



IBM Research

# Real-time Speech Transcription Service to Improve Non-native Speaker's Listening Comprehension



IBM China Research Lab  
IBM TJ Watson Research Center  
IBM Human Ability & Accessibility Center

## Overview

---

- Globalization is driving many people to communicate in non-native languages, but the comprehension of a non-native spoken language often poses many difficulties:
  - Speech-to-Speech translation is a solution, but often not feasible at the moment due to the technological difficulty of machine translation.
  - We propose to provide **real-time transcription** synchronized with the audio stream to improve non-native speaker's comprehension.
  - The state-of-the-art real-time transcription system can achieve about 10% or even less WER if an in-domain language model is used.
- Does real-time transcription REALLY help?
  - We first designed and implemented an user study experiment to validate the idea.
- Real-time speech transcription service
  - We have developed a prototype system to provide real-time transcription in various scenarios.

## User study: Experiment design

---

- 1-way scenario of remote computer-mediated communication
  - Native speakers, Chinese listeners
- Mixed design (2\*3)
  - Between-subject variable: Communication Modality (2 conditions)

*Audio Only (A) vs. Audio + Video (A+V)*

- Within-subject variable: Transcription (3 conditions)

*No Transcription (NT) vs.*

*Transcription synchronous with speech (T) vs.*

*All Transcripts remained (RT)*

- Measures
  - Comprehension performance
  - User experience evaluation
  - Cognition resource allocation

# User study: Interface

## Audio Only

## Audio + Video

RT: All transcripts remained on the screen

T: Transcription synchronous with the speech

Please listen to the following Audio with subtitles.

When we talk about a modern company



when the melting of Greenland ice sheet  
that it's no longer reversible  
We can't stop it  
All right we're all in the same boat  
What about developing and developed countries  
There has been a lot of debate between the two camps  
as to whose responsibility it is to cut first

RT: All transcripts remained on the screen

Please listen to the following Audio with subtitles.

When we talk about a modern company  
we usually have managers employees products  
research and development or marketing in mind  
However in reality a company is not just made up of these elements  
There are other things that make a company what it is  
Today we're going to look at some other aspects of a company  
Let's first take a look at of

T: Transcription synchronous with the speech



In others the staff members work more privately

NT: No transcription



## User study: Participants

- 48 Chinese university students participated as paid volunteers.
- All were Chinese native speakers and NON-English-majors, majoring in a variety of disciplines including CS, Biology, Finance, EE, Accounting, Law and etc.
- English Level:
  - All passed CET-6 (College English Test- band6, national standard test), mandatory in most major universities to get a master's degree and much preferred by local and multinational companies for recruiting now.
  - The avg. years of learning English is 11, mostly starting from Primary school.
  - Though CET-6 is a reasonable basis for potential communication in English, English listening still poses difficulty\*.

### Demographics

Between-subject	N	Gender		Avg. Age	Education	
		M (25)	F (23)		Undergraduate	Graduate
Audio only	24	13	11	22.5	14	10
Audio + Video	24	12	12	24.8	11	13

\* Jiang, Q.X., Survey on Needs and Difficulties of Student English Learning, Sino-US English Teaching, 2005, Volume 2, No.10 (Serial No.22)

## User study: Task

- Each participant was asked to listen to (A) or watch (AV) 6 English clips, and then:
  - Answer comprehension questions and report confidence of the answers accordingly after finishing each clip.
  - Fill in evaluation questionnaires after all clips and comprehension questions are finished.
  - Examples of task
    - Dialogue, T: <http://9.186.102.47:8080/EEC-Demo1>
    - Lecture, RT: <http://9.186.102.47:8080/EEC-Demo2>
- The same Latin-Square design was implemented for each modality (A,AV).

		Within		
		L1&D3	L2 & D1	L3 &D2
Audio Only [or AV] (n=24)	Group 1	NT	T	RT
	Group 2	NT	RT	T
	Group 3	T	NT	RT
	Group 4	T	RT	NT
	Group 5	RT	NT	T
	Group 6	RT	T	NT

NT: No Transcription

T: Transcription synchronous with the speech

RT: All transcripts remained on the screen

## User study: Materials

---

- Test Materials
  - 6 English clips of 3.5 minutes' long in average and of general topics.
    - 3 dialogue clips cut from an English TV show (2 or 3 persons for each): D1-D3.
    - 3 lecture clips recorded by invited native speakers (1 person for each): L1-L3.
  - 5 comprehension questions for each clip.
    - Balanced in question form (choice/essay) and question type (local/global).
- The materials were validated/adjusted through pilot experiments.
  - Two rounds: 6\*2 native Chinese university students (all passed CET-6)
  - Deleted clips with inadequate accent/speaking rate and topic.
  - Deleted and redesigned questions with too high or too low accuracy.
    - All questions have a difficulty level (accuracy) in the range of 0.3-0.7\*.
  - 4 conditions were tested in pilot (RT + A/AV, NT+A/AV). Pilot results showed that transcription helped accuracy.

\* Jin Y., Psychology Measurement. East China Normal University Press, 2005

## User study: Measurements

---

- Comprehension Scores
  - Comprehension performance: response accuracy
  - Comprehension confidence: “to what extent you are confident in your answer to the question?” (5-point likert scale)
- User Experience Questionnaire (5-point likert scale)
  - Usefulness: “I think transcription is helpful to my understanding.”
  - Willingness to use: “I would love to use transcription next time.”
  - Preference: “I like transcription.”
- Self-reported Cognitive Resource (attention) Allocation
  - “How did you allocate your attention to the following different information sources when you received information from the clips in each of the three conditions?” (in percentage respectively and full score 100%)

# User study: Procedure

---

**1**

## Pre-Training

- The experimenter briefs participants about the procedure and show them system interface before the experiment starts.
- Participants are encouraged to do their best.

**2**

## Comprehension Test

Each participant needs to finish 6 English clips in different conditions. For each clip, the participant needs to:

- Go through the topic and questions within 1.5 minutes with a count-down on the screen. ([Pro-active information seeking](#))
- Listen to or watch the clip. (the clip begins automatically and is played for only once)
- Answer questions. (Questions come out once at a time, and confidence level choices are shown after each question is answered)

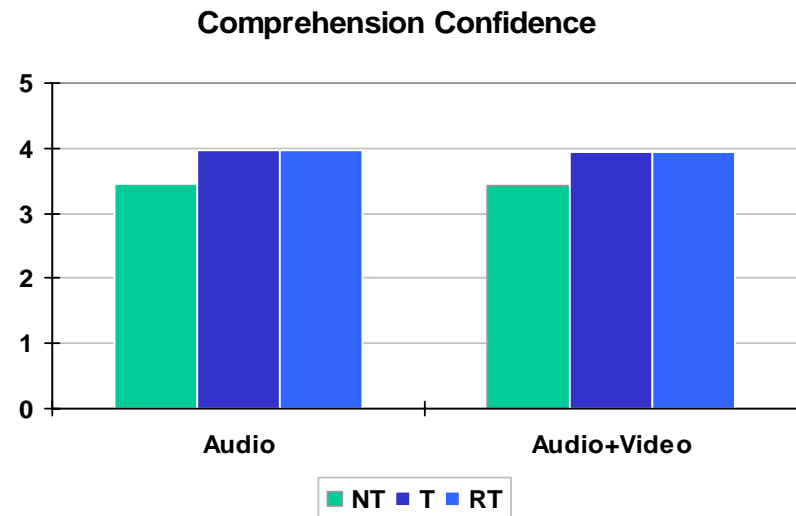
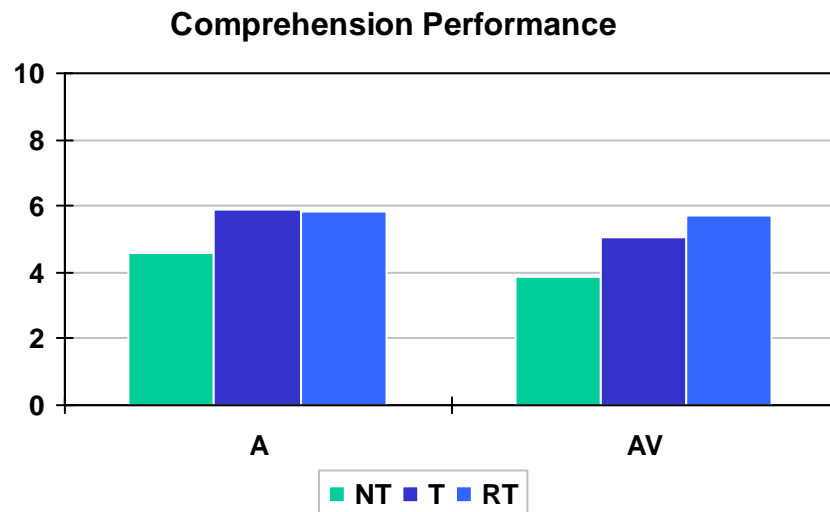
**3**

## Questionnaire Filling

After finishing all clips, participants are asked to fill in two questionnaires (user experience & cognitive styles).

## User study: Result – Comprehension Scores

- ANOVA test results:
  - Main effect of Modality (between-variable) is marginally significant on comprehension performance ( $F(1,47)=3.82, p=.053$ )
  - Main effect of Transcription (within-variable) is significant on both comprehension performance ( $F(1,46)=11.28, p<.01$ ) and confidence ( $F(1,46)=13.69, p<.01$ )
  - No interaction effect.
- Post-hoc t-test shows that real-time transcription (both T&RT) significantly improves both comprehension performance and confidence, in both audio-only and audio + video condition.



## User study: Result – User Experience Scores

---

- Both in A & AV, participants reported a satisfying user experience with Transcription (1=Strongly disagree, 5=Strongly agree, with positive statements of user experience)
- There is no significant difference in User Experience between A & AV.

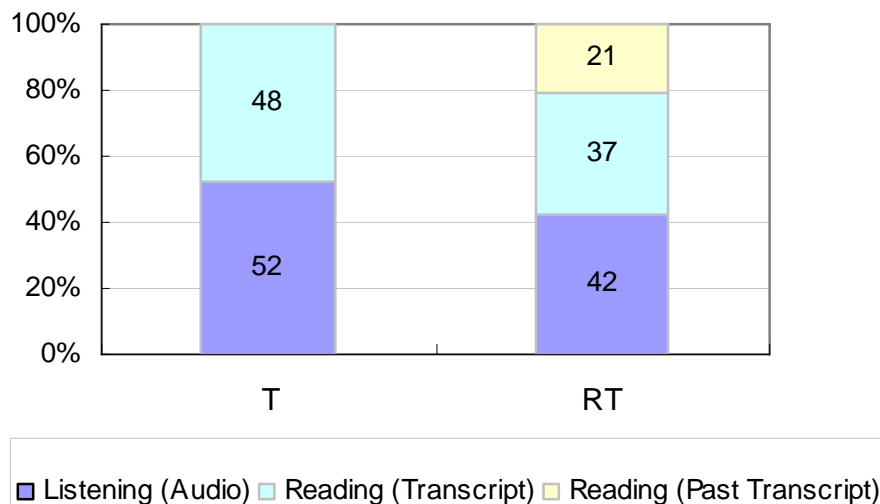
**User Experience with transcription**

	<b>Audio only</b>	<b>Audio + Video</b>
User Experience	<b>4.04</b>	<b>4.32</b>
<b>Usefulness</b>	<b>4.17</b>	<b>4.54</b>
<b>Preference</b>	<b>3.92</b>	<b>4.29</b>
<b>Willingness to use</b>	<b>4.04</b>	<b>4.13</b>

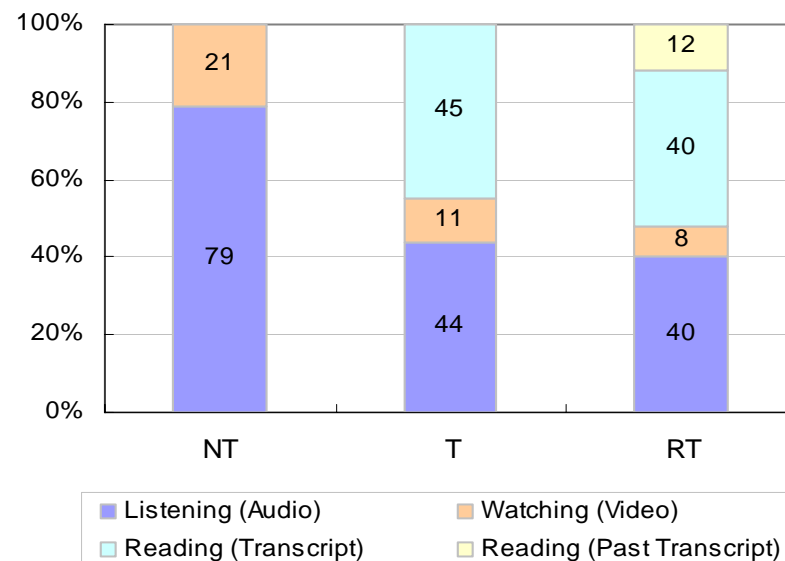
## User study: Result – Cognitive Resource Allocation

- **Resource Allocation Assessment:** “how did you allocate your attention to the following different information sources (in percentage respectively and full score 100%) when you received information from the clips in each of the three conditions?”
- In A condition, participants allocated about a half of resources (attention) on Transcription.
- In AV condition, participants allocated a very small portion of attention on Video but still about half on Transcripts.

Audio Only Attention Allocation



Attention Allocation Audio + Video



## User study: Conclusions

---

- Real-time transcriptions (T & RT) are able to improve **comprehension performance and confidence** significantly both in Audio-only (A) and Audio + Video (AV) conditions.
- **Video doesn't improve performance**; actually, comprehension performance in A is marginally better than that in AV.
  - It is compatible with pervious findings. (" video makes people more satisfied with the work, but it doesn't help the quality of the work. One exception is negotiation task "\*.)
  - But T is still effective in AV condition.
- Both in A & AV, participants reported a **satisfying user experience with Transcription**.
- T takes about one half of **attention** for comprehension in A; in AV, T still accounts for a big portion and video becomes minor.
  - Users tend to focus more on sources that have a direct effect on performance.

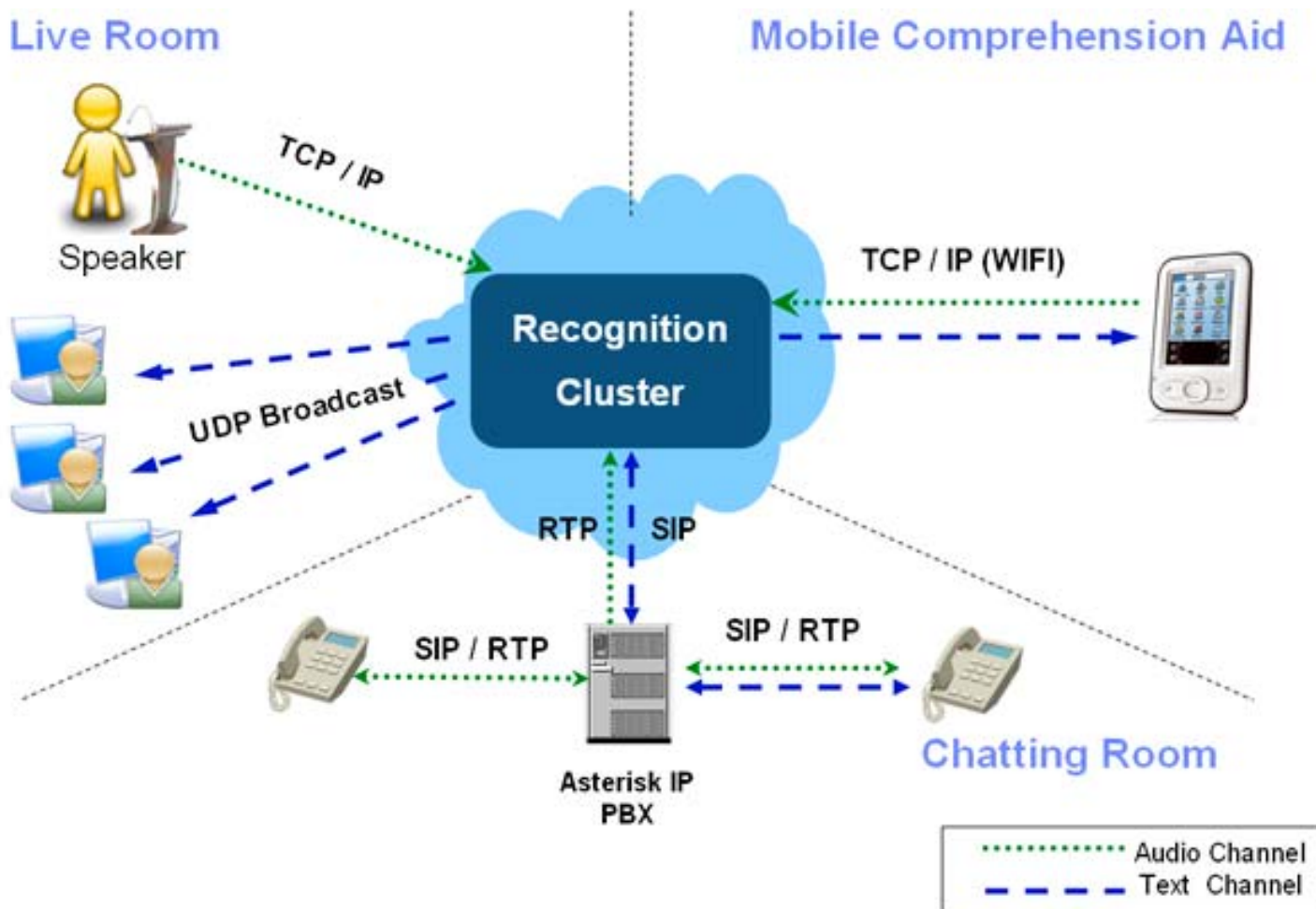
\*E.S.Veinott, X.Fu, J.Olsen, et al. "Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other," CHI 99.

## Real-time transcription service: Overview

---

- Based on the user study result, we implemented a prototype system to provide real-time service to improve non-native speaker's comprehension in a multilingual communication.
- Three applications were developed to meet the requirements of different communication scenarios:
  - *Live Room* for one-way communications: there is a primary speaker and a number of listeners that can be located in different places.
  - *Chatting Room* for long-distance two-way communications: two or more participants engage in a discussion.
  - *Mobile Comprehension Aid* for face-to-face communications: can happen at any time, any place.

# Real-time transcription service: System paradigm



## Real-time transcription demo

---

- The demo would be a client-based dictation machine.
  - BCC-Direct-Speech

# Accessing the Future:

A global collaborative exploration for  
accessibility in the next decade

July 20-21, 2009  
Boston, MA



## Conference tracks:

- Universal Design and Accessibility Standards
- Patient-centered Collaborative Care
- Accessible Online Workplaces and Communities
- Travel & Transportation



---

Session ID: AAC-3034

Real-Time Speech Transcription Service to Improve  
Non-Native Speakers' Listening Comprehension

Presenter:  
Ali Sobhi

Author:  
Dan-ning Jiang

Date: Friday, March 20, 2009

Start Time: 9:20 AM

End Time: 10:20 AM

Location: Renaissance - International C