Bridging different languages, countries, and cultures by Speech-to-speech Translation Research

Satoshi Nakamura

Director, Augmented Human Communication Laboratory,

Graduate School of Information Science,

Nara Institute of Science and Technology, Japan

(Formerly ATR and NICT)

History of Speech Translation Research in Japan

1986 — 1992 NAIST 1999 — 2006 — 2008 — 2010

ATR — ATR — ATR — NICT

Read Speech

- Syntactically correct
- Clear utterance
- Limited domain
  Ex. “Conference Registration”

Daily Conversation

- Standard expression
- Unclear utterance
- Limited domain
  Ex. “Hotel Reservation”

Wider and Real Domain

- Wider and real domain
  “International Travel”
- Realistic expressions
- Noisy speech
- J-E, J-C speech translation

+ More languages for translation

•21 multilateral text translation

VoiceTra

A-STAR

- Multilateral translation for 8 Asian languages
- Network-based S2ST

2011

C-STAR

Multilateral translation for 7 world languages

IWSLT

- Evaluation Campaign of S2S technologies

chnology

Hand-made

Corpus-based Technology

Large scale corpus
+ Machine learning

Speech Recognition

- 1986 at ATR
  - HMM + n-gram : Collaboration with CMU (K. Lee, Sphinx)
  - Spectrogram Reading : Collaboration with MIT (Victor Zue)
  - Neural Network : Dr. Alex Waibel from CMU

- TDNN:
  - Time Delay Neural Network
    "Phoneme Recognition Using Time-Delay Neural Networks", A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, IEEE Transactions on Acoustics, Speech and Signal Processing, March, 1989

    → The IEEE 1990 Senior Best Paper award of the IEEE Acoustics, Speech and Signal Processing Society

  - MS-TDNN: Multi-scale Time Delay Neural Network
    "Modularity and Scaling in Large Phonemic Neural Networks", A. Waibel, H. Sawai, K. Shikano. IEEE Transactions on Acoustics, Speech and Signal Processing, December, December 1989

TDNN: Time Delay Neural Network

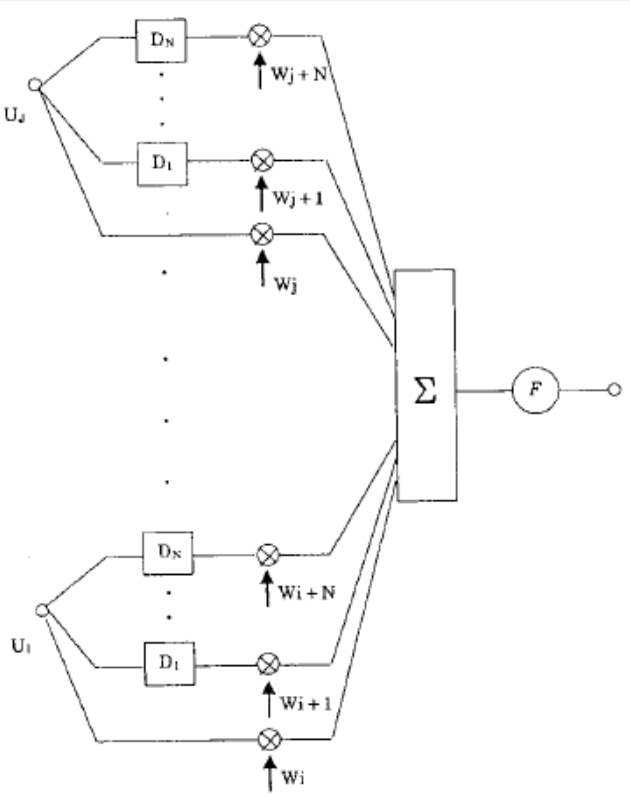


Fig. 1. A Time-Delay Neural Network (TDNN) unit.

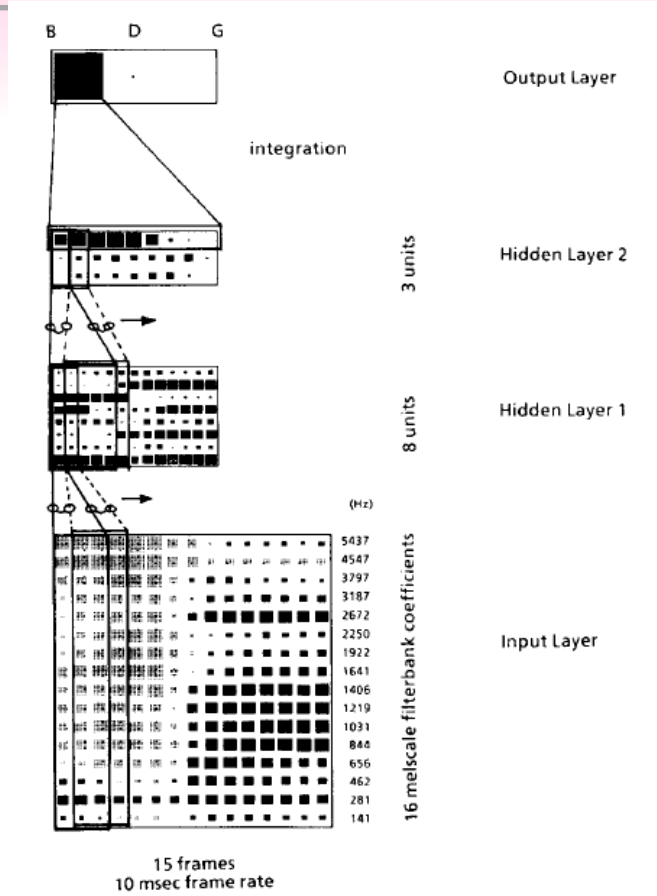
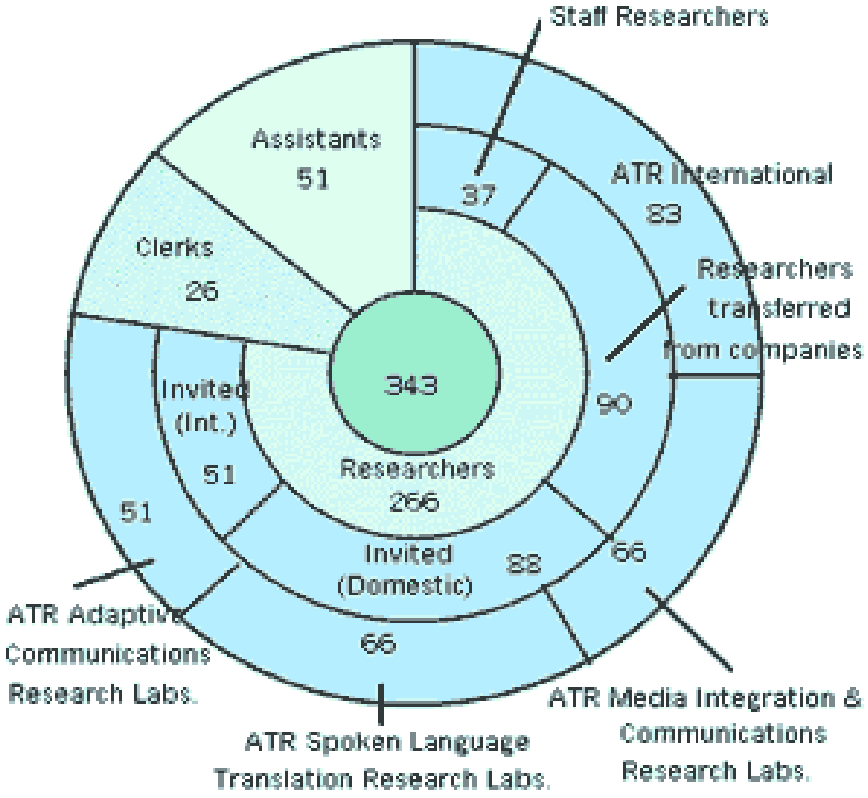


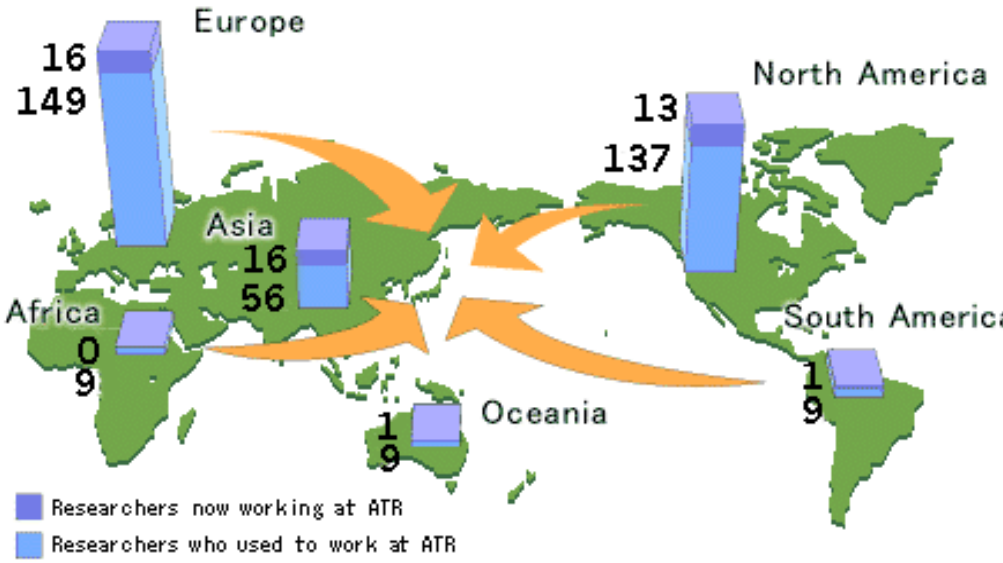
Fig. 2. The architecture of the TDNN.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, “Phoneme Recognition Using Time-Delay Neural Networks”, IEEE Transactions on Acoustics, Speech and Signal Processing, March, 1989

Y. LeCun and Y. Bengio, “Convolutional Networks for Images,Speech, and Time-series,” in The Handbook of Brain Theory and Neural Networks. MIT Press, 1995

ATR as of 2002





14 M Euro/ Year for Each Lab.

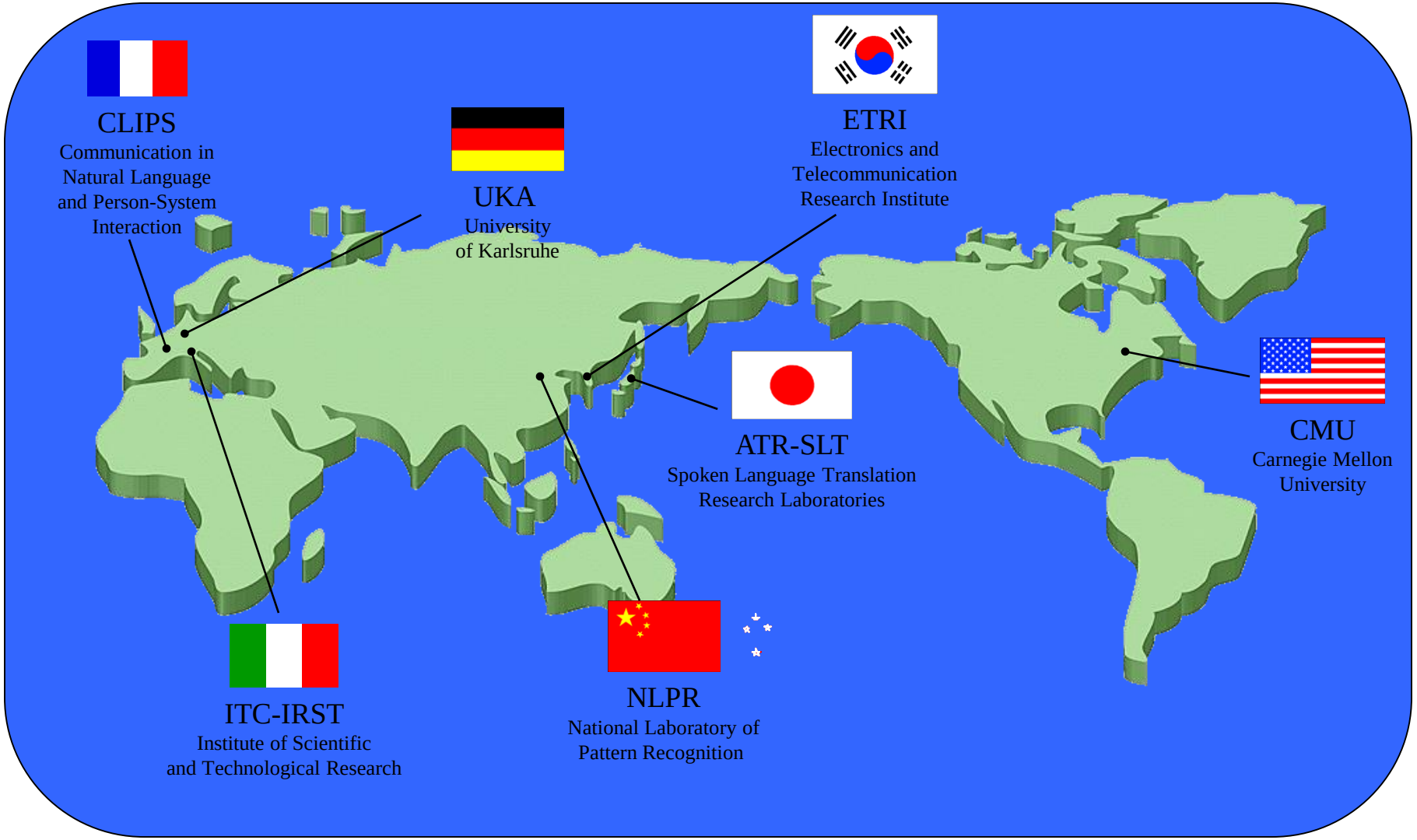
ATR bridged many international researchers !

[Researchers from around the world at ATR]
18% of ATR researchers are non-Japanese.

C-star partners (2002)

(Consortium for Speech Translation)

C-STAR bridged many international research institutes !



Speech-to-speech translation system

Speech-to-speech translation bridged speech and NLP researchers!

Japanese

「私は学校に行く: Watashi wa Gakko he iku」

Multilingual Speech Recognition → Spoken Language Translation → Multilingual Speech Synthesis

English

I go to school

Statistical Speech Recognition

Watashi wa Gakko he iku

Statistical Machine Translation + Multi-engine approach

I go to school

Corpus-based Speech Synthesis

Corpora

- Large Scale Japanese Speech Corpora
- Large Scale Japanese Text Corpora
- Large Scale Parallel Corpora of Japanese and English
- Large Scale English Text Corpora
- Large Scale English Speech Corpora

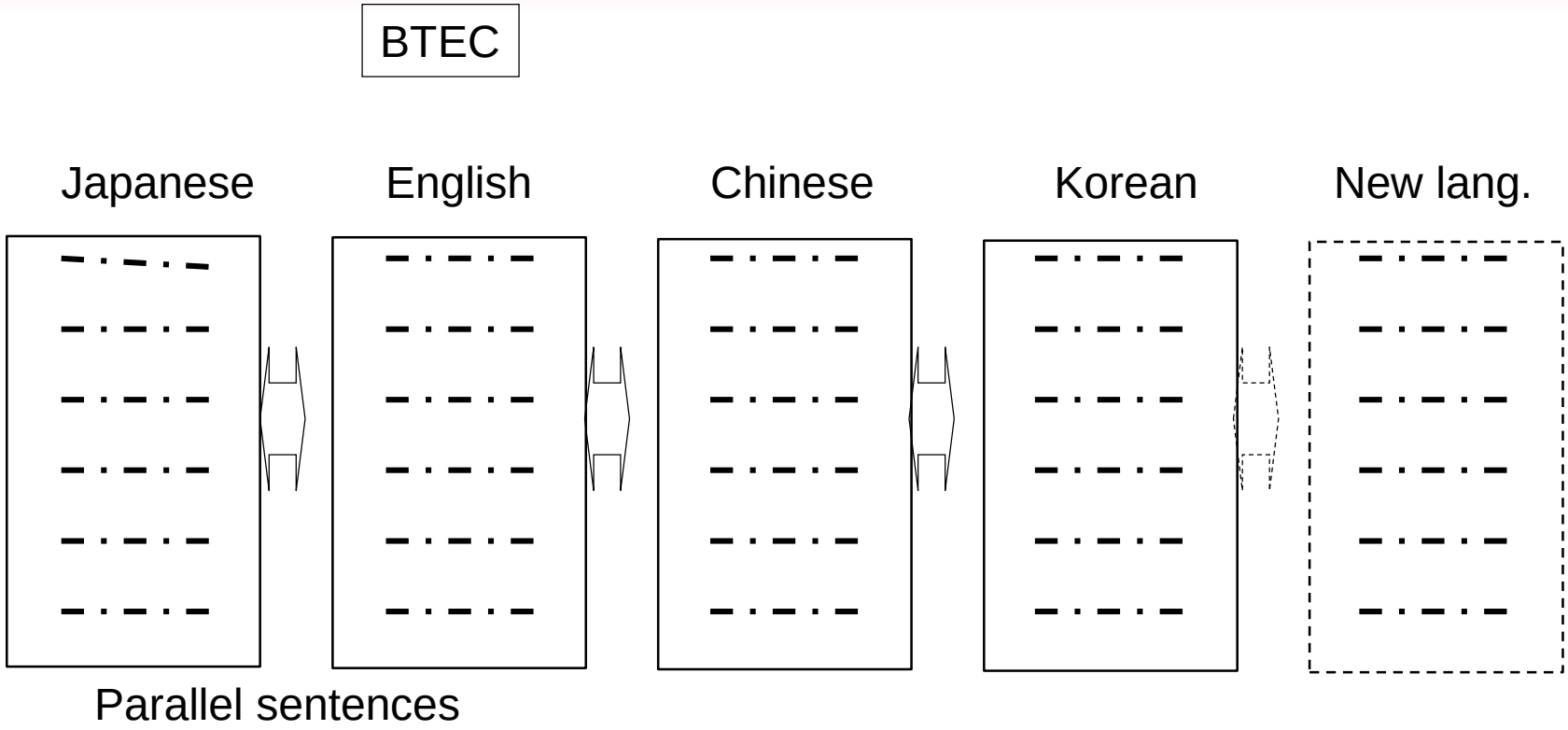
ATR BTEC Corpus

Basic	Trouble	Shopping	Move	Stay
12.2% (7)	12.1% (20)	10.0% (13)	8.4% (8)	8.2% (11)
• greet someone • ask a question • state one's purpose • …	• luggage • emergency • medicine • assistance • …	• buy something • gather information • price • wrapping • …	• transportation • buy a ticket • rental car • trouble • …	• make/change a reservation • check-in • trouble • …

Sightseeing	7.7% (11)	**Study Overseas**	1.6% (14)
Restaurant	7.3% (11)	**Drink**	1.3% (4)
Communication	6.4% (6)	**Exchange**	1.2% (5)
Airport	5.5% (14)	**Snack**	1.2% (4)
Business	5.3% (26)	**Beauty**	0.8% (5)
Contact	4.0% (6)	**Go Home**	0.6% (4)
Airplane	3.6% (11)	**Research**	0.1% (12)
Homestay	2.3% (11)		

Basic Travel Expression Corpus: Parallel Sentences

Spoken Language Communication Research Laboratories



Parallel sentences bridged many languages!

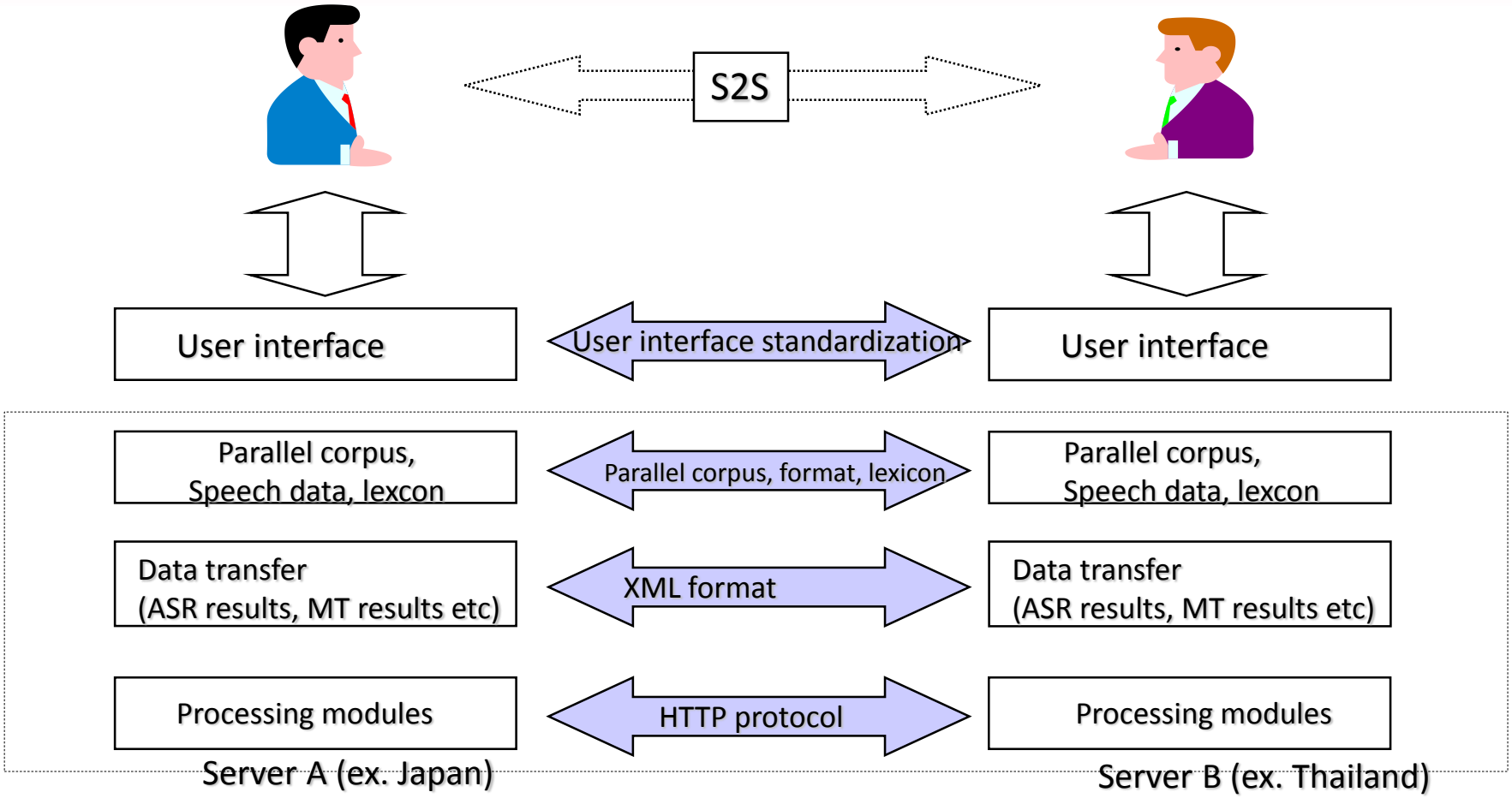
C-star Data Preparation for IWSLT 2004

Status of Shared Corpus
(uploaded to CSTAR server)

		training corpus		evaluation corpus	
		sentences	*paraphrases*	*sentences*	*multiple references*
Japanese	(ATR)	**162k**	—	**506 sen**	**≤ 16 [5x3 sen]**
English	(ATR)		—		**≤ 16 [5x3 sen]**
Korean	(ETRI)		**309k**		**≤ 2 [1x2 sen]**
Chinese	(NLPR)		—		**≤ 3 [1x3 sen]**
Italian	(IRST)	**48k**	**7k**		**≤ 6 [1x3 sen]**
Spanish	(CMU)	**6k + α?**	**2k**		**≤ 10 [2x5 sen]**
German	(UKA)	**β?**	—	—	—
French	(CLIPS)	**γ?**	—	—	—

IWSLT bridged many international researchers!

Standardization at S2ST



Standardization activity of Network-based S2ST protocol at ITU-SG16

- Activity start for standardization of Network-based S2ST at ITU-T SG16
- Session period：2009- 2010
- NICT is the editor for S2ST standardization at ITU-T SG16, WP2 Q21/22

Document	Title	Scope
F.745	Functional Requirements for Network-based S2ST	- Definition of Network-based S2ST - Functions and service requirements of network-based S2ST
H.625	Architectural Requirements for Network-based S2ST	- Requirements of S2ST architecture - Definition of interface for Network-based S2ST

- Not only language conversion but also the potentially added module like sign language are taken into account：

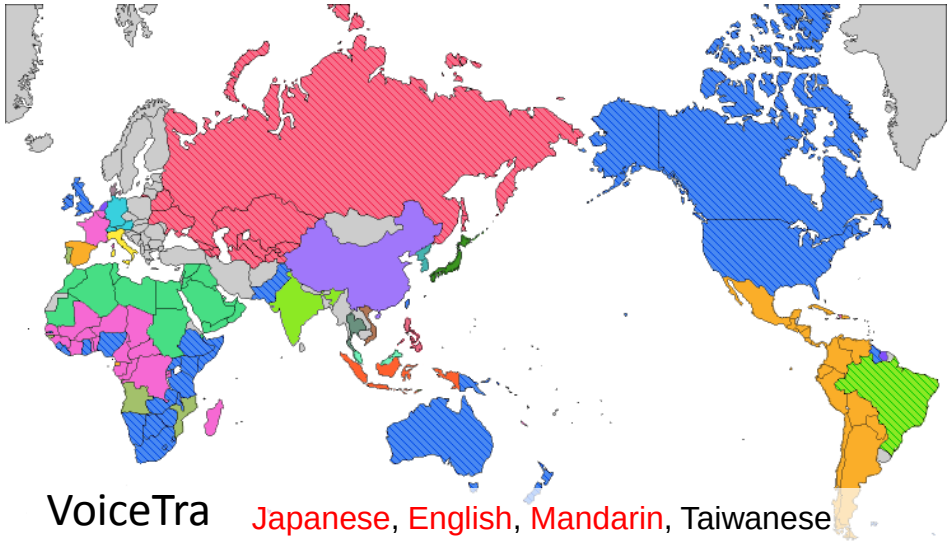
  S2ST -> Modality conversion

ITU-T standards bridge more languages and services!

VoiceTra, TexTra on iPhone (2010)

- **A new speech translation software “VoiceTra” is available for iPhone at AppStore (released on July 29, 2010).**
- **21 languages (including Ja, En, Ch, Ko) are covered, while 6 languages (including Ja, En, Ch) are voice enabled.**
- **So far approx. 800,000 download and 10 million accesses have been achieved after the software release. (2012)**

* Text-translation application, TexTra is released at the same time.



VoiceTra

Japanese, English, Mandarin, Taiwanese Mandarin, German, French, Dutch, Danish, Italian, Spanish, Portuguese, Brazilian Portuguese, Russian, Arabic, Hindi, Indonesian, Malay, Thai, Tagalog, Vietnamese, Korean

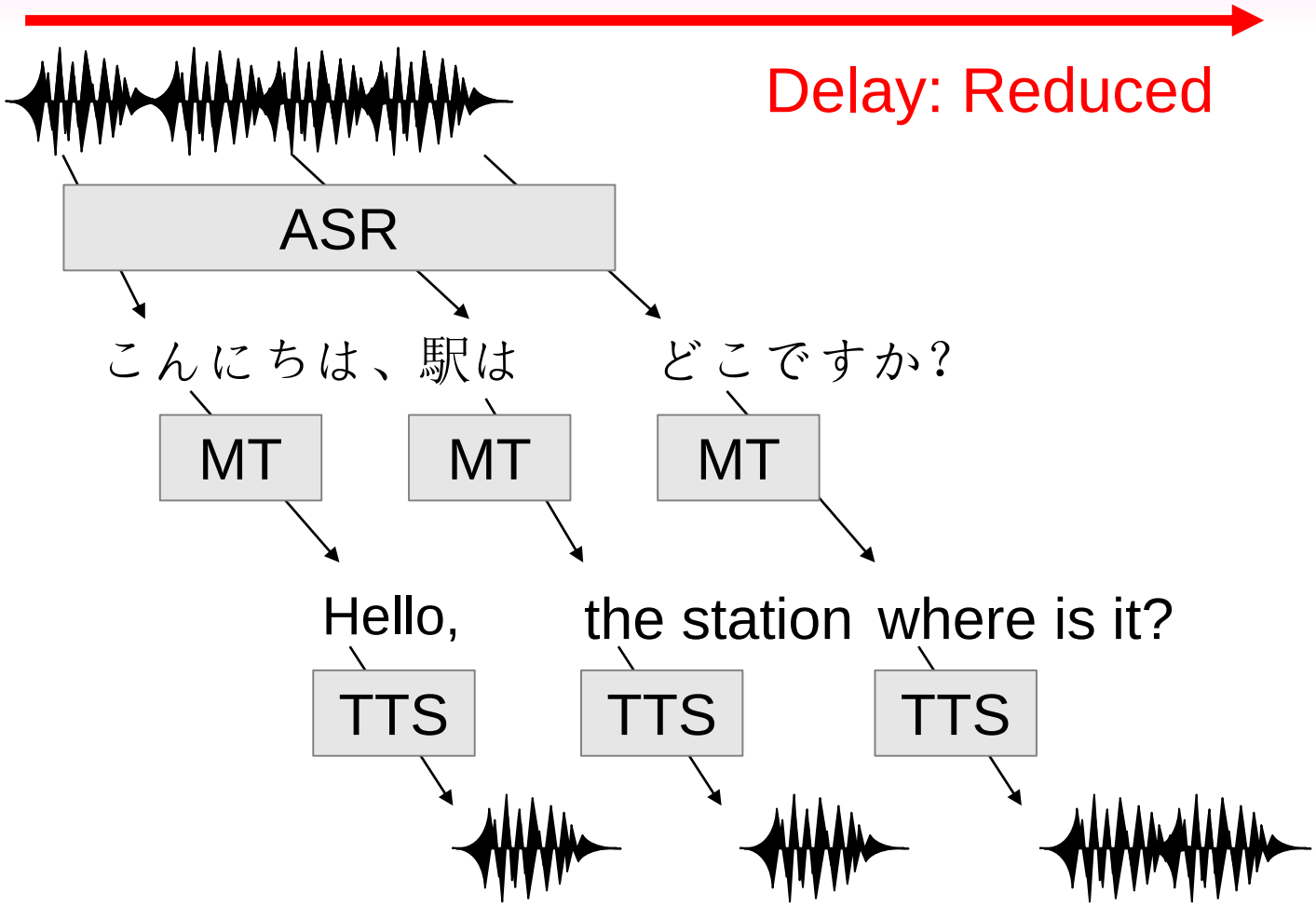
※ Language in red can be input/output in voices.

※There is no text input support for Hindi or Vietnamese.

Global Communication Project in Japan

- 2020 Olympic and Paralympic Game in Tokyo
- Incresing incoming tourists
  - 20 million international tourists to Japan in 2015
- Language suppoet service in Japan
  - Initiated Ministry of Information and Communication, Internal Affairs of Japan, MIC
  - National Institute of Information and communications technology, NICT
  - Global Communication Project Consortium, GCP
    - Research and Development Group (Nakamura, NAIST)
    - Technology Transfer Group (Usami, KDDI)
  - Industry Consortium
    - Panasonic organizes contract project funded by MIC.
      - NTT, Fujitsu, NEC, …..
- Target: Speech-to-speech translation service for
  - Shopping, sightseeing, living, troubles, disaster management
  - Stores, hotels, sightseeing spots, hospital

Simultaneous Incremental Speech Translation



But, this is not easy!

Can We Do the Same in Speech Translation Systems?

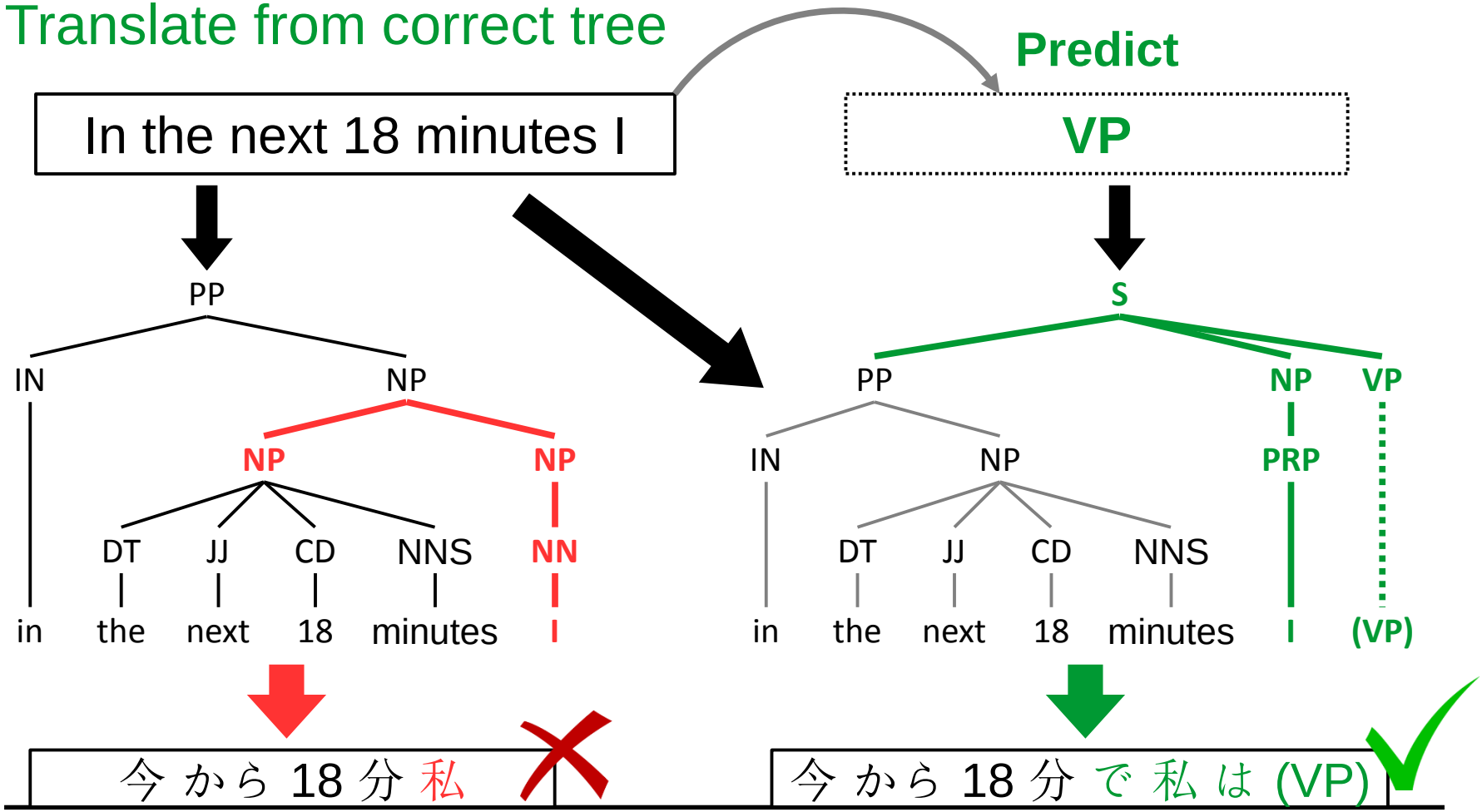
Four problems:

- Segmentation: When do we start translating?
- Prediction: Can we predict things that haven't been said?
- Rewording: Can we reword sentences to be conducive to simultaneous translation?
- Evaluation: How do we decide which results are better?

Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents [Oda+ ACL15]

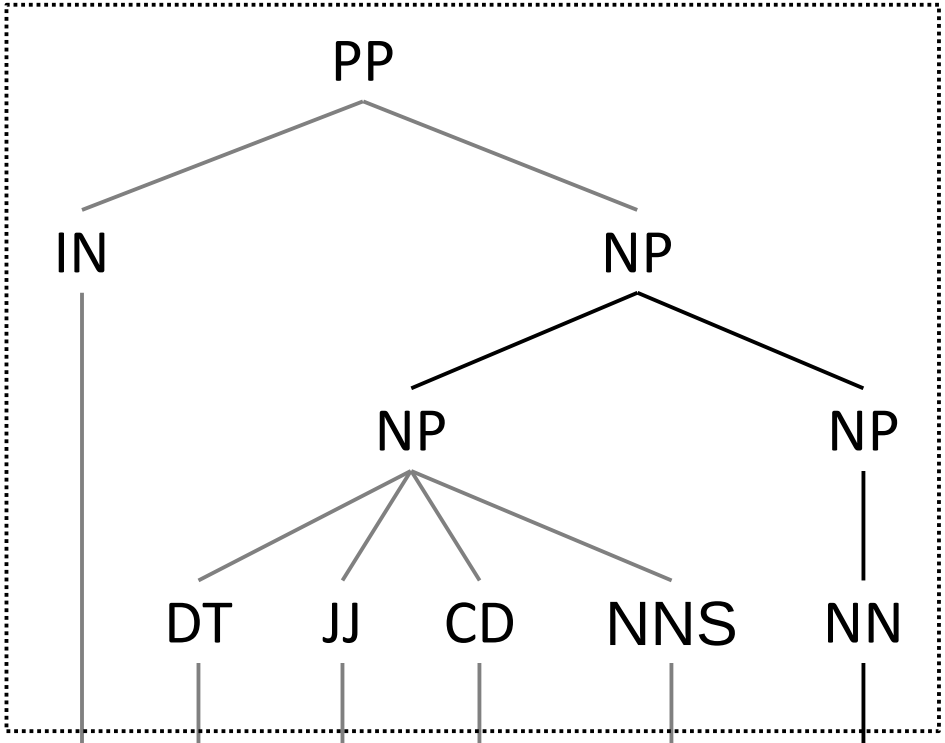
- Predict unseen syntax constituents

- Translate from correct tree



Syntax Prediction Process

1. ***Parse*** the input as-is



2. ***Extract*** features

Word:R1=I	ROOT=PP
POS:R1=NN	ROOT-L=IN
Word:R1-2=I,minutes	ROOT-R=NP
POS:R1-2=NN,NNS	...
...	

3. ***Predict*** the next tag (linear SVM)

VP ... 0.65
NP ... 0.28
nil ... 0.04
...

4. ***Append*** to sequence

5. ***Repeat*** until nil

in	the	next	18	minutes	I	VP			nil

Input translation unit

Summary and Future Directions

- Speech-to-speech translation research
  - bridged
    - Different languages, international researchers, researchers in different fields, different services, international research institutes
  - InterACT and C-Star contributed to the S2ST history.

- New Research Direction of Speech Translation at NAIST
  - Simultaneous incremental speech-to-speech translation
  - Emotion, para-linguistics, face, and gesture translation
  - Neural MT modeling

- Future Work
  - Para-linguistics and discourse structure
  - Context and situation
  - Background and specific domain knowledge
  - Semantic structure and semantic analysis
  - Communicability through the speech translation

NAIST joined interACT (2012)