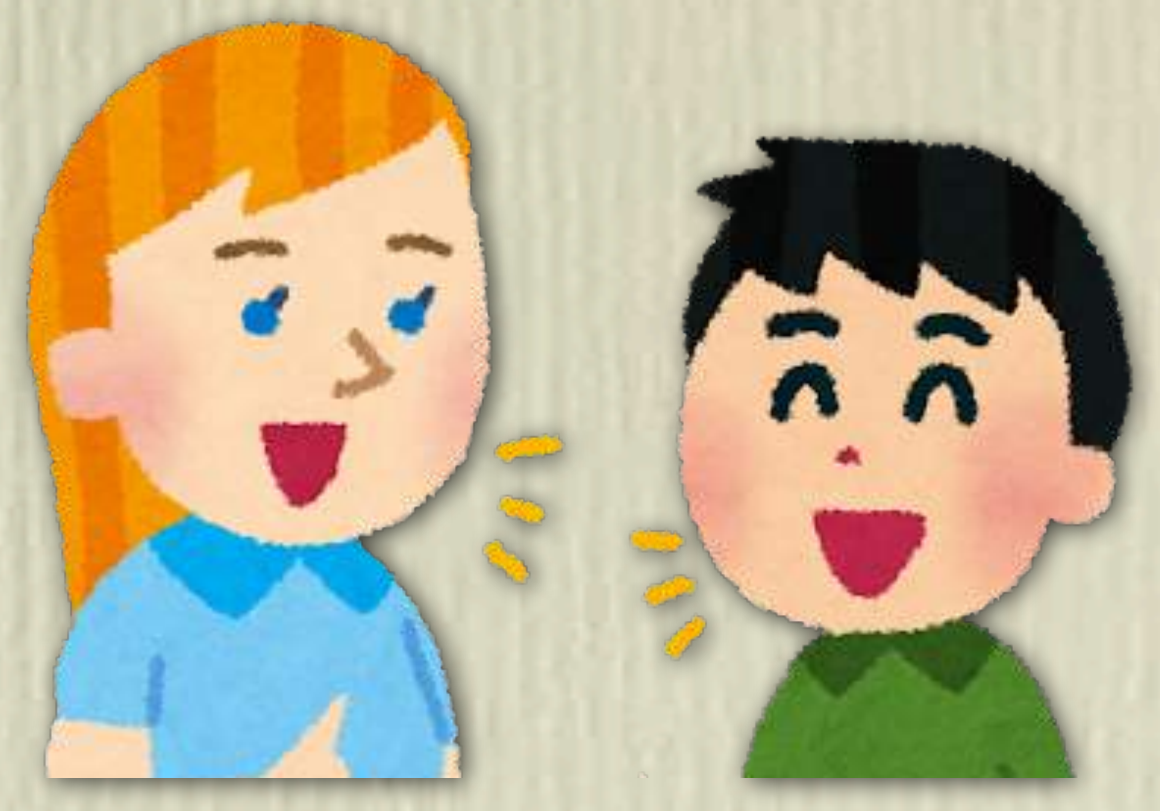
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies

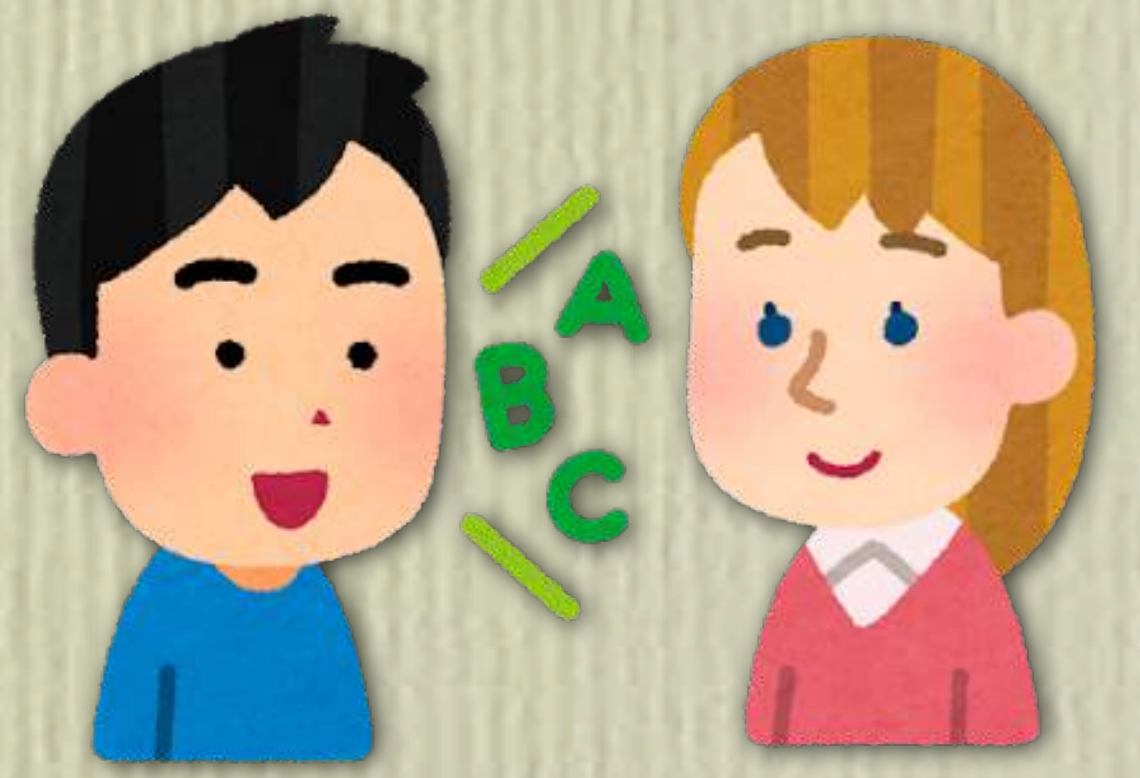


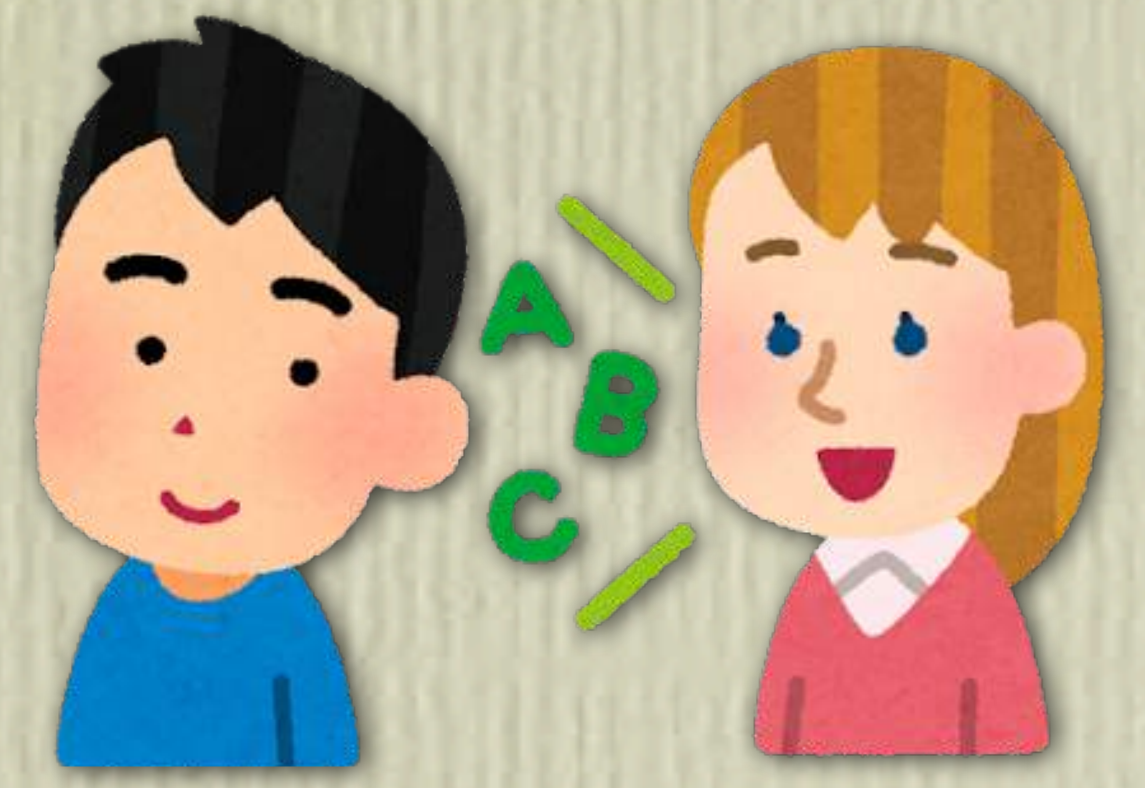
with speech recognition technologies ✔

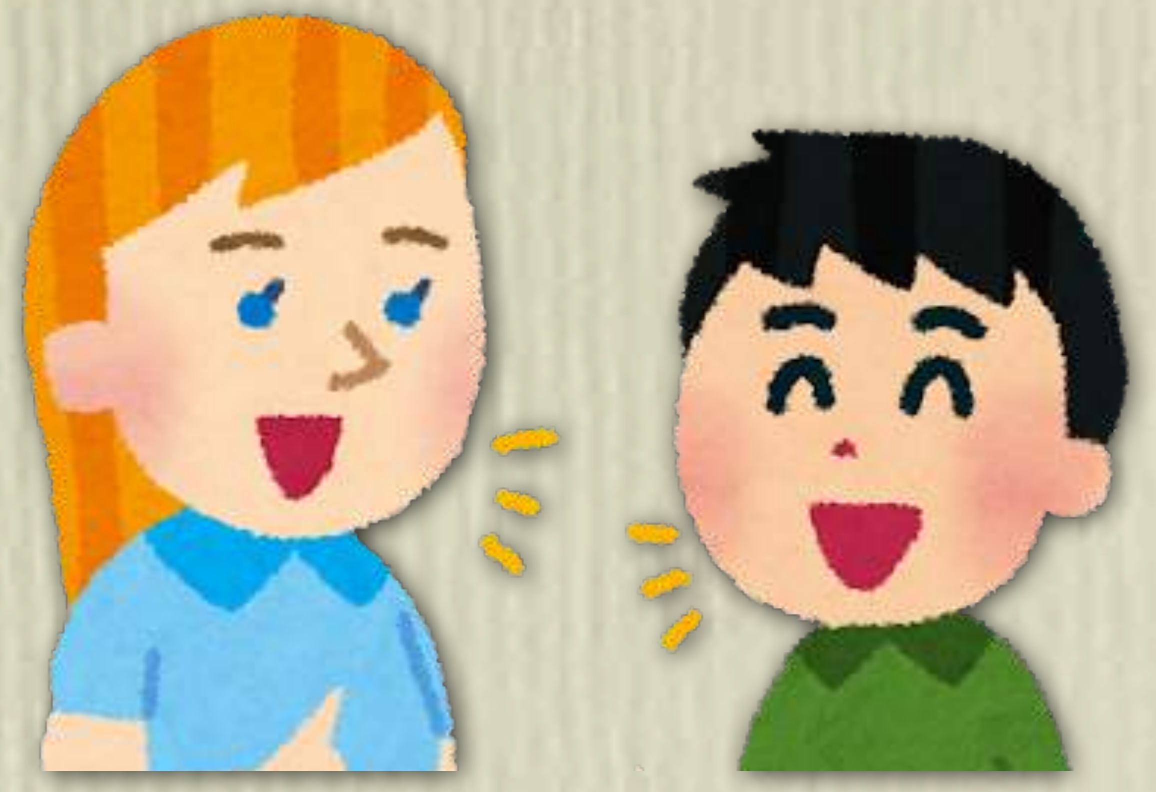
with new speech technologies being developed in our new project

Conversation is a multi-task speech activity.

- Listening, understanding, and speaking running almost together

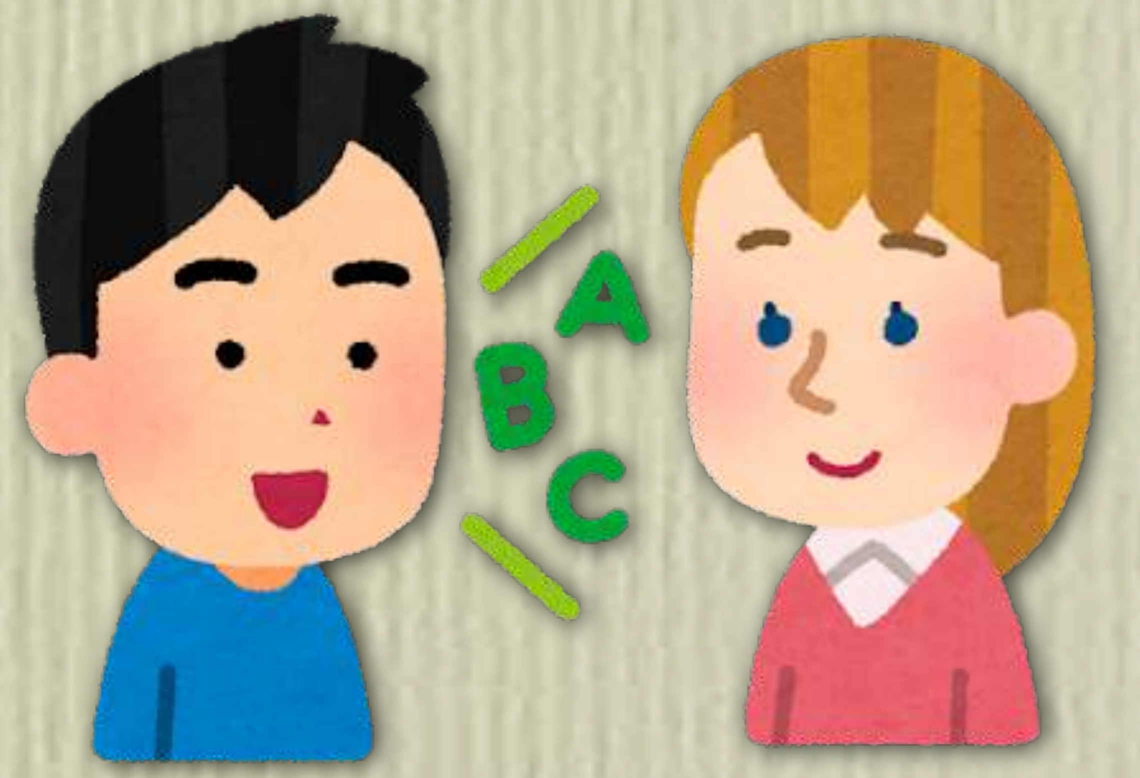


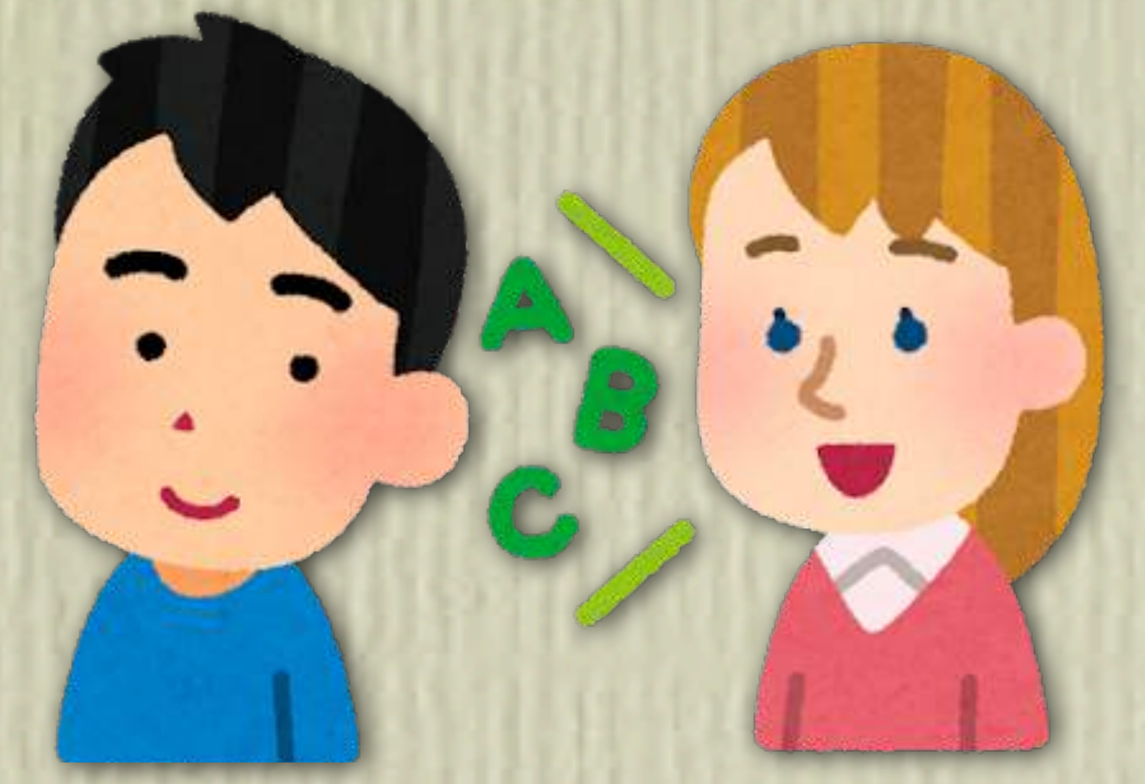


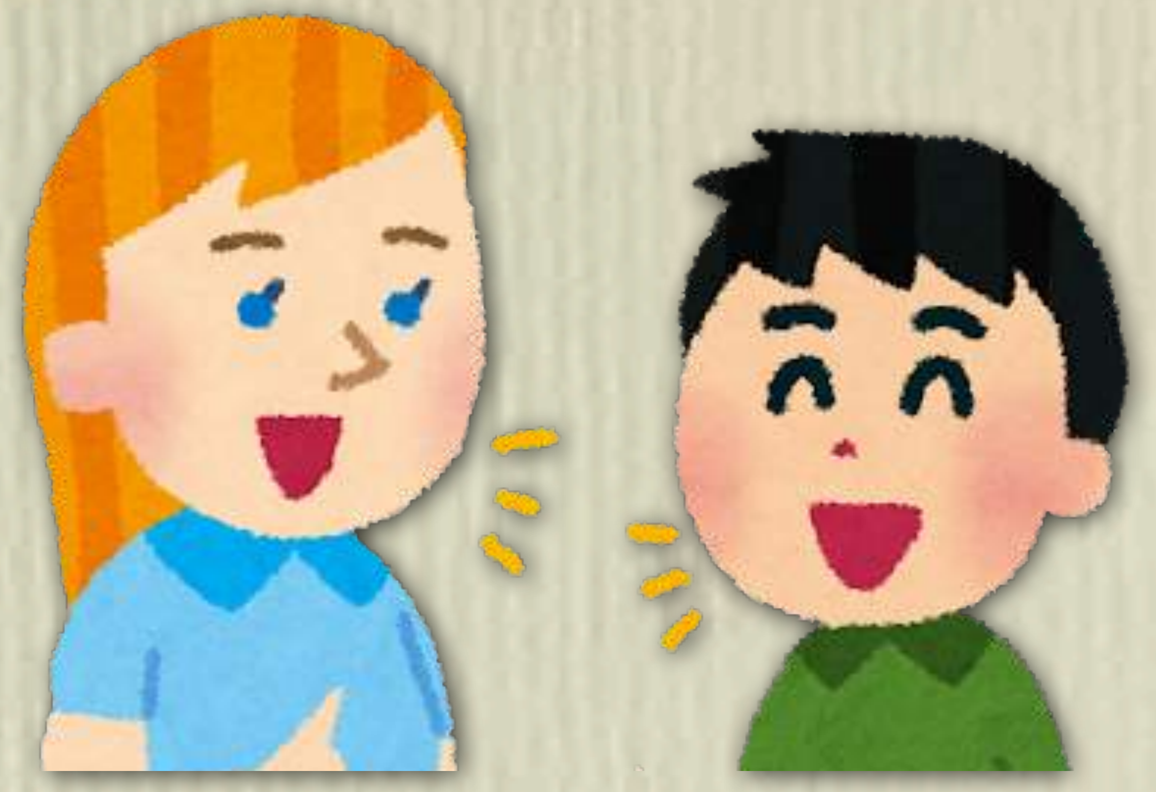


Conversation is a multi-task speech activity.

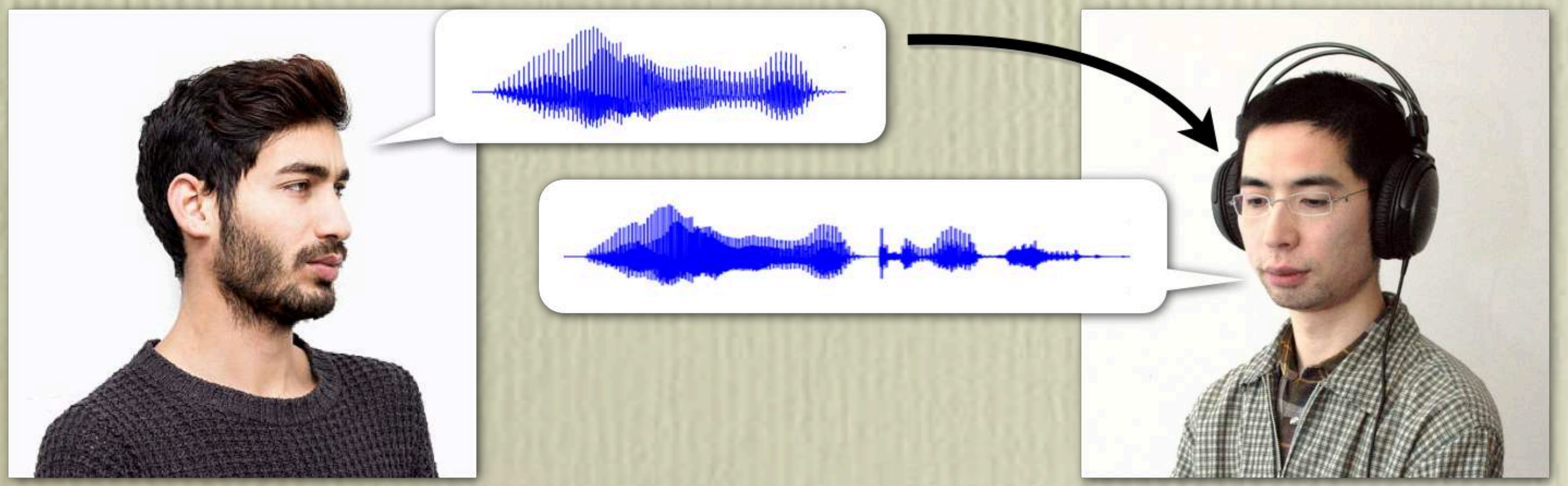
- **Listening, understanding, and speaking running almost together**







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

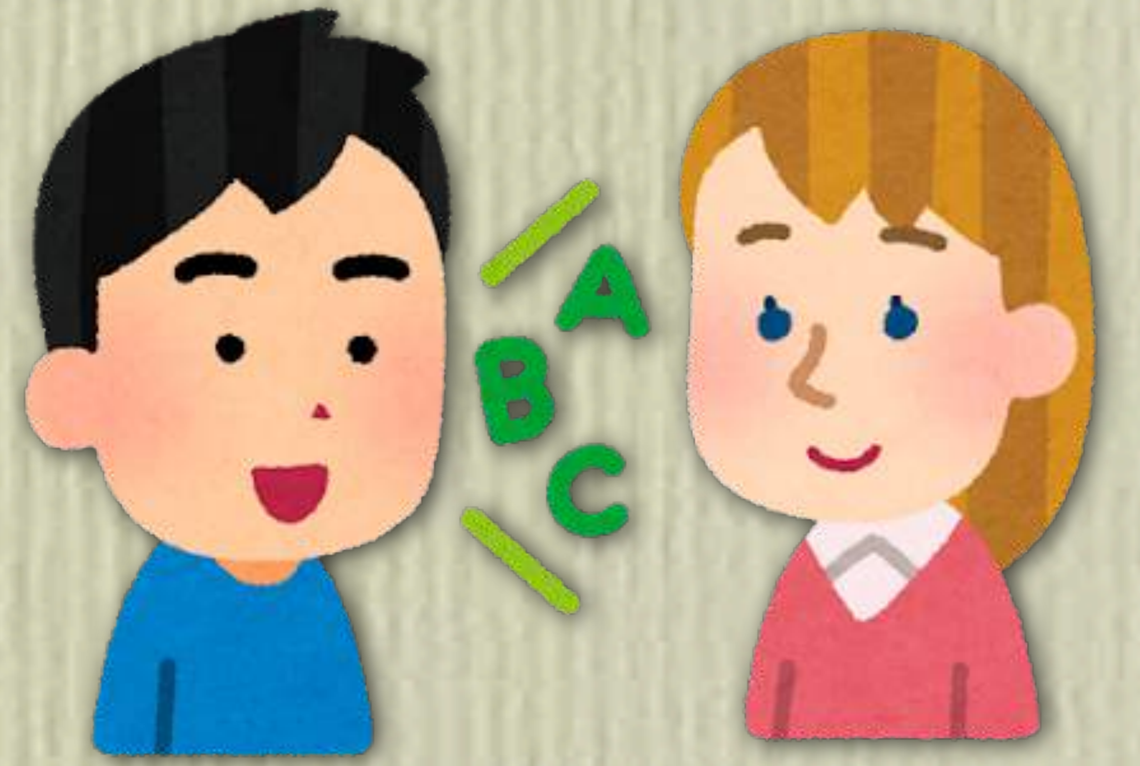


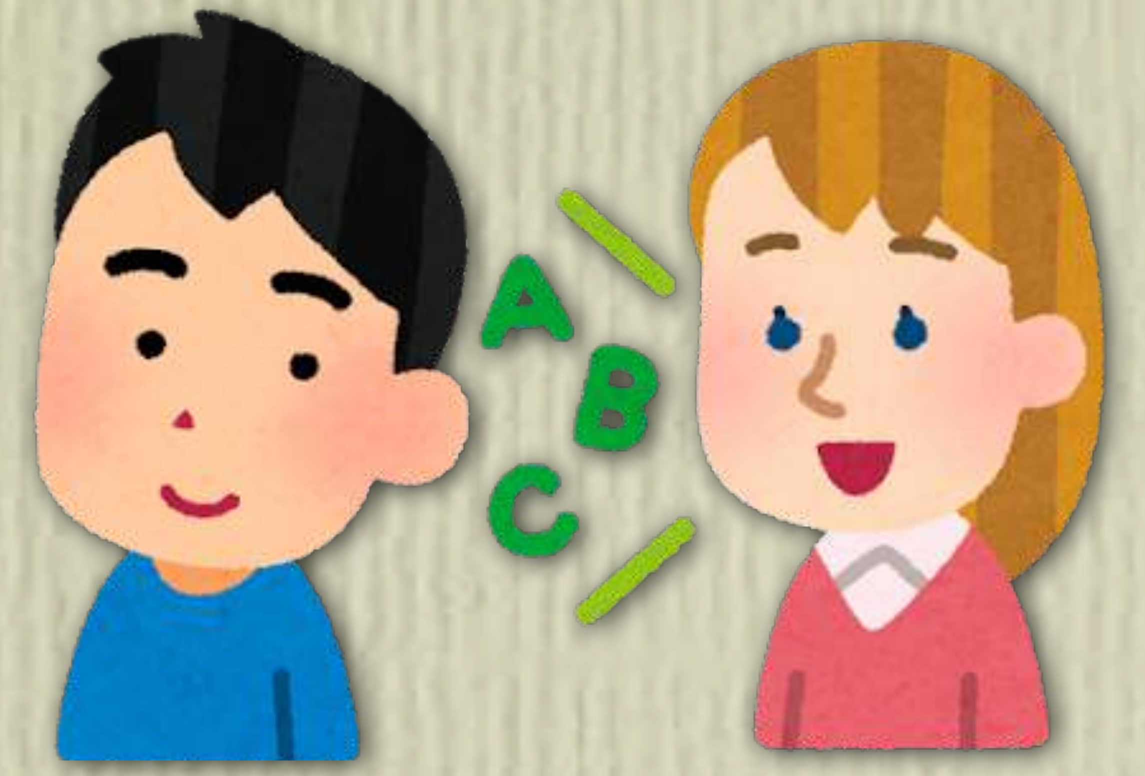
native

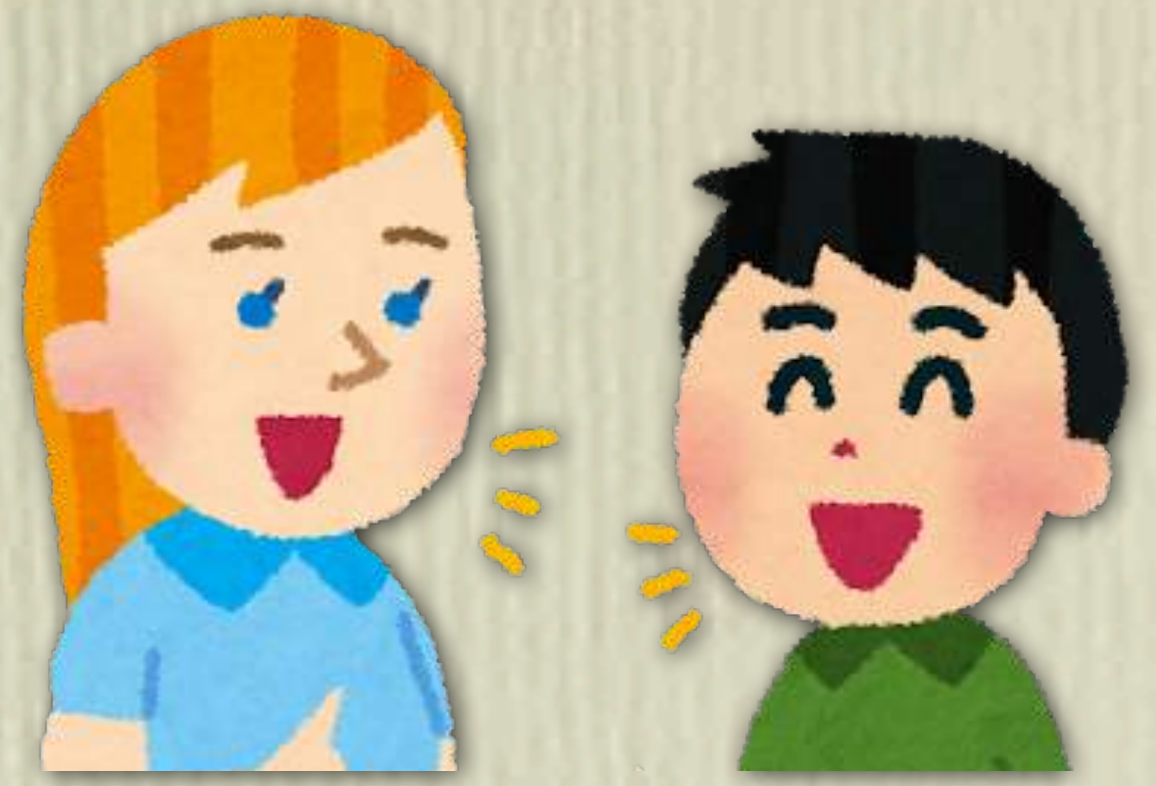
learner

Conversation is a multi-task speech activity.

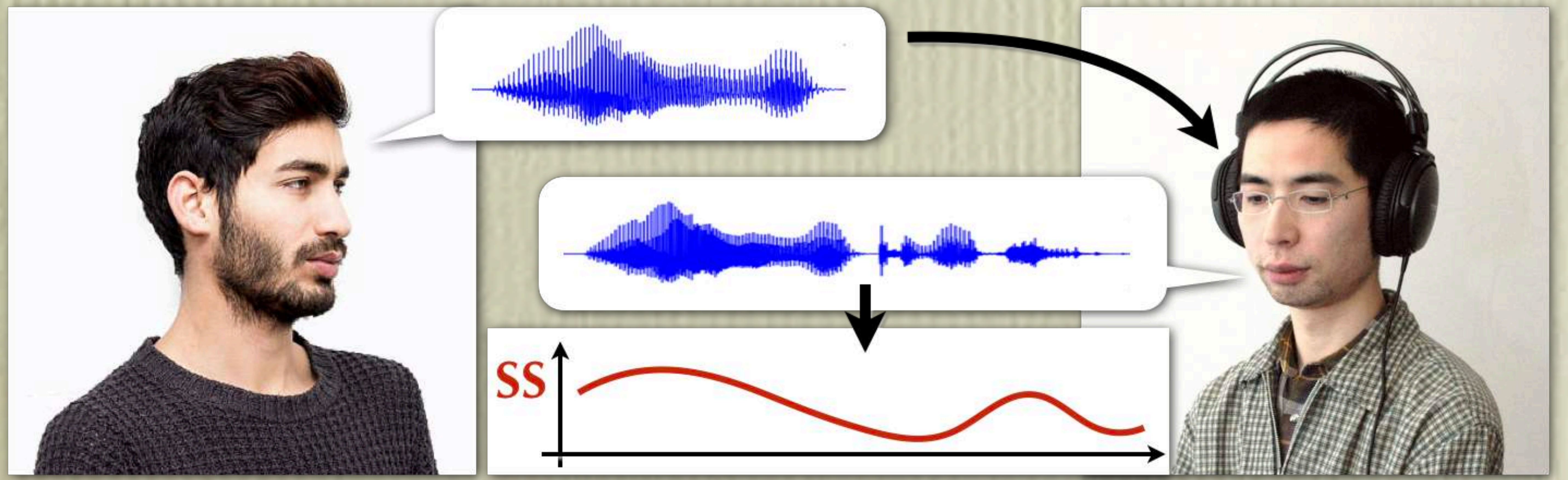
- **Listening, understanding, and speaking running almost together**







- **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

Data collection and teachers' manual rating

- **Collection of samples from 125 university or college students**
  - 4 passages = 55 sentences
  - Four repetitions and 27,500 utterances all together
- **Sentence selection for manual rating**
  - 10 sentences were selected based on syntactic complexity and pronunciation difficulty.
    - 10 sentences = **27 clauses** = 3,375 (= 27 x 125) **clause-based** utterances all together
  - Fourth shadowings were rated manually by a unit of clause.
  - Strategies for rating
    - How correctly phonemes are produced **(P)**.
    - How correctly prosody is produced **(S = Supra-segmental = Prosody).**
    - Whether each word sounds as if it is produced after identifying that word **(C = Correctness)**.
    - 5-step scale (1–5) and the total score **(P+S+C)** varies from 3 to 15.
  - Raters
    - 3 bilingual (AE+J) teachers of English

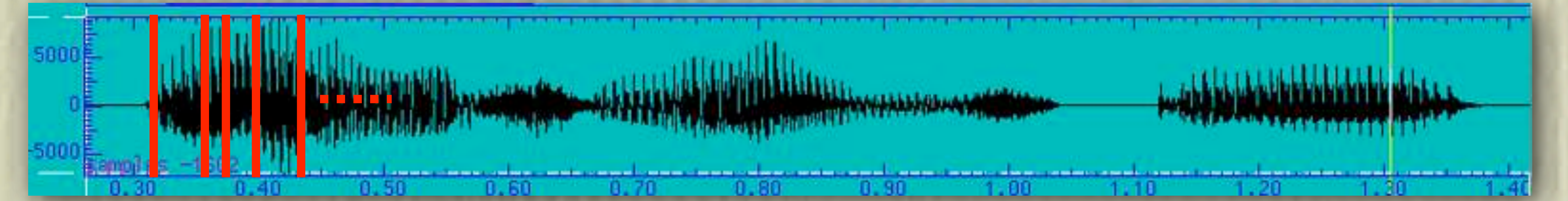
Spectrogram is converted to posteriogram

- **Phoneme posterior probabilities calculated by DNN**
  - A front-end module of current ASR systems.

- **DNN can be viewed as strong abstraction.**
  - Spectrogram is acoustic representation, including extra-linguistic features.
  - Posteriogram is phonetic/phonemic representation, suppressing those features.

DNN-based calculation of GOP

- GOP = Goodness Of Pronunciation



+phonemes intended by the model speaker

time ↓

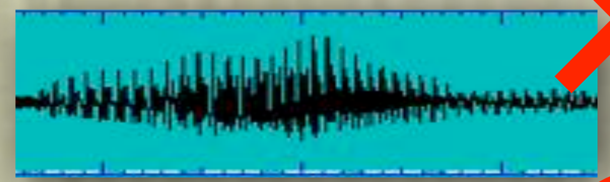
Frame	Phoneme
1	a
2	a
3	u
…	…
1232	sil

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

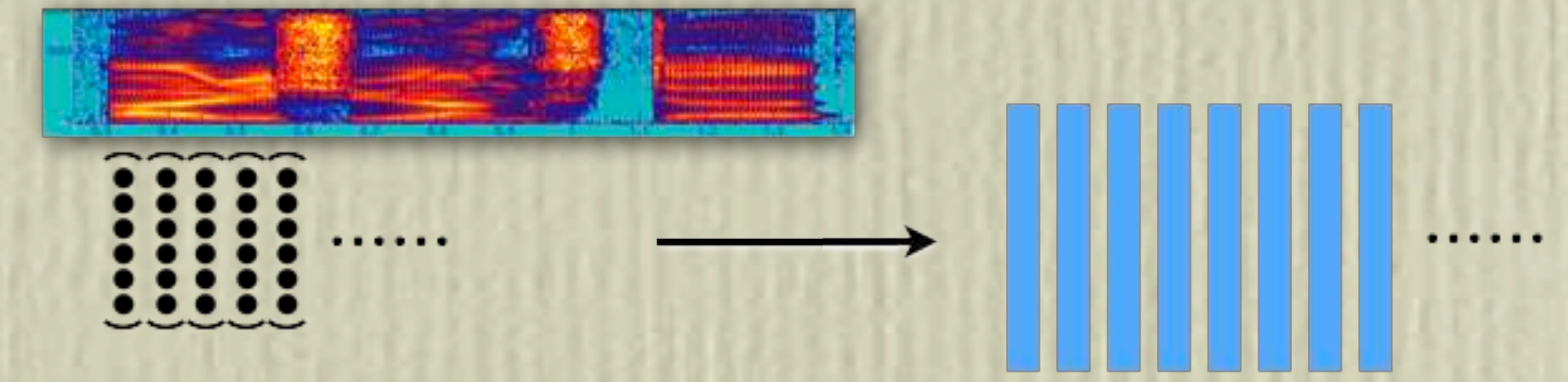
$$\mathrm{GOP} = \frac{0.8 + 0.7 + 0.4 + \cdots + 0.9}{1232} = 0.63$$

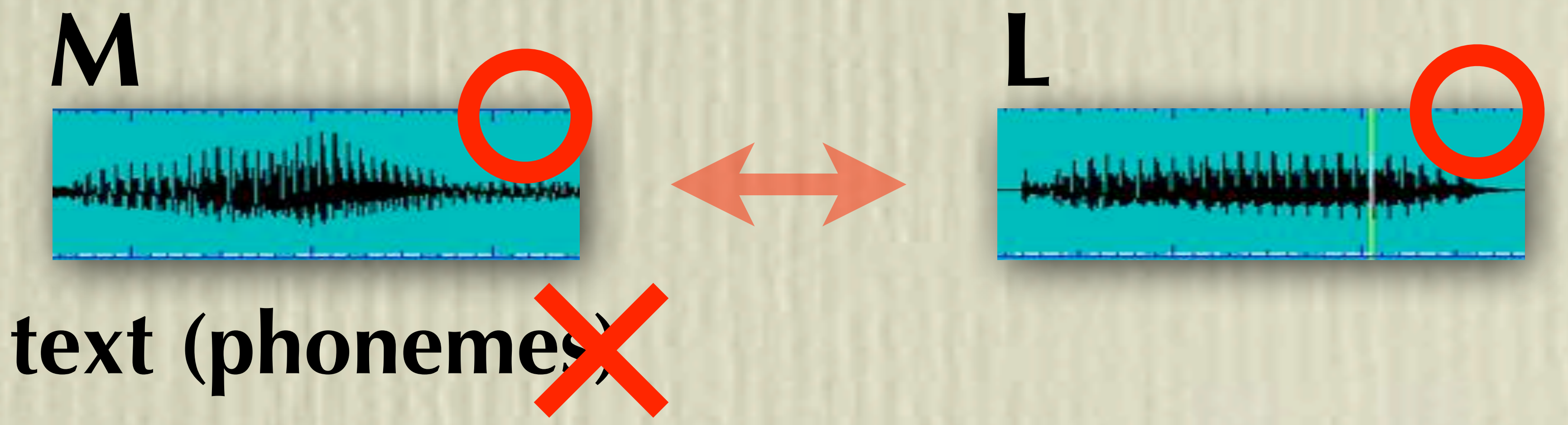
DNN-GOP

Another method for utterance comparison

The two utterances are compared.

- 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

- Clause-based scores are averaged to give sentence-based and speaker-based scores.

Regression model to predict human scores

- Variants of DNN-GOP and some other features or scores are prepared 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

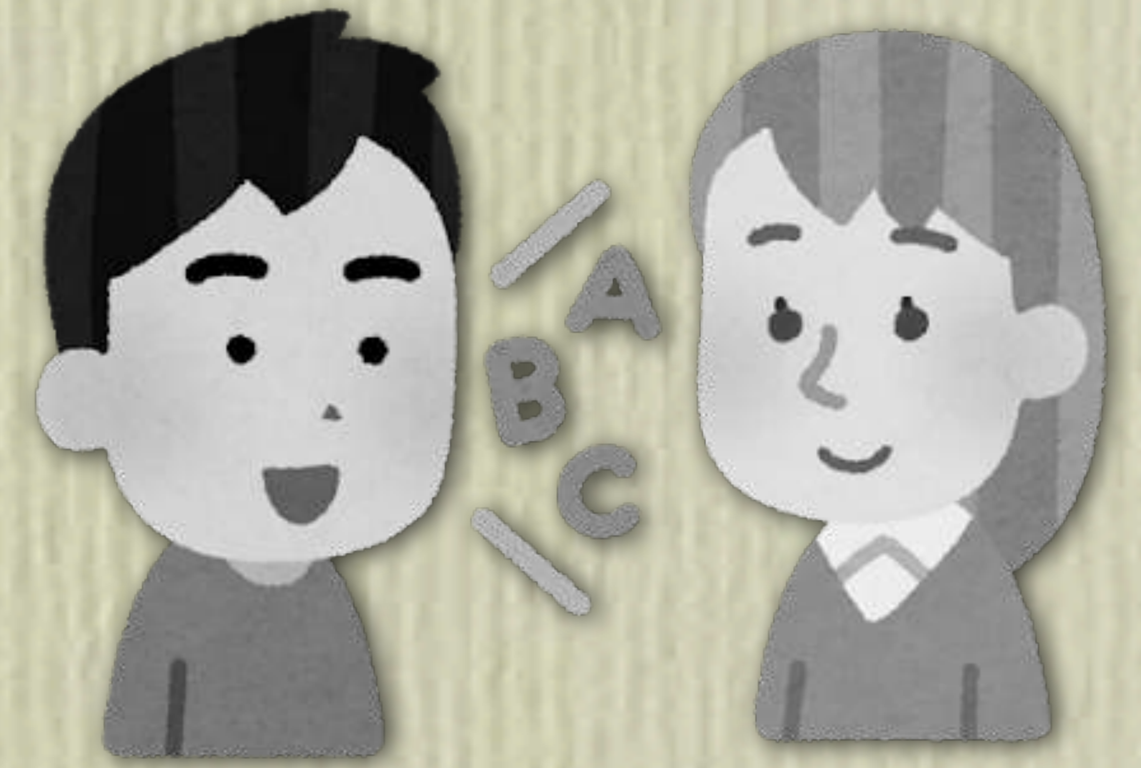
models	P	S	C	P+S+C
Lasso	0.68	0.73	0.65	0.77
SVR	0.70	0.73	0.68	0.78
Random Forest	0.67	0.68	0.61	0.74
inter-rater	0.58	0.54	0.74	0.75

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

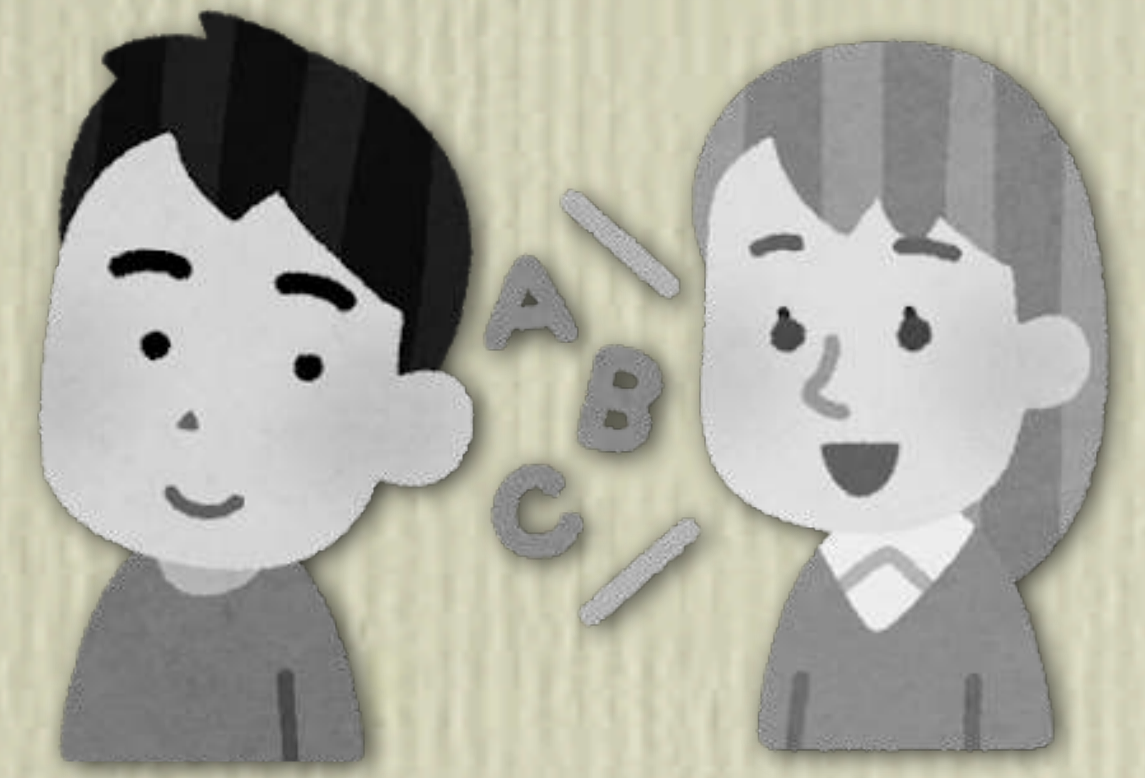
Outline of the presentation

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

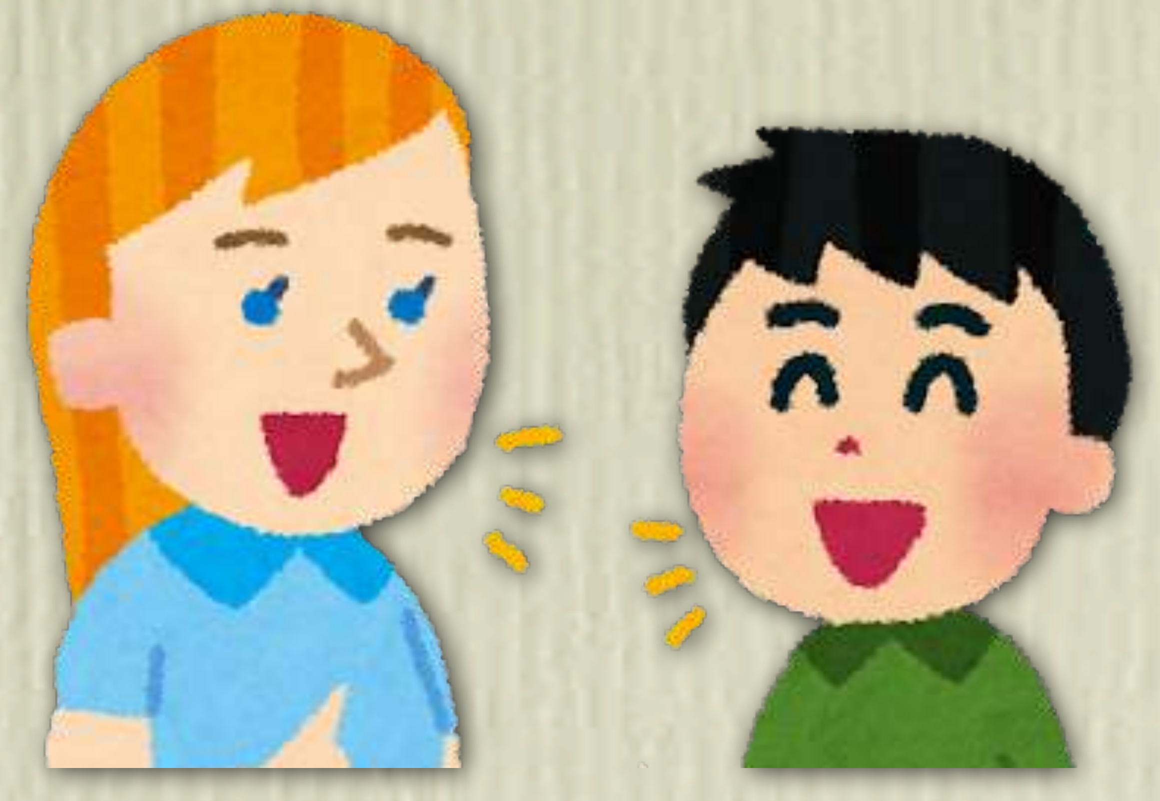
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



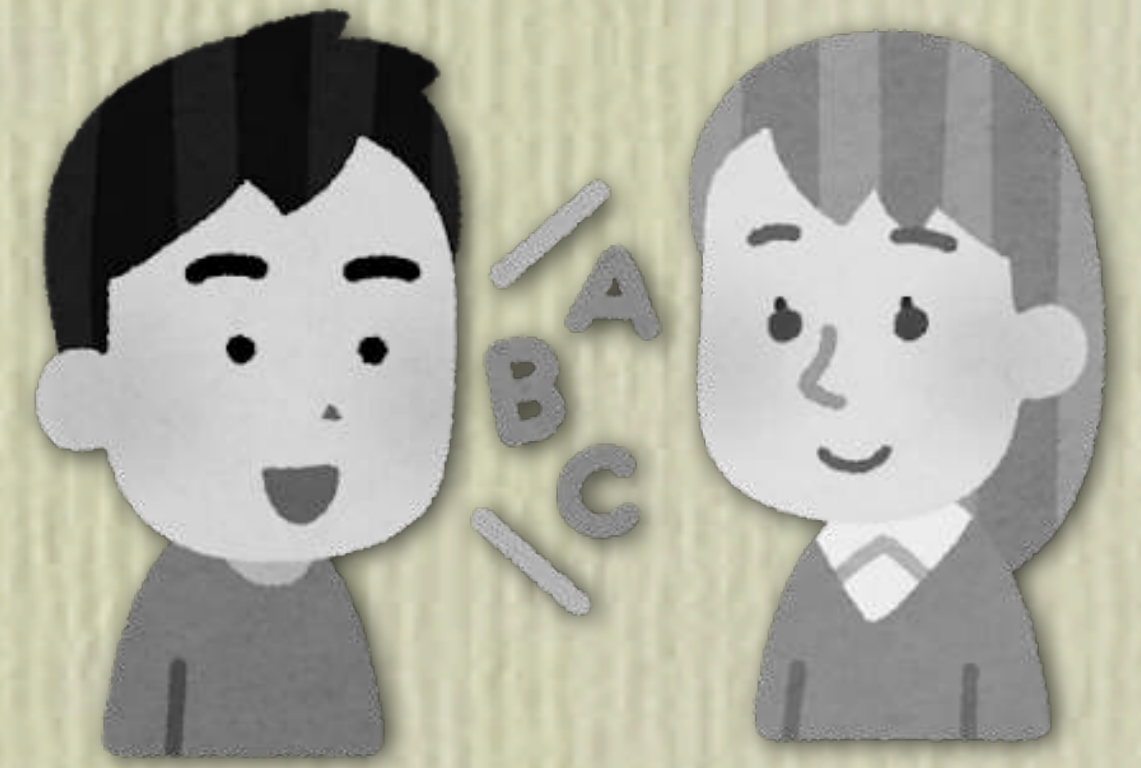
with speech recognition technologies ✔

with new speech technologies being developed in our new project

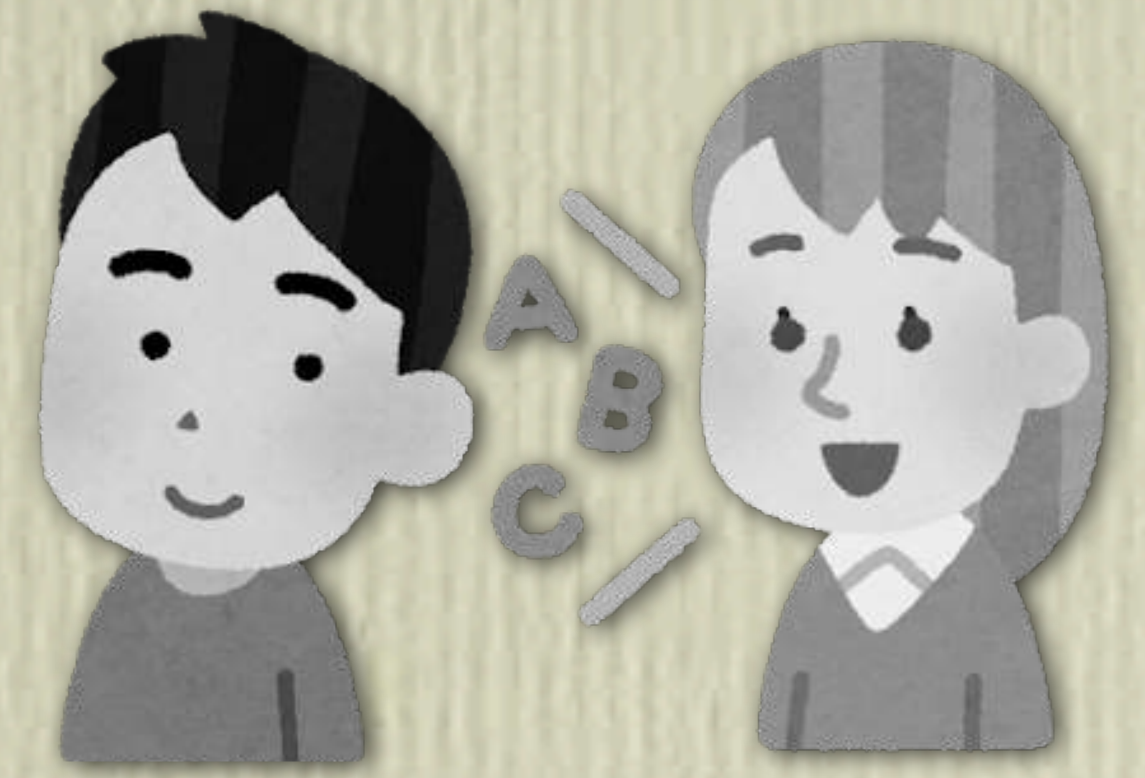
Outline of the presentation

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

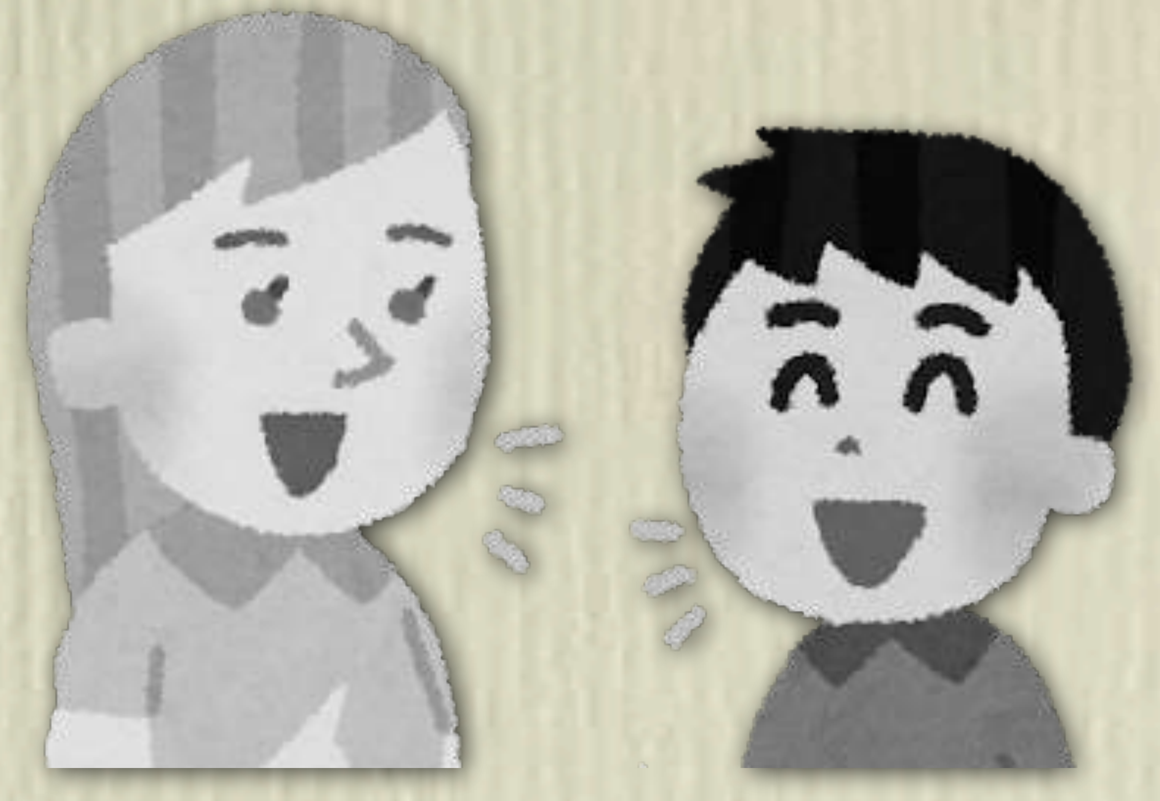
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies

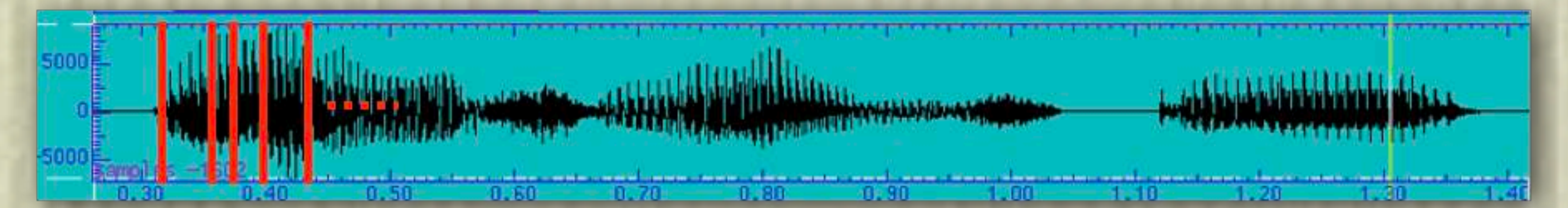


with speech recognition technologies

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

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

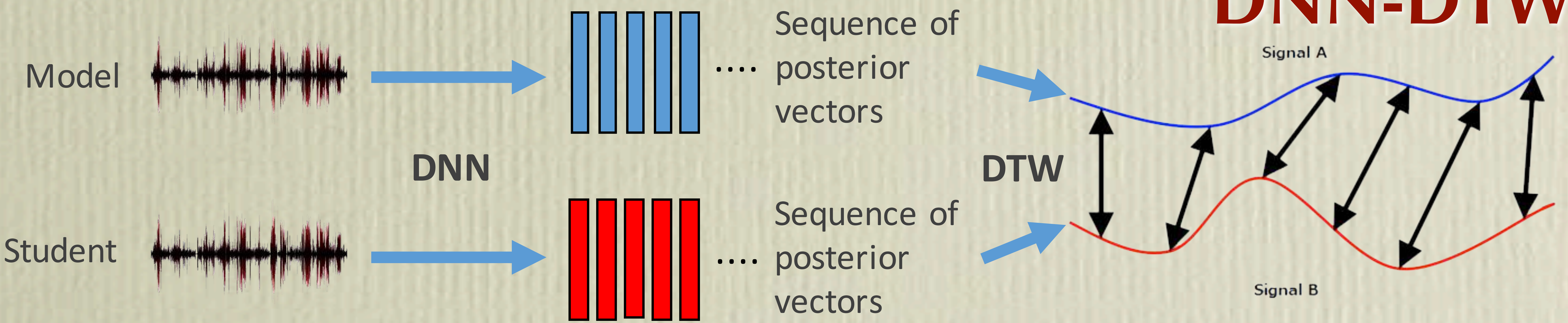
phoneme

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

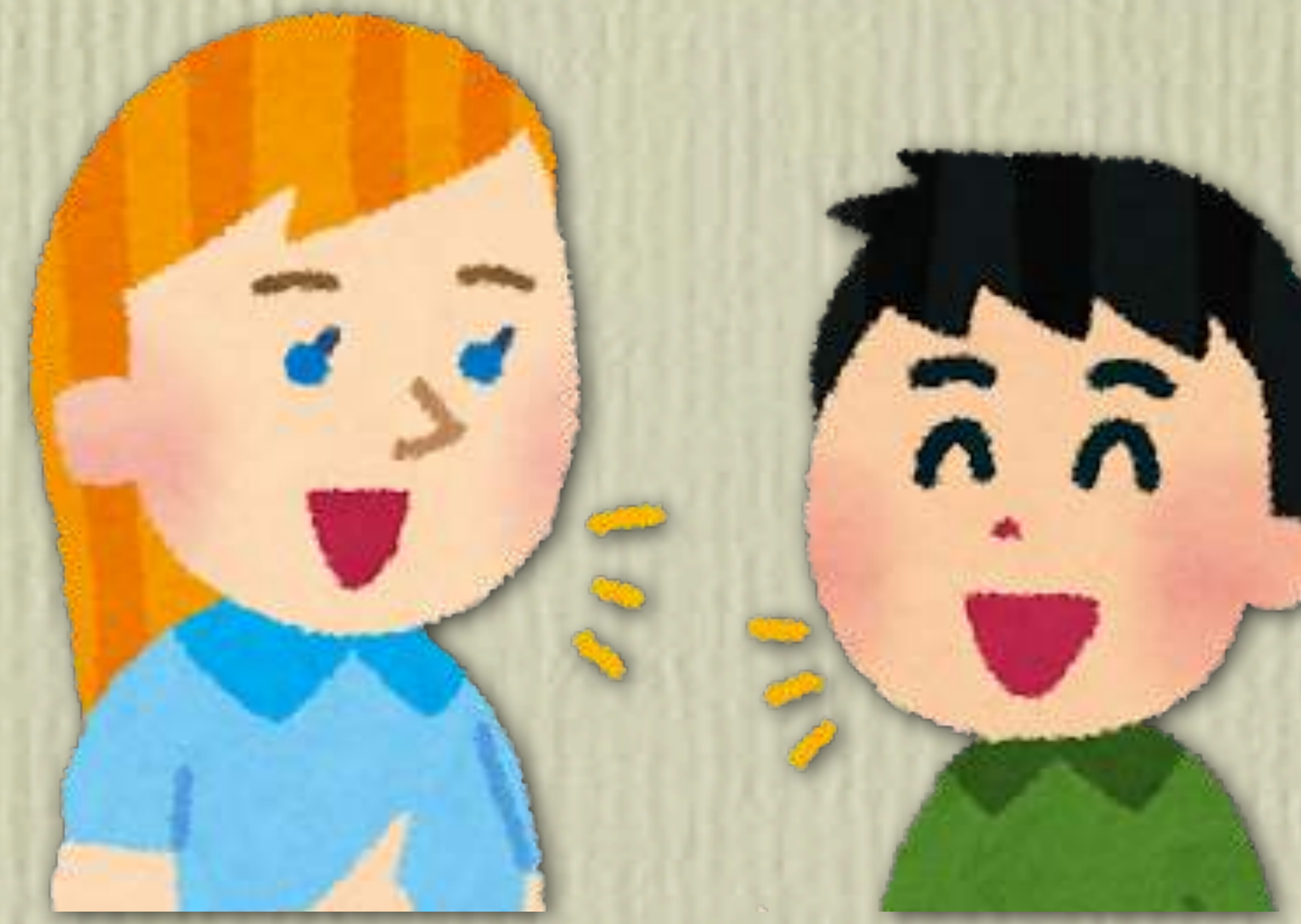
DNN-GOP

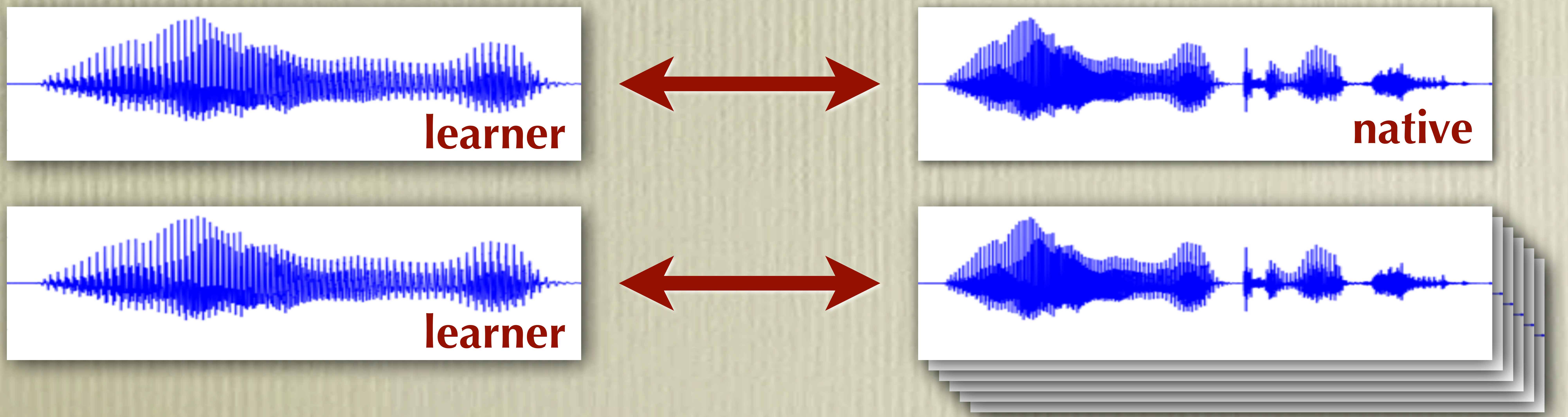
DNN-DTW



Accented but intelligible or comprehensible enough

- What kind of accented pron. is acceptable or not acceptable?

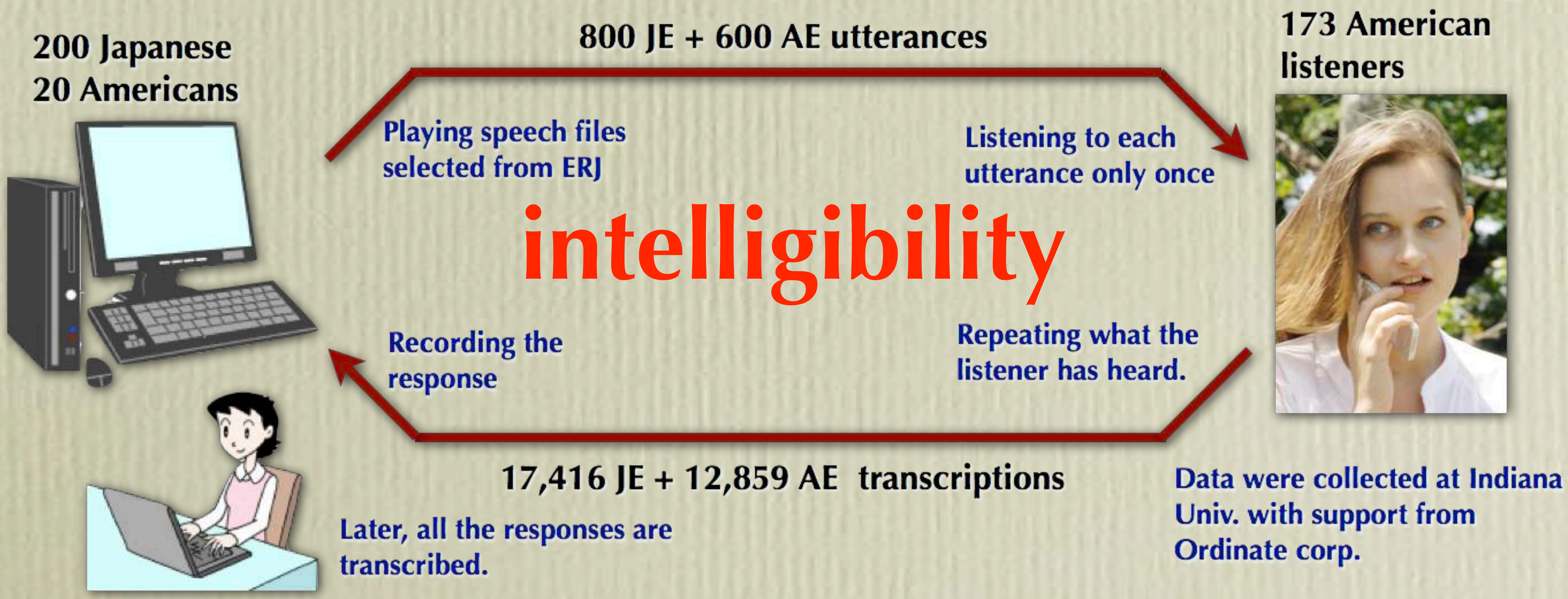




Americans encounter Japanese English for the fist time.

How intelligible is JE to Americans with no exposure to JE? [Minematsu+'11]

- Japanese and Americans (GA) read aloud sentences written by native speakers.
- JE (100 males and 100 females, 800 utt.) + AE (10 males and 10 females, 600 utt.)
  - All the recordings were judged as correct by the speakers themselves.
- Other 173 Americans listened to each utterance only once and repeat it.
  - Topic and speaker always varied from utterance to utterance.
- Technical staff transcribed carefully the repetitions (20 repetitions on avg. / utterance).



How correctly were JE repeated by Americans?

"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
- 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 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 was retracted with an apology
- the misquote was retracted with an apology
- the misquote 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
- instructed with an apology
- an apology
- someone was distracted with an apology
- is destructed apologize
- [Q]
- uh something was obstructed and needs an apology [N]
- distracted with an apology
- sammy's co:at was obstructed with apology
- @ is extracted with an apology
- @ was instructed with an apology
- the m(isquote)- mister was di- distracted with an apology
- attracted with # *polar
- the misquote was constructed with an apology #
- sammy was instructed with a apology [N]
- someone was instructed with an apology

goo.gl/jUAehX

How correctly were JE repeated by Americans?

“T … tracted with an apology.”

$P_I(w \mid o)$

$P_S(w \mid o)$

goo.gl/jUAehX

How correctly were JE repeated by Americans?

"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
- 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 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 was retracted with an apology
- the misquote was retracted with an apology
- the misquote was retracted with an apology

$P_l(w | o)$

- 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
- someone was distracted with an apology
- is destructed apologize
- [Q]
- uh something was obstructed and needs an apology [N]
- distracted with an apology
- sammy's co:at was obstructed with apology
- @ is extracted with an apology
- @ was instructed with an apology
- the m(isquote)- mister was di- distracted with an apology
- attracted with # *polar
- the misquote was constructed with an apology #
- sammy was instructed with a apology [N]
- someone was instructed with an apology

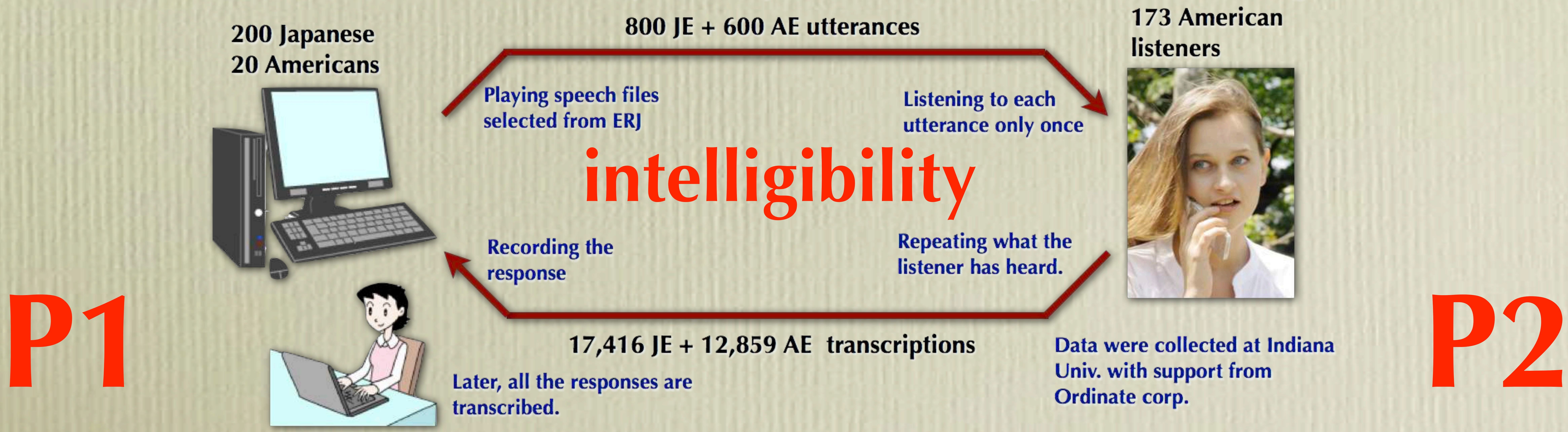
$P_s(w | o)$

goo.gl/jUAehX

Americans encounter Japanese English for the fist time.

How intelligible is JE to Americans with no exposure to JE? [Minematsu+'11]

- Japanese and Americans (GA) read aloud sentences written by native speakers.
- JE (100 males and 100 females, 800 utt.) + AE (10 males and 10 females, 600 utt.)
  - All the recordings were judged as correct by the speakers themselves.
- Other 173 Americans listened to each utterance only once and repeat it.
  - Topic and speaker always varied from utterance to utterance.
- Technical staff transcribed carefully the repetitions (20 repetitions on avg. / utterance).



Intelligibility and comprehensibility

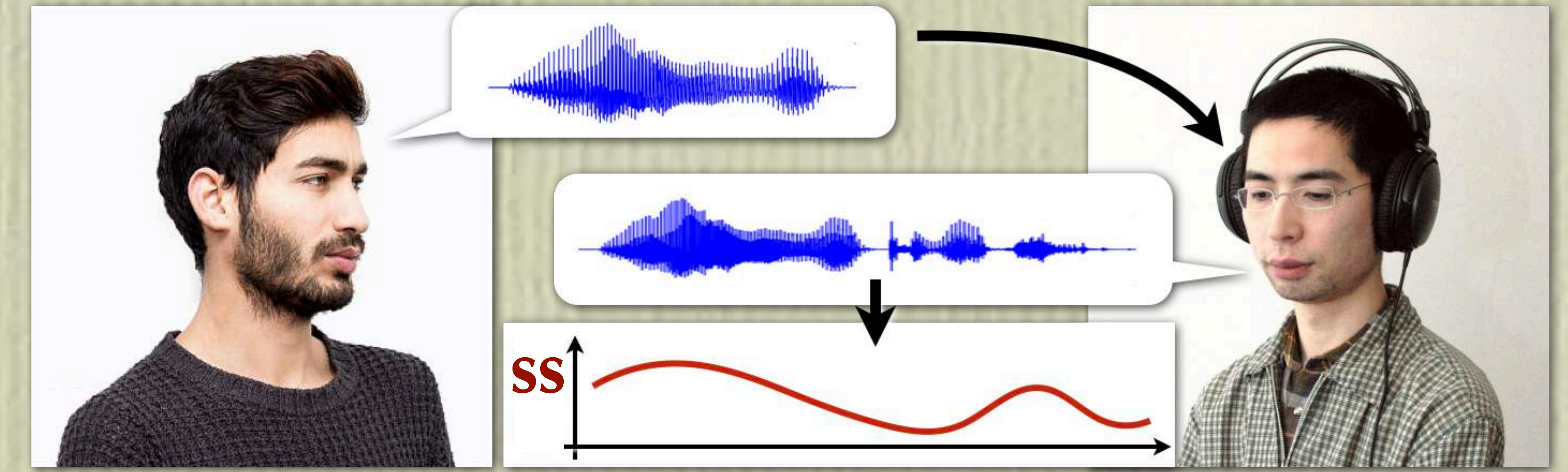
Intelligible/comprehensible enough pronunciations [Derwing+'09]

- Intelligibility
  - How many words in a given utterance can be identified correctly?
  - Measured *objectively* by native listeners' transcription or oral repetition.
  - Focuses on the results of listeners' recognition process. ➙ offline
- Comprehensibility
  - How easily, i.e., how smoothly, the content of a given utterance can be understood?
  - Measured *objectively* by monitoring brain activities or size of pupils
  - Focuses on how the recognition process is running. ➙ online

Both metrics are strongly related to tolerance and lenience of listeners.

Shadowing = repeating without waiting and deep guessing

General form of shadowing

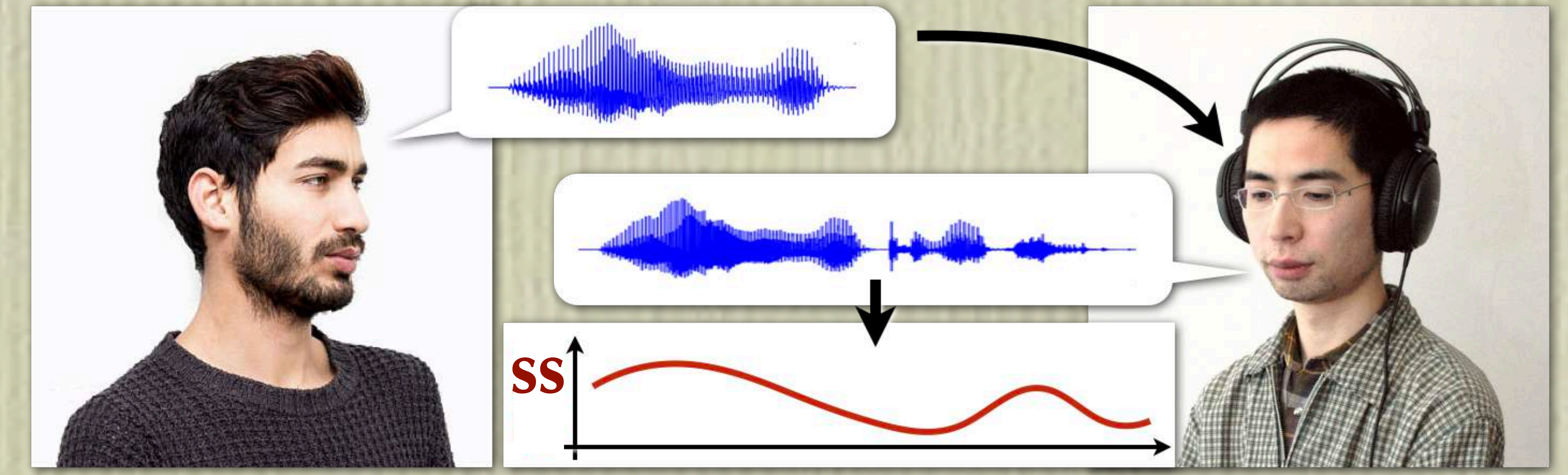


native ss=smoothness of shadow learner

- Objective measurement of SS (shadowability)
  - Luo+2009, Luo+2010, Kato+2012, Yamauchi+2014, Shi+2016, Yue+2017, Kabashima+2018

Shadowing = repeating without waiting and deep guessing

General form of shadowing

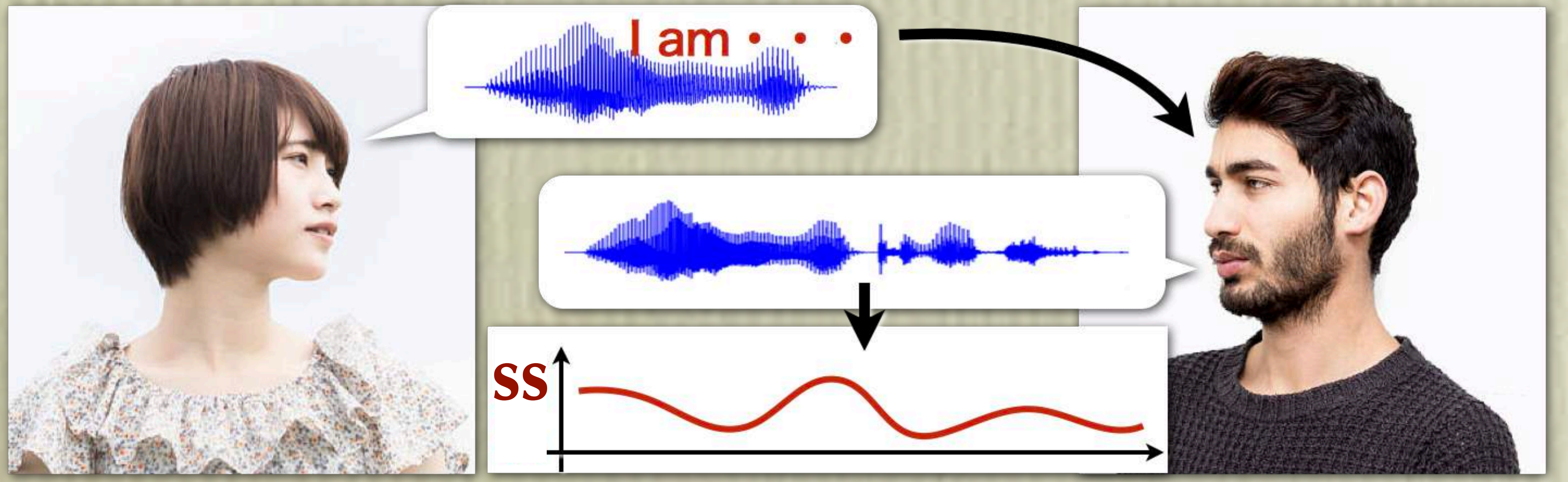


native ss=smoothness of shadow learner

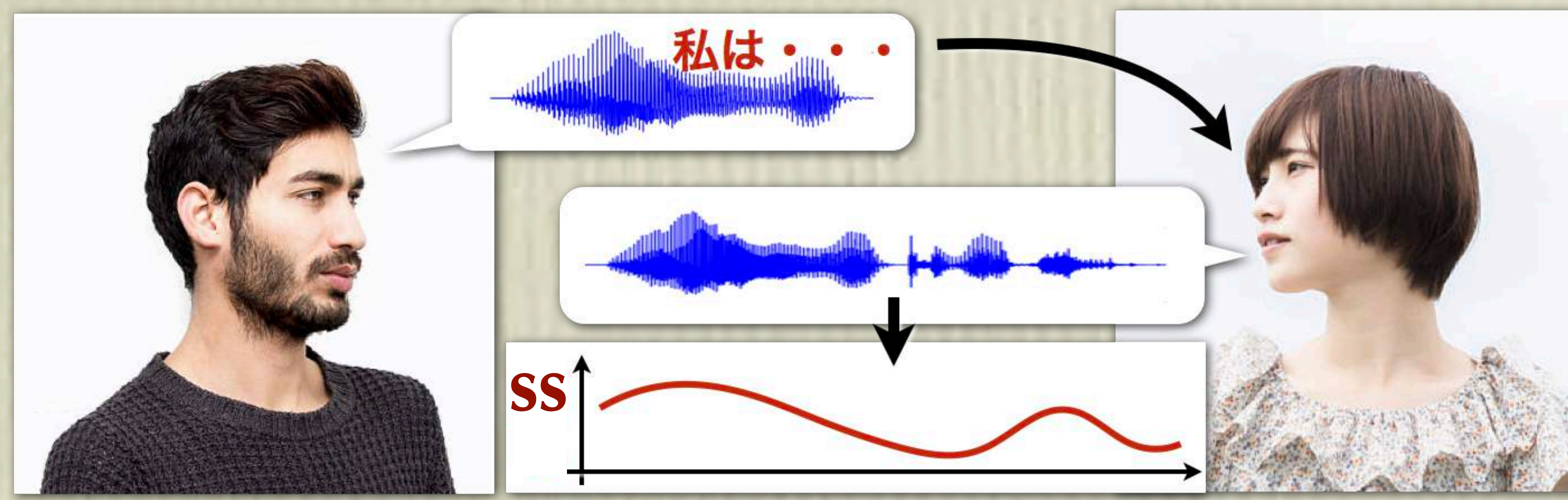
- Objective measurement of SS (shadowability
  - Luo+2009, Luo+2010, Kato+2012, Yamauchi+2014, Shi+2016, Yue+2017, Kabashima+2018

Shadowing = simultaneous reproduction of words intended by the learner, in a native pronunciation, not imitation of accented pronunciations.

Proposed (inverse) form of shadowing



learner native



learner native

Collection of non-native and native Karaoke readings

L1 = Vietnamese and L2 = Japanese

- Slow utterances are easy to shadow.
- Speaking rate control was introduced to recording.

Karaoke-style recording with speaking rate controlled

- Model utterances from the CD of a textbook are used as reference.
- Color of the text changes according to the speaking rate of the model utterances.

Examples of readings and shadowings

Texts and model utterances from CD

Read by Vietnamese

Read by Japanese

Utterances read by Vietnamese

Utterances read by Japanese

Shadowed by Japanese listeners

Shadowing utterances by Japanese listeners

Shadowing utterances by Japanese listeners

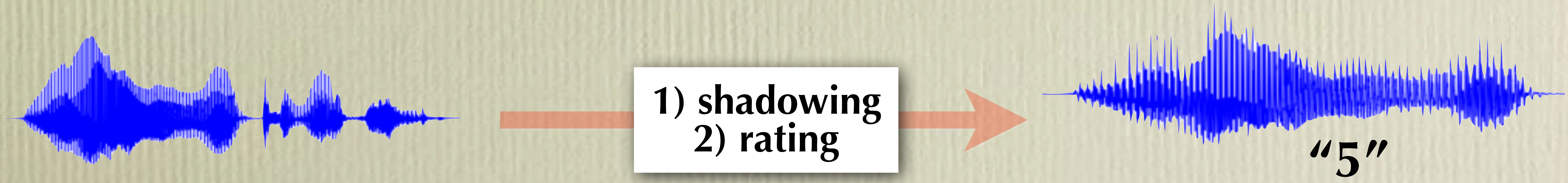
Conditions of the experiments

Non-native and native speakers and their Karaoke-style readings

- Vietnamese learners of Japanese (INT x 3, ADV x 3) **VJ**
- Native speakers of Japanese x 6 **NJ**
- 164 phrase utterances from the CD of the textbook ➛ 96 **VJ** + 68 **NJ**

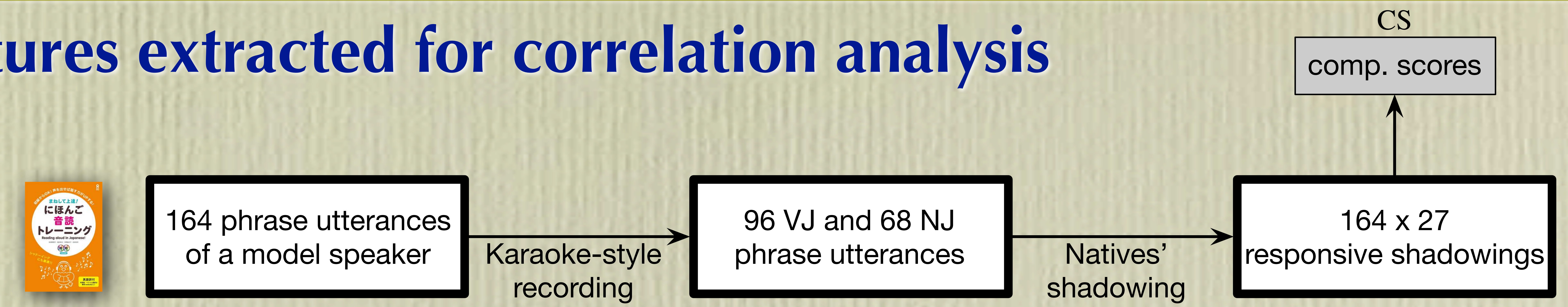
Native shadowers and their tasks

- 27 native speakers of Japanese with normal hearing
- Asked to reproduce in a native pronunciation what was heard as simultaneously as possible, not to imitate accented pronunciations.
- Asked to rate comprehensibility of a given phrase utterance using a 7-degree scale.



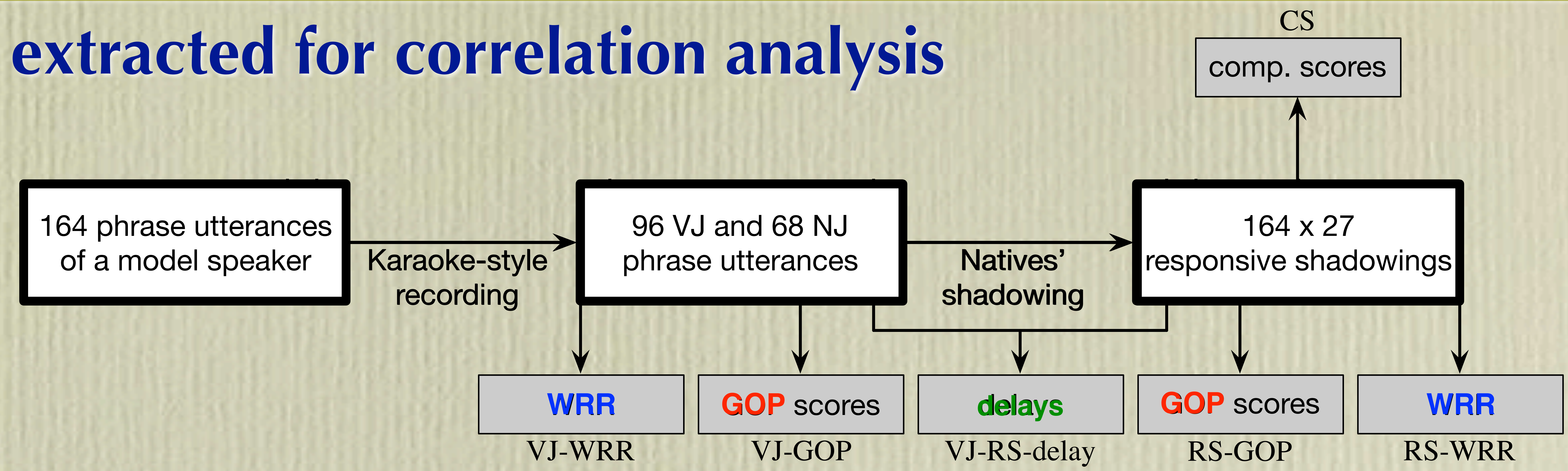
Experiments

Features extracted for correlation analysis



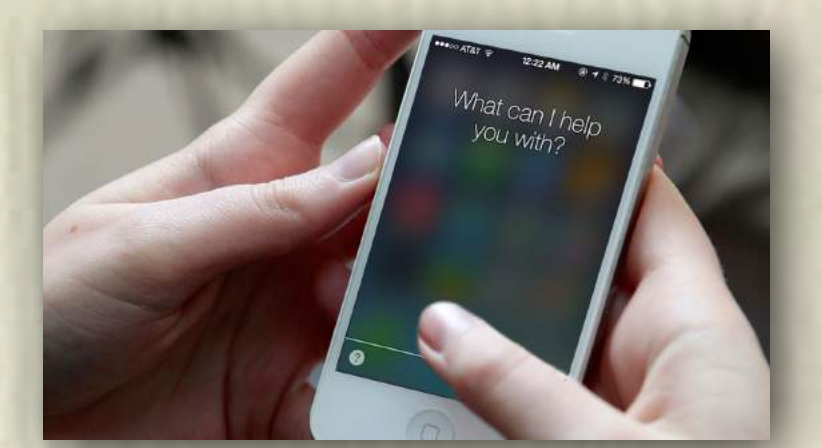
Experiments

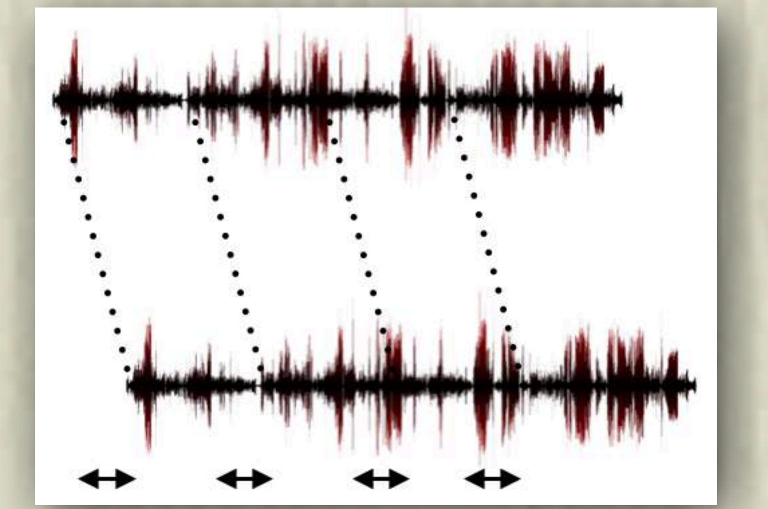
Features extracted for correlation analysis



MS = Model Speech, RS = Responsive Shadowing

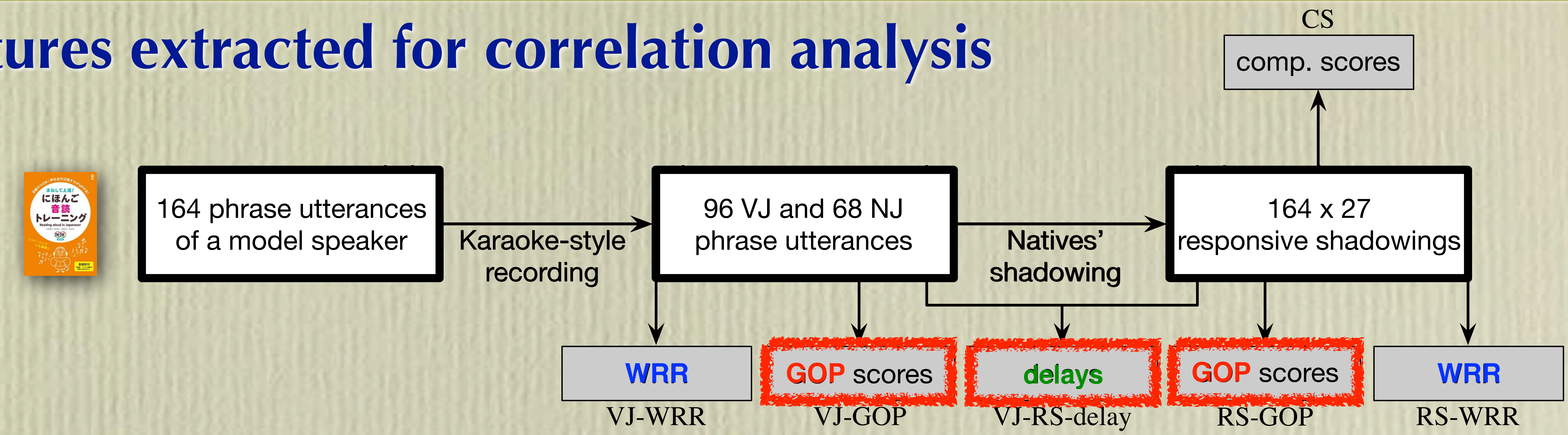
- GOP (Goodness of Pronunciation) [Yue+2017]
  - Accuracy of articulation, calculated as phoneme-based posterior probabilities
- WRR (Word-based Recognition Rate)
  - Performance of the ASR system that is used as baseline system in the Japanese ASR community
- Delay of shadowing
  - Calculated as averaged phoneme boundary gap between VJ and RS





Experiments

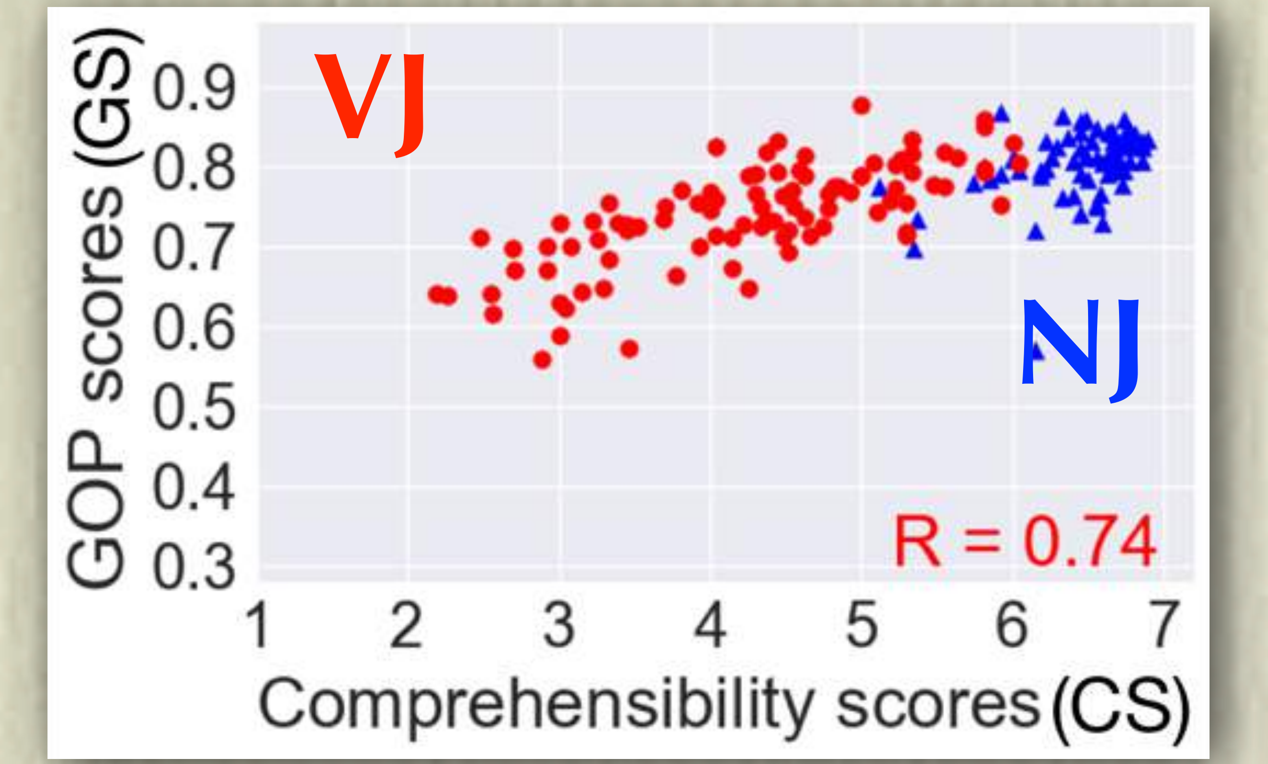
Features extracted for correlation analysis



MS = Model Speech, RS = Responsive Shadowing

Results of correlation analysis between the features and CS

VJ-GOP	VJ-WRR	
0.58	0.47	
RS-GOP	RS-WRR	VJ-RS-delay
0.74	0.53	-0.59



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

Examples of readings and shadowings

Texts and model utterances from CD

Read by Vietnamese

Read by Japanese

Utterances read by Vietnamese

Utterances read by Japanese

Shadowed by Japanese listeners

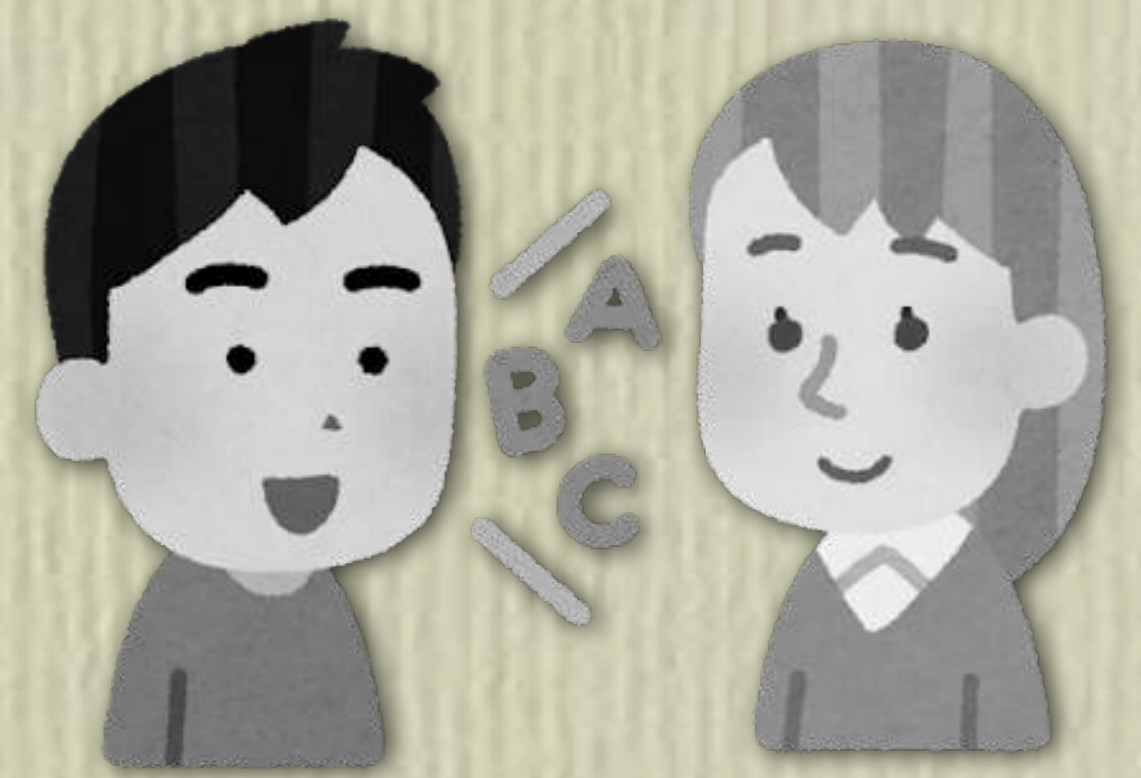
Shad nces
by J ers

Shadowing utterances by Japanese listeners

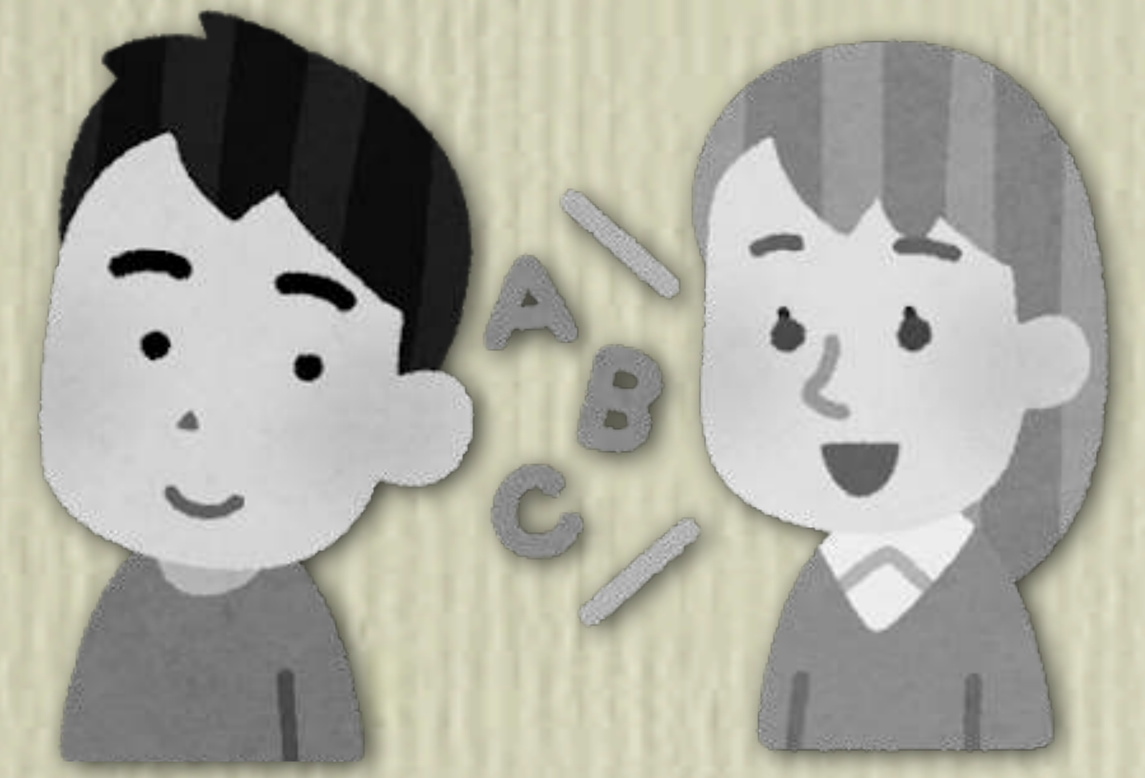
Outline of the presentation

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

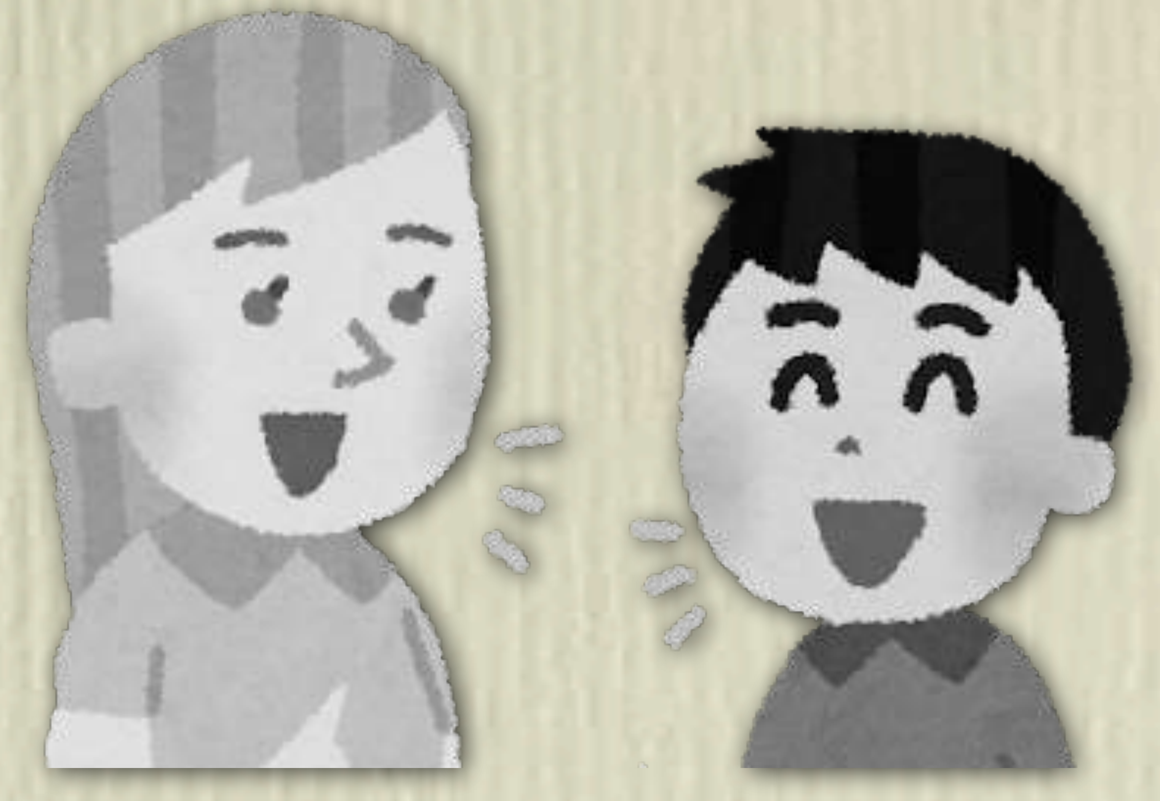
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies



with speech recognition technologies

with new speech technologies being developed in our new project ✓

Out poster at AAAL2019

INTER-LEARNER SHADOWING WITH SPEECH TECHNOLOGIES ENABLES AUTOMATIC AND OBJECTIVE MEASUREMENT OF COMPREHENSIBILITY OF LEARNERS' UTTERANCES

Nobuaki Minematsu*, Yusuke Inoue*, Daisuke Saito*, Yutaka Yamauchi**, Kumi Kanamura*** (*UTokyo, **Soka Univ., ***Nagoya Univ. Economics)

RESEARCH QUESTION AND RELATED WORKS

Two different goals of pronunciation training
- Native-like vs. intelligible/comprehensible enough pronunciations
- Intelligibility and comprehensibility [Derwing+'09]
  - I: How many words are correctly identified in given L2 utterances?
    → Focuses on results of listeners' process of understanding → **offline**
  - C: How easily or smoothly given L2 utterances are understood?
    → Focuses on how the understanding process is running while listening → **online**

How to calculate intelligibility/comprehensibility of L2 speech?
- **Subjective** assessment
  - "How many words do you think you identified correctly?" 100%, 80%, 60%, 50%,,,,,
  - "How easily did you understand the message?" Very easily, easily, rather easily,,,,,
- **Objective** assessment
  - L2 utterances were transcribed by native speakers **after** hearing them. Objective intelligibility = word-level correct transcription rate [Bernstein'03]
  - Listeners' behaviors are observed using physiological sensors. **Listening effort** observed using EEG (Electroencephalogram) [Song+'18] **Cognitive load** observed as the size of pupils using an eye-tracker [Govender+'18] Both methods are too expensive to be used in classrooms.

\+ =

Q: How to calculate comprehensibility in an inexpensive way?
- Objective assessment on intelligibility [Bernstein'03, Minematsu+'11]
  - Transcription or **oral repetition** is done **after** hearing L2 utterances.
  - Transcription or repetition allows waiting and guessing. → **offline**
- Objective assessment on comprehensibility using **modified** repetition [Inoue+'18]
  - Repetition with almost no waiting or guessing → **shadowing!!!**
  - Native listeners' **reverse** shadowing of L2 utterances!!! → **online**

Conventional form of shadowing — Proposed form of shadowing

I will visit・ I will visit・

SS = Smoothness of Shadowing → GOP
- Learners' SS is automatically predicted. [Luo+'10][Yue+'17][Kabashima+'18]

NATIVE LISTENERS' REVERSE SHADOWING

Recording of L2 read sentences or phrases
- Vietnamese learners of Japanese (L1= , L2=)
  - 6 intermediate and 6 advanced learners
  - Read sentences were also collected from natives as reference.
- Sentences were selected from an intermediate-level textbook.
  - Using CD audios of the textbook, Karaoke-style recording was done with speaking rate controlled. よろしくおねがいいたします。
  - All the learners and native speakers read aloud 164 phrases from the textbook.

Native listeners' reverse shadowing [Inoue+'18]
- 27 native shadowers shadowed 96 VJ read phrases and 68 NJ read phrases.
- After each shadowing, subjective rating of comprehensibility was done. (→ CS)

Overview of experiments examining native listeners' reverse shadowing

164 phrase utterances of a model speaker → Karaoke-style recording → 96 VJ and 68 NJ phrase utterances → Natives' shadowing → 164 x 27 responsive shadowings → CS → comp. scores
WRR (VJ-WRR), GOP scores (VJ-GOP), delays (VJ-RS-delay), GOP scores (RS-GOP), WRR (RS-WRR)

Correlation analysis between CS and speech features
- WRR: Word Recognition Rate, calculated using a speech recognition system
- GOP: Goodness Of Pronunciation, obtained as posterior probabilities of the phonemes intended by learners calculated using a speech recognition module.
  - Widely used in CALL as baseline speech feature for pron. assessment.
- Delay of shadowing: average phoneme boundary gap bet. VJ and RS

Corr. of the speech features to CS

VJ-GOP	VJ-WRR	
0.58	0.47	
RS-GOP	RS-WRR	VJ-RS-delay
0.74	0.53	-0.59

- RS-GOP >> VJ-GOP
- RS-WRR > VJ-WRR
- Native listeners' reverse shadowing is by far more informative than learners' speech to predict its comprehensibility.

Pros and cons of native listeners' reverse shadowing
- Pros: inexpensive and directly focusing on listeners' immediate impressions
- Cons: always requires native listeners as shadowers, which may need payment.
  → Can be solved by inter-learner shadowing?

INTER-LEARNER SHADOWING

SS = 60
SS = "B+"
Logical feedback

Intuitive feedback

← utterances ←······ shadowings

ILS = A learner shadows others and is shadowed by others.
- Any learner is a native of a language and shadows others learning that language.
  - Inter-supportive framework among learners of different languages.
  - A shadowing version of Lang-8 (http://lang-8.com).

ILS can feedback "honest" impressions on L2 utterances.
- Many teachers tend to be generous (and politically correct) to learners.
- Generous (politically correct) shadowing is difficult and probably impossible to do.
  - Waiting and guessing is not allowed in shadowing.

ILS may be able to solve the problem of "lack of exposure".
- Many learners do not have good chances to be exposed to native listeners.
- Pronunciation of any learner is the most comprehensible pronunciation to him/her.
  - Any learner is the most generous assessor of his/her own pronunciation.
  - Difficult to guess how smoothly others can or cannot listen to him/her.

ILS may be able to generate more "intuitive" feedback.
- Natives' shadowability can be fed back **logically** using scores or statements.
- Foreign accented utterances with the same level of shadowability can be fed back to learners, which can work as much more **intuitive** feedback.
  - Chinese-accented Japanese utterances are used to explain how his pronunciation is viewed from native speakers in terms of comprehensibility (shadowability).
  - So intuitive that accented speech feedback might discourage learners.

ILS requires collaboration and support from learners and teachers!!
- We may start soon a huge collection of L2 utterances and natives' shadowings.
  - Please give us your name card to keep you updated!!!
- Natives' shadowings can be viewed as **spoken annotation of comprehensibility**.
  - Any AI-based speech technology requires a huge corpus with annotation.

Examples of natives' *transcription* of AE and JE
- Speakers: natives and Japanese learners with low proficiency
- Transcribes: natives who have **never** talked with Japanese
- Each of read sentences was presented only once.

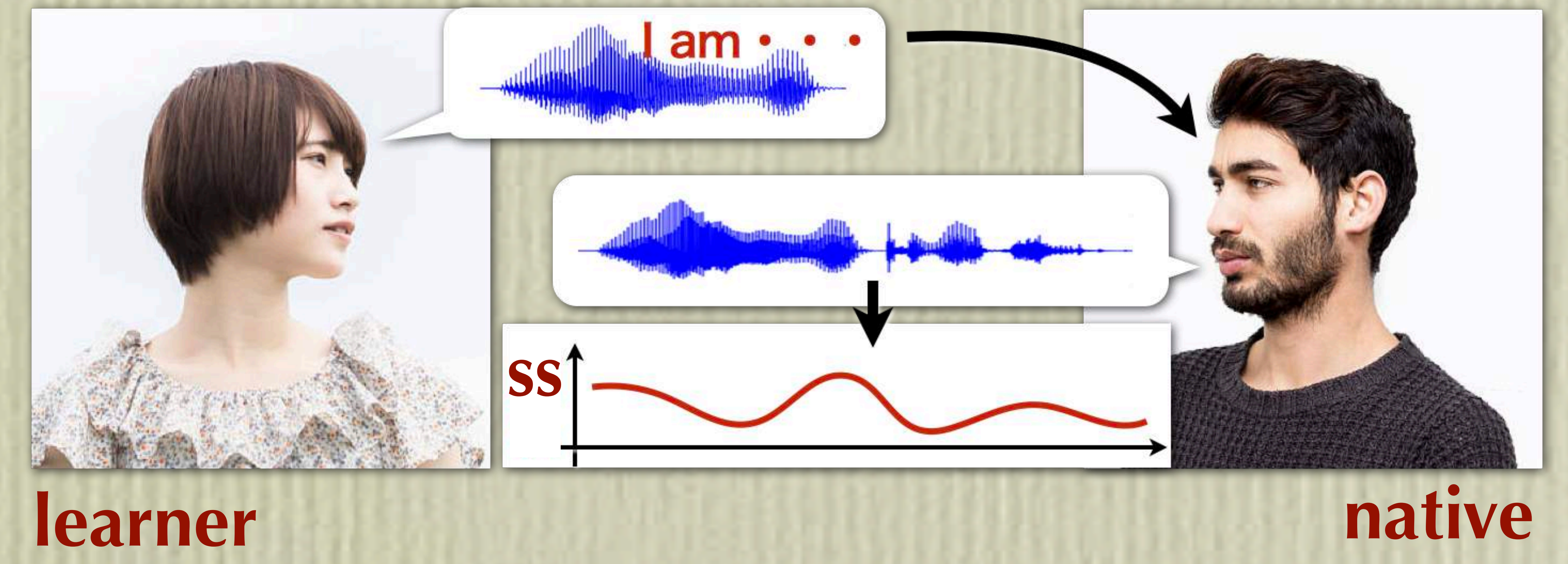
NATIVE — PH_121
NATIVE — PH_151
NATIVE — PH_301

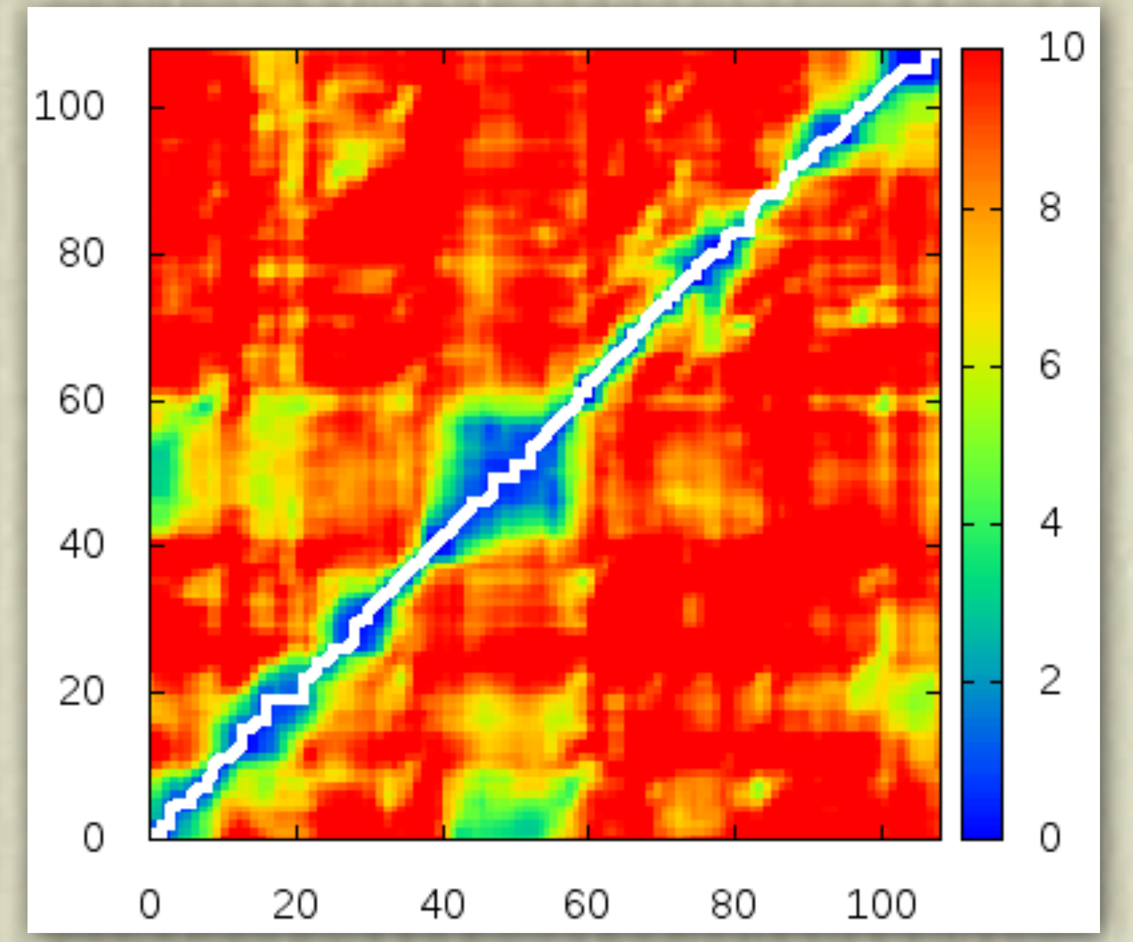
- More examples of transcription are found at [QR] with .

A simpler approach to calculate smoothness of shadowing

Proposed (inverse) form of shadowing

- DNN-GOP is calculated at every frame, that can be viewed as **spoken annotation**!!
  - Annotations (labels) should be collected with simpler and more reliable techniques.



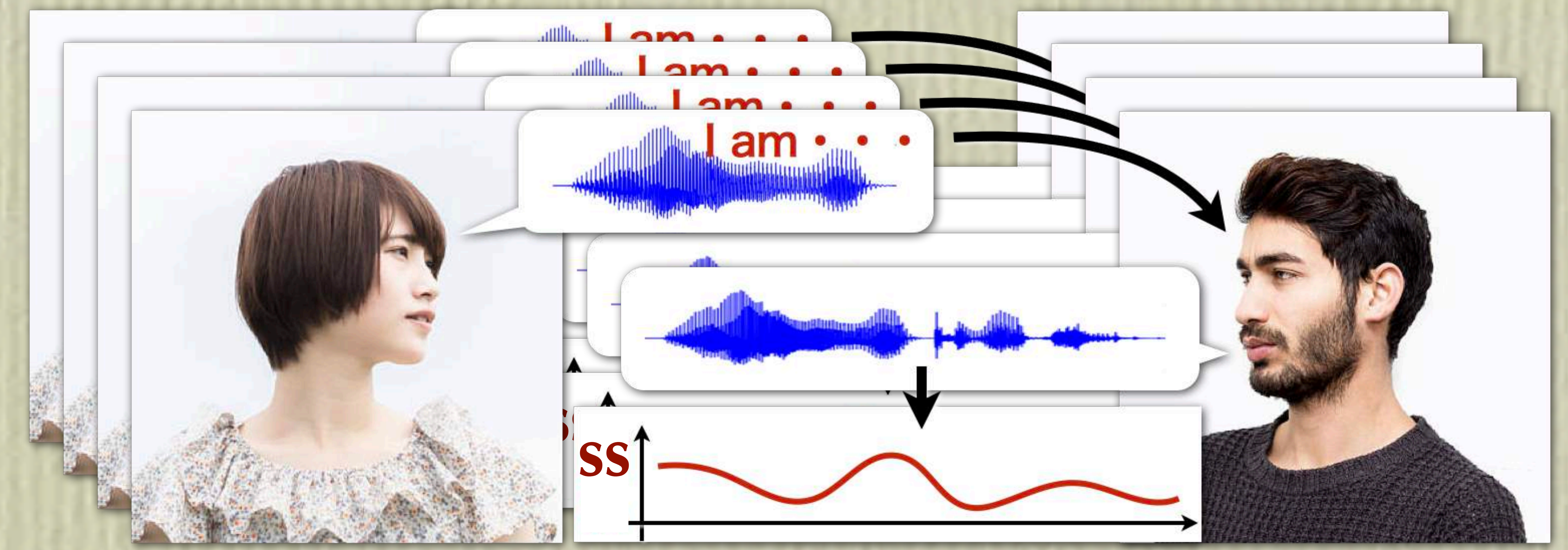


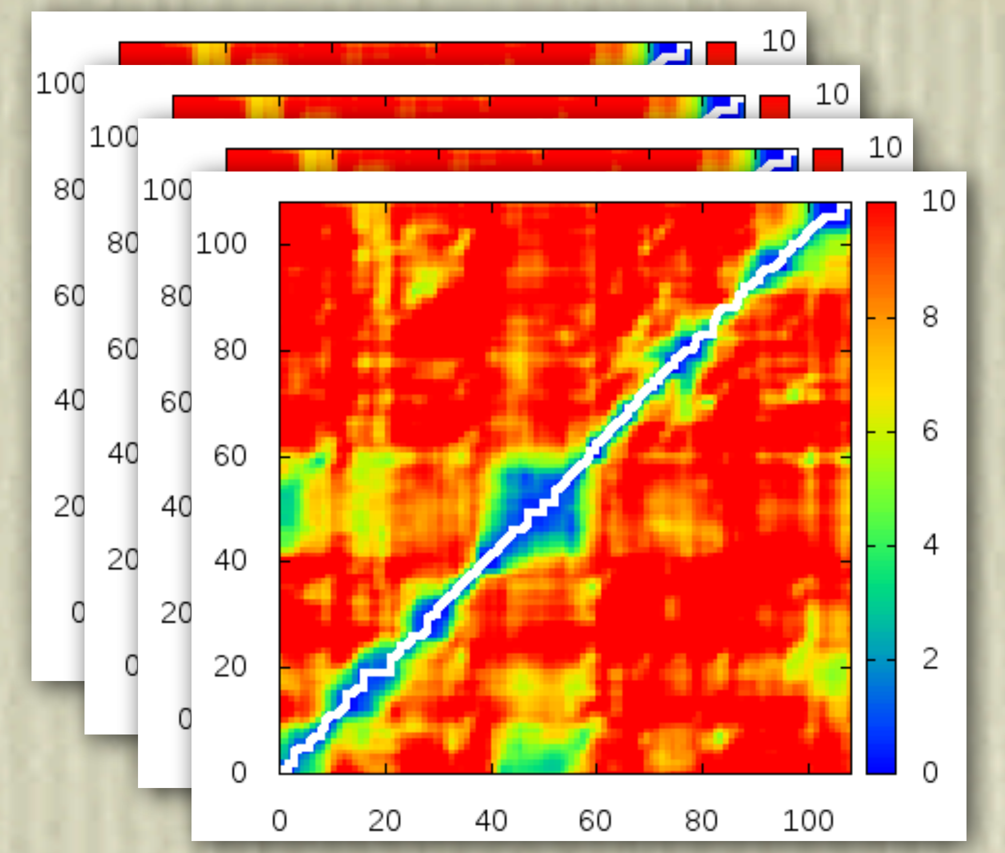
A much simpler and reliable approach

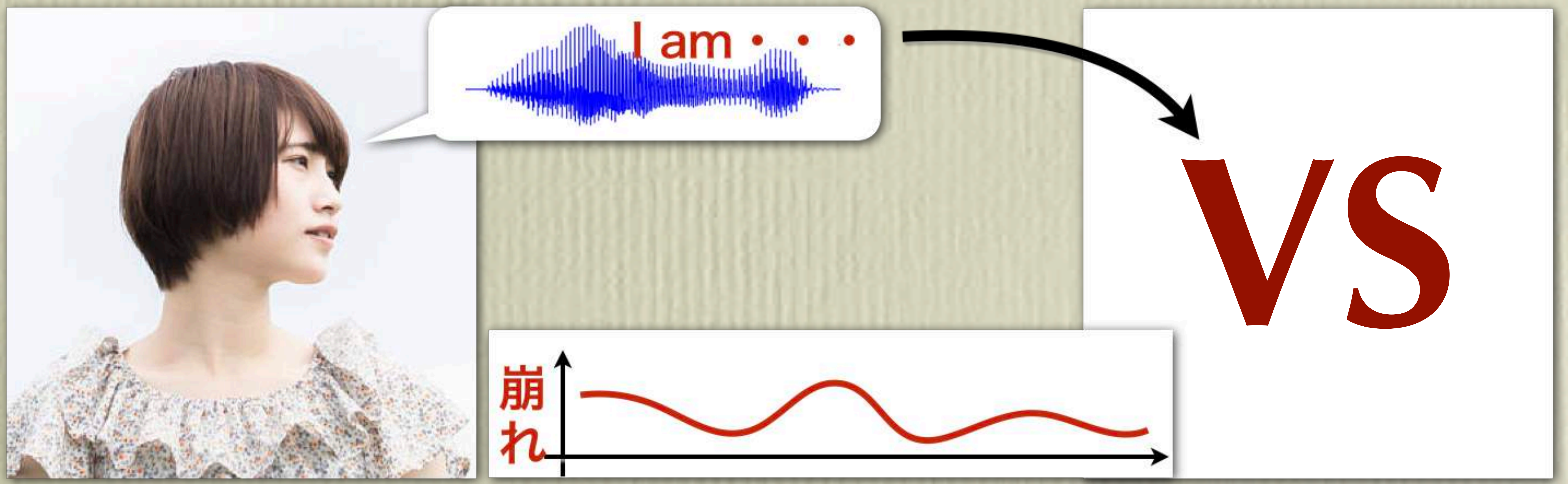
- A native listener is asked to ***shadow*** an utterance given from a learner.
- The shadower is asked to ***read*** the text that was read by that learner.
  - Read speech = **most prepared** speech, shadowed speech = **least prepared** (hastened) speech
- DTW between the two speeches will give us a sequence of smoothness of shadowing.

What will be possible with a huge amount of data?

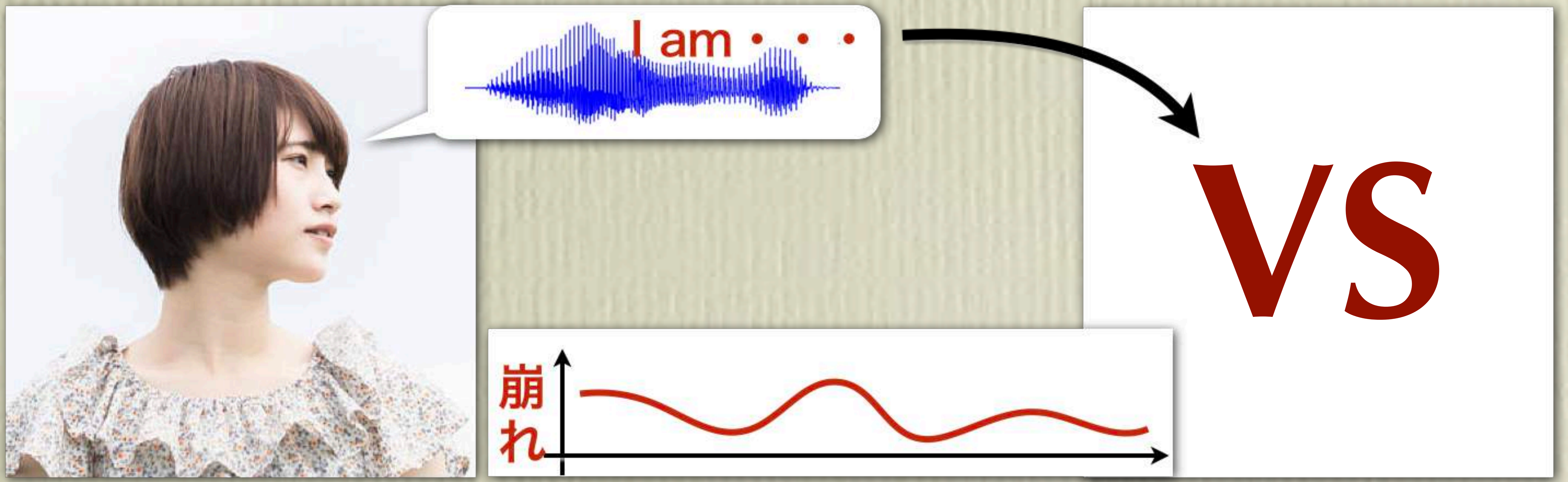
L2 utterances with spoken annotations







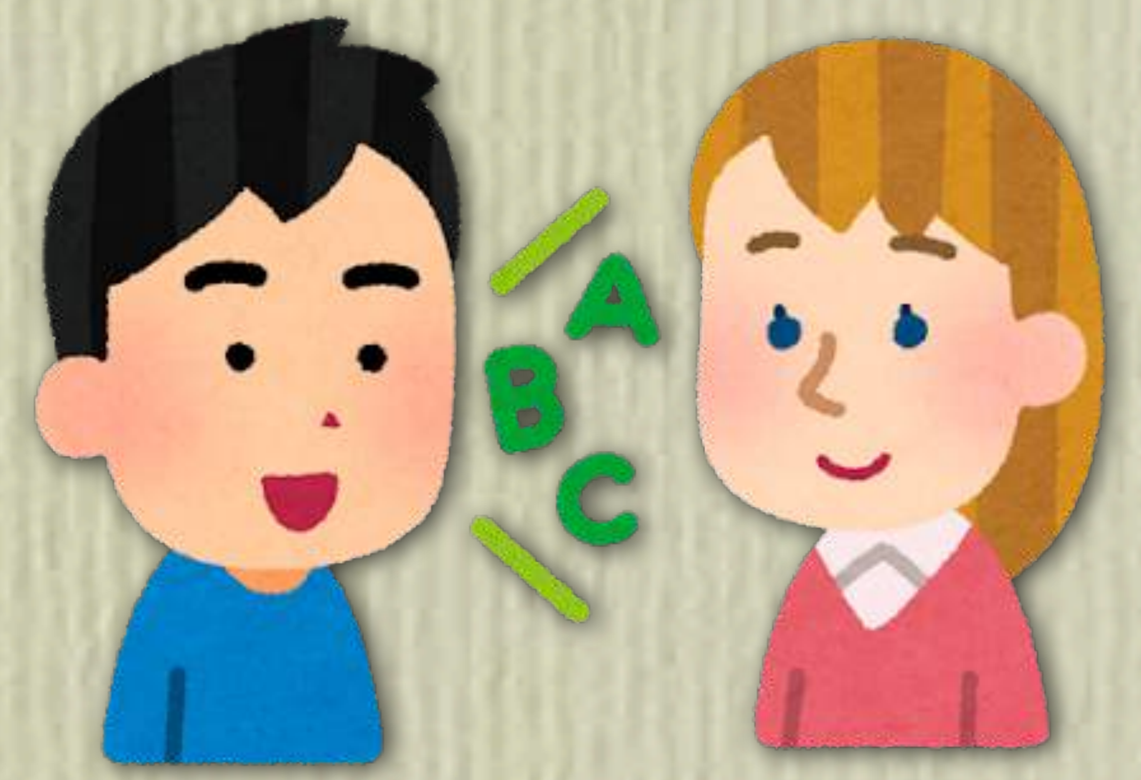
Data collection and system development



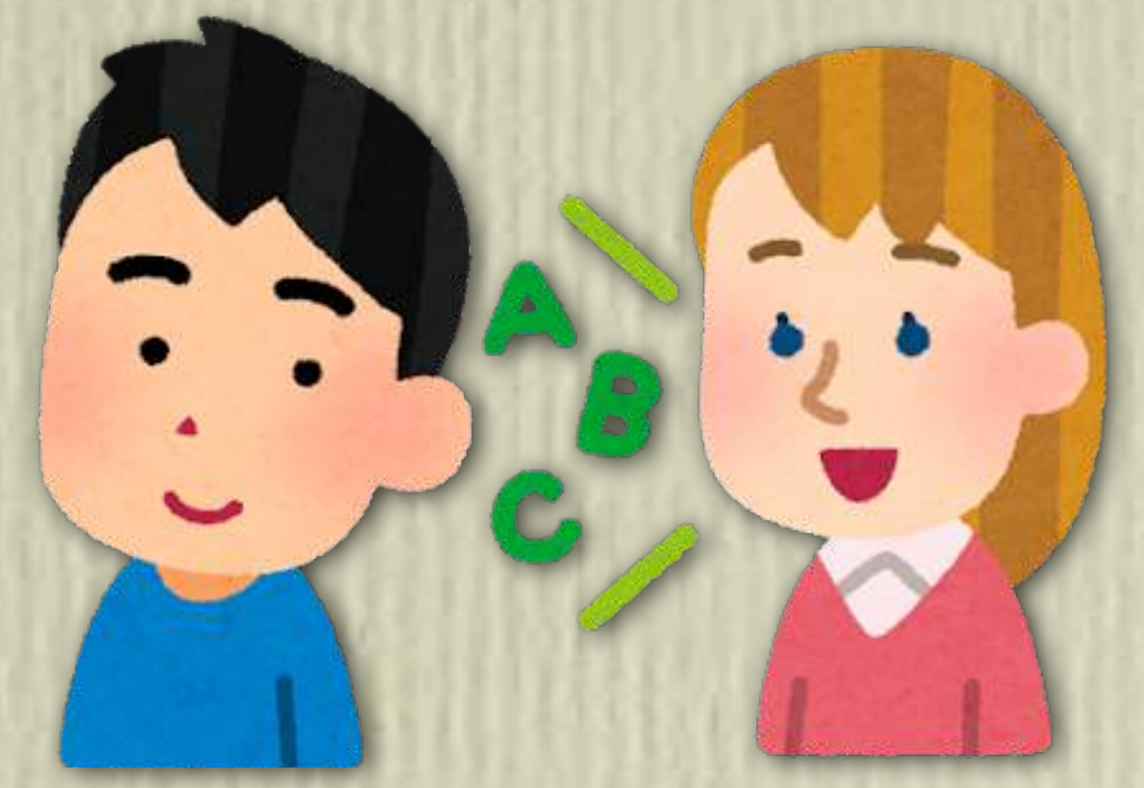
Conclusions

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

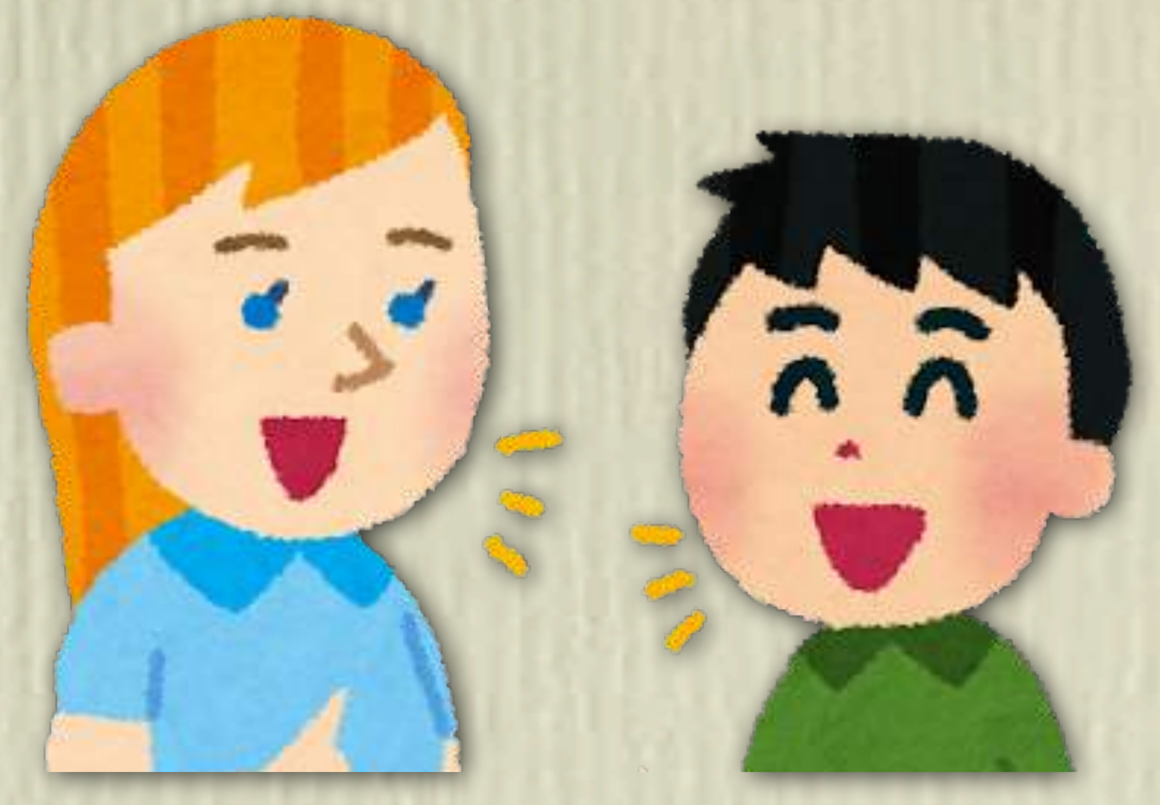
- Computer-Aided Language Learning with speech technologies



with speech synthesis technologies



with speech analysis technologies

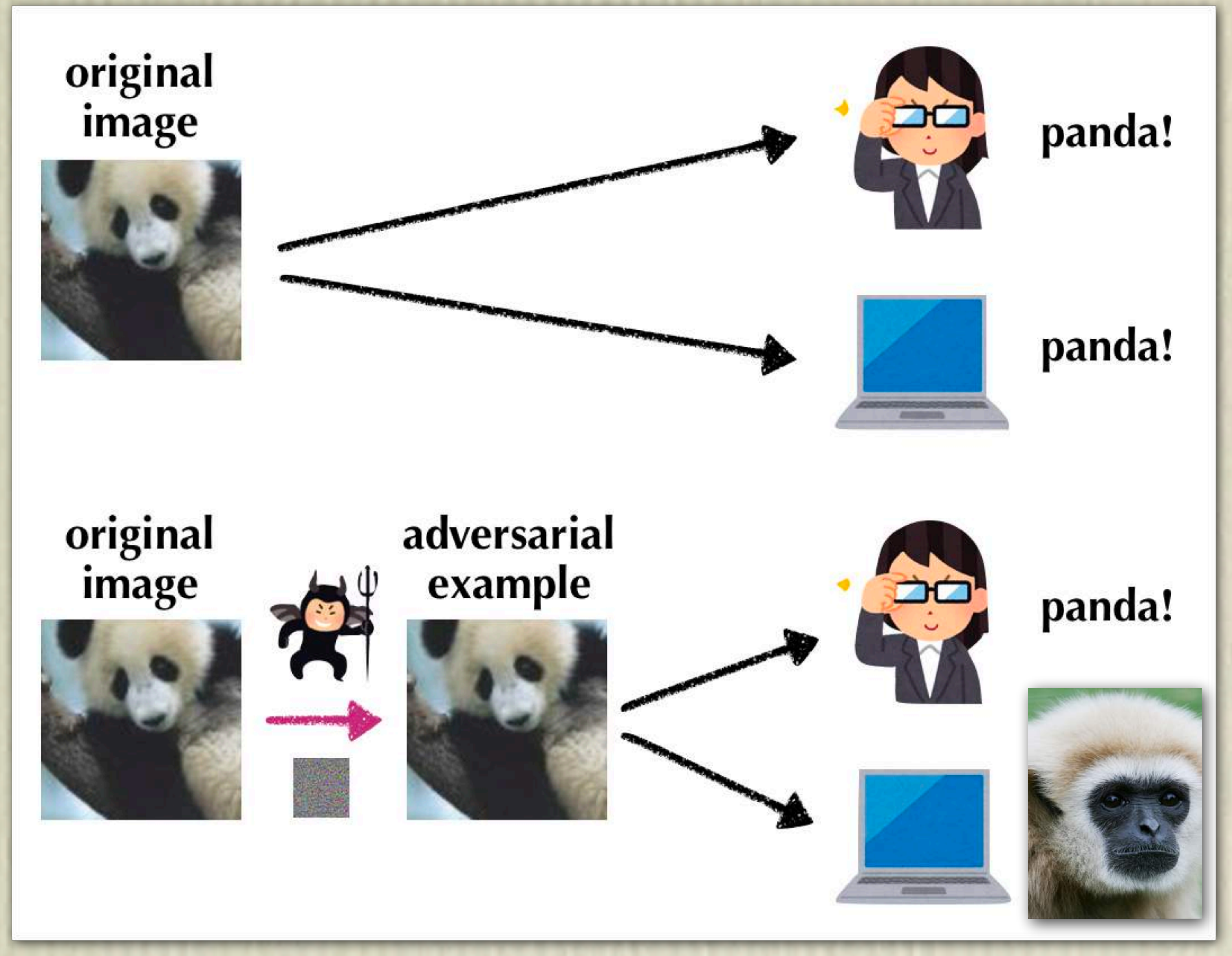


with speech recognition technologies

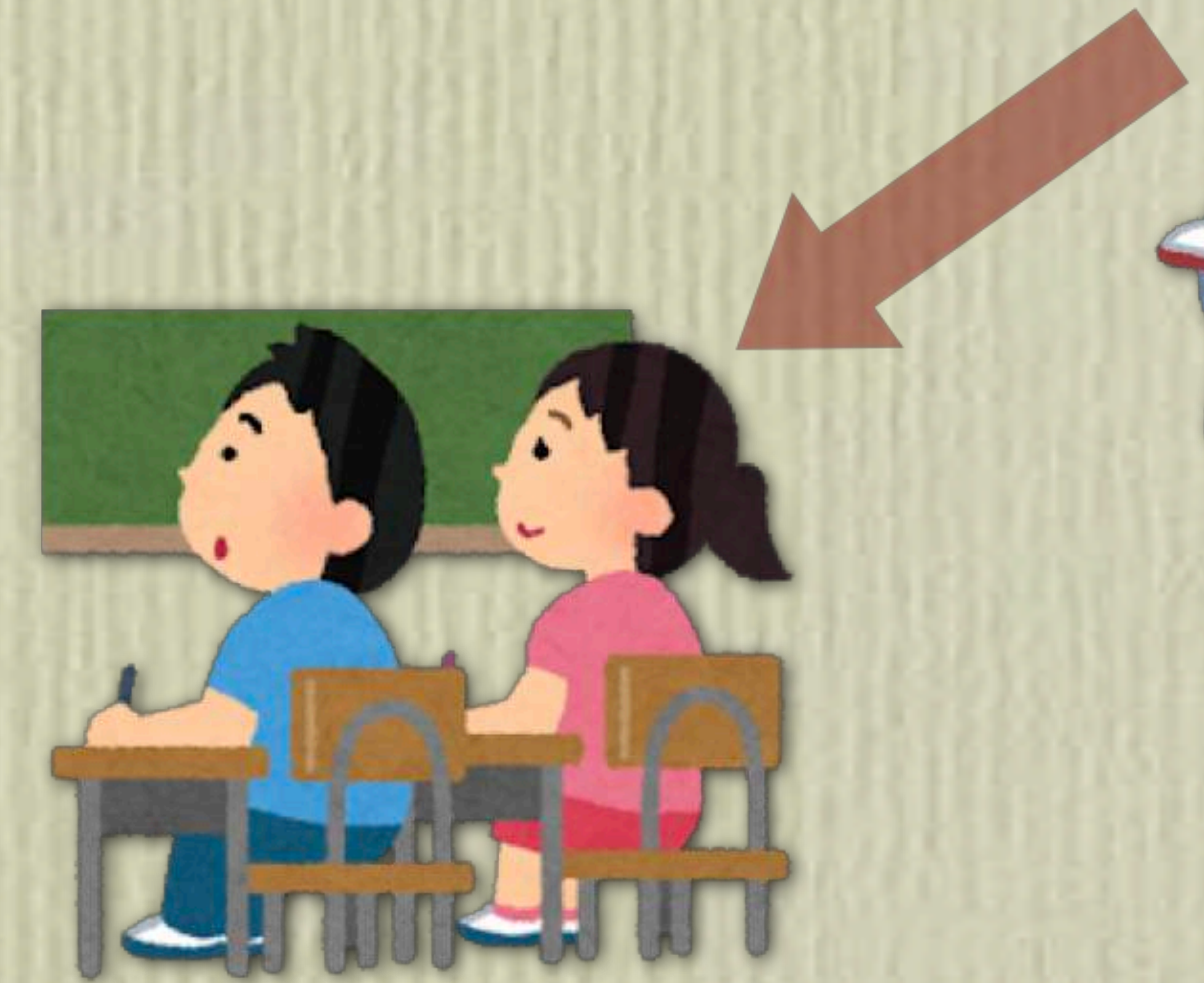
with new speech technologies being developed in our new project

Conclusions

Listening drills with



Conclusions



$$P_{I}(w \mid o)$$

Thank you, any questions?