From Research to Product, Transforming the Impossible to the Expected

Eric Chang, Ph.D.

Senior Director, Technology Strategy

Microsoft Research Asia

Thanks to:
Alex Acero, Behrooz Chitsaz, Li Deng, Qiang Huo, Chin-Hui Lee, Mark Liberman, Yao Qian, Matt Scott, Frank Seide, Frank Soong, Ivan Tashev, Dong Yu, Lijuan Wang , Chris Wendt

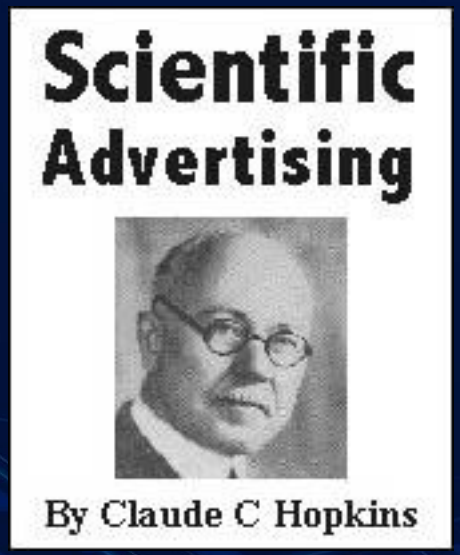
Microsoft Research

Talk Outline

- Introduction
- Factors for Success
  - Reliability
  - Delivered value
  - Frequency of use
- Paths from Research to Product
  - Kinect Based Speech Recognition
  - Speech as a 1st Class Data Type
  - Bridging the Language Barrier
- Opportunities for Research
- Summary

Convincing People to Brush

- Within 10 years > 50% people brushing in US



Why That Tartar

If You Keep Teeth Clean?

All Statements Approved by High Dental Authorities

It is Due to Film

TARTAR shows that teeth are not kept clean. The basis is a slimy film. If you removed it daily tartar would not form.

That film on your teeth causes most tooth troubles. It is ever-present, ever-forming. You can feel it with your tongue.

The film is what discolors, not the teeth. It holds food substance which ferments and forms acid. It holds the acid in contact with the teeth to cause decay.

Millions of germs breed in it. They, with tartar, are the chief cause of pyorrhea.

This film is viscous, so it clings. It gets into crevices and stays. The ordinary dentifrice does not dissolve it. The tooth brush leaves much of it intact. That is why the best-brushed teeth so often discolor and decay.

Every dentist knows this. Dental science has for years sought a way to combat that film. That way has now been found. And, for daily use, it is embodied in a dentifrice called Pepsodent.

We ask you to write for a free 10-day Tube and learn what it means to your teeth.

Watch It Disappear

Get this free tube of Pepsodent and use like any tooth paste. Note how clean the teeth feel after using. Mark the absence of the slimy film. See how the teeth whiten as the fixed film disappears. You will know in a few days what clean teeth mean.

Pepsodent is based on pepsin, the digestant of albumin. The film is albuminous matter. The object of Pepsodent is to dissolve it, then to constantly combat it.

The way seems simple but for long it seemed impossible. Pepsin must be activated, and the usual method is an acid harmful to the teeth.

Then the invention of a harmless activating method made this application possible. And it seems to solve the problem of this tooth-destroying film.

Pepsodent has been proved under able authorities by many clinical tests. Leading dentists all over America have come to endorse and adopt it. Now we urge you to try it.

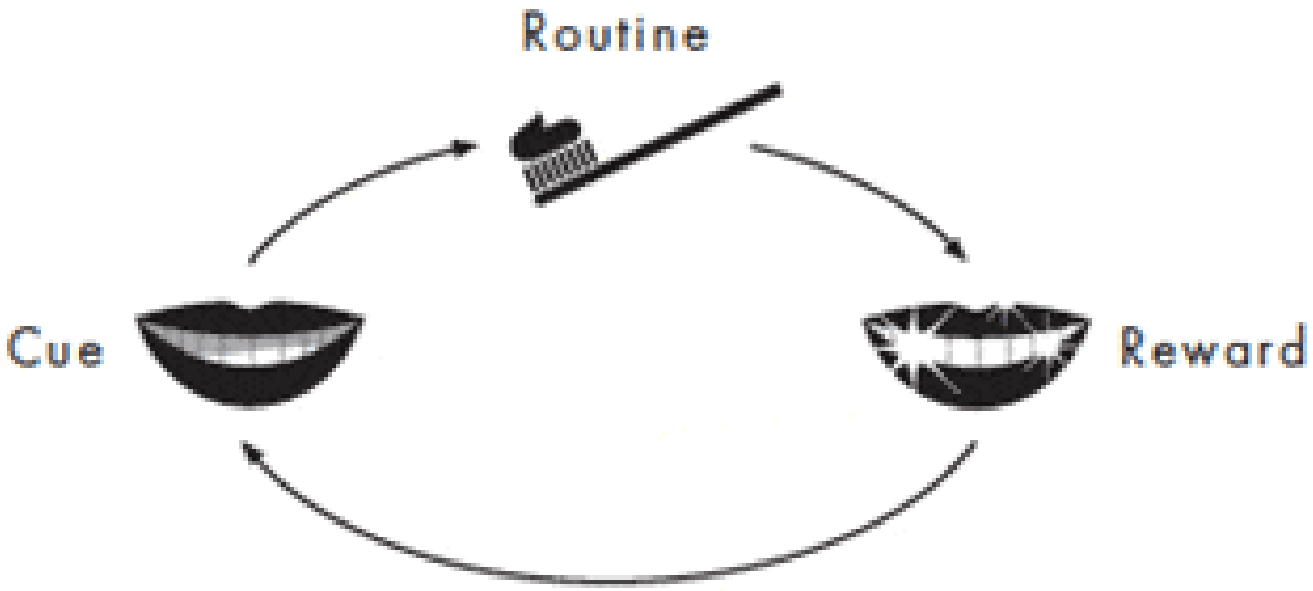
Pepsodent PAT. OFF. REG. U.S.

The New-Day Dentifrice

Now advised by leading dentists. Druggists everywhere are supplied with large tubes

Microsoft Research

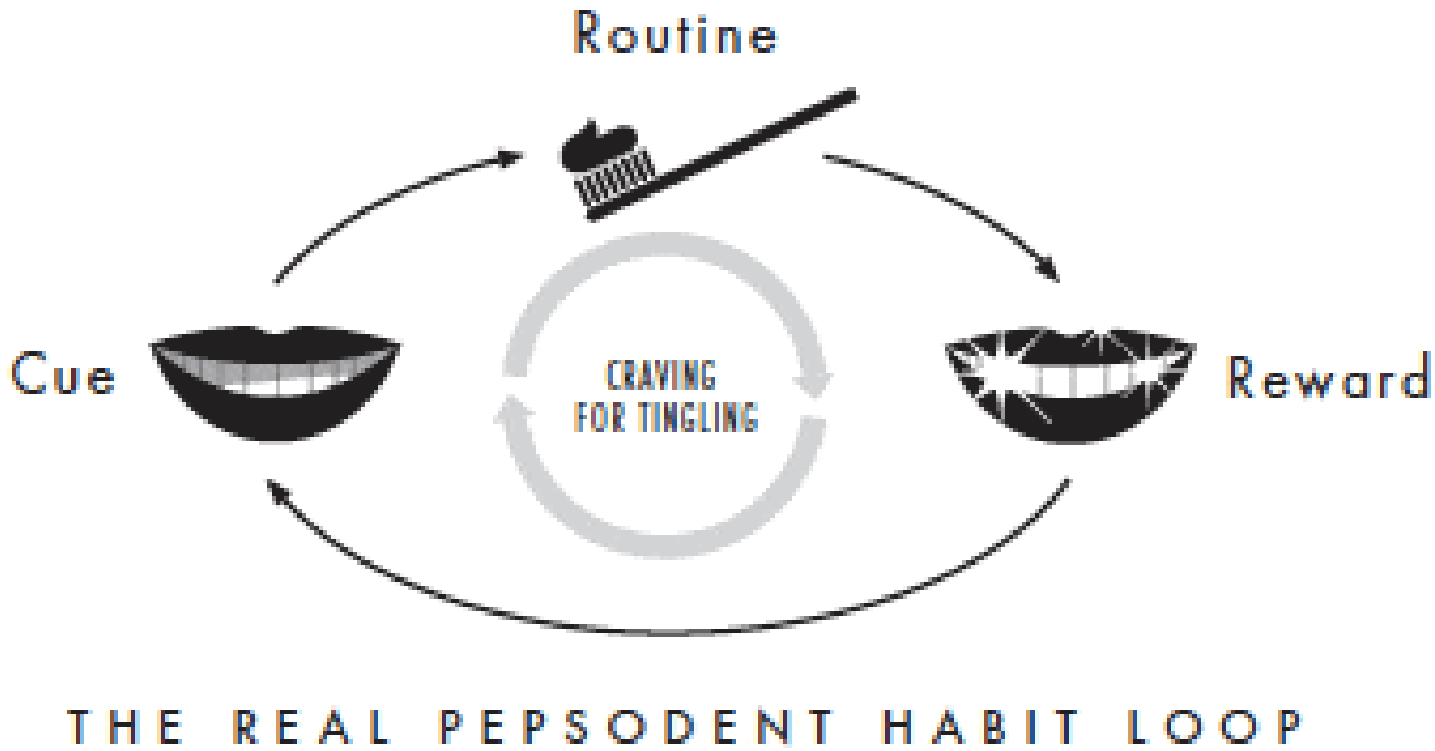
Convincing People to Brush



HOPKINS' CONCEPTION
OF THE PEPSODENT HABIT LOOP

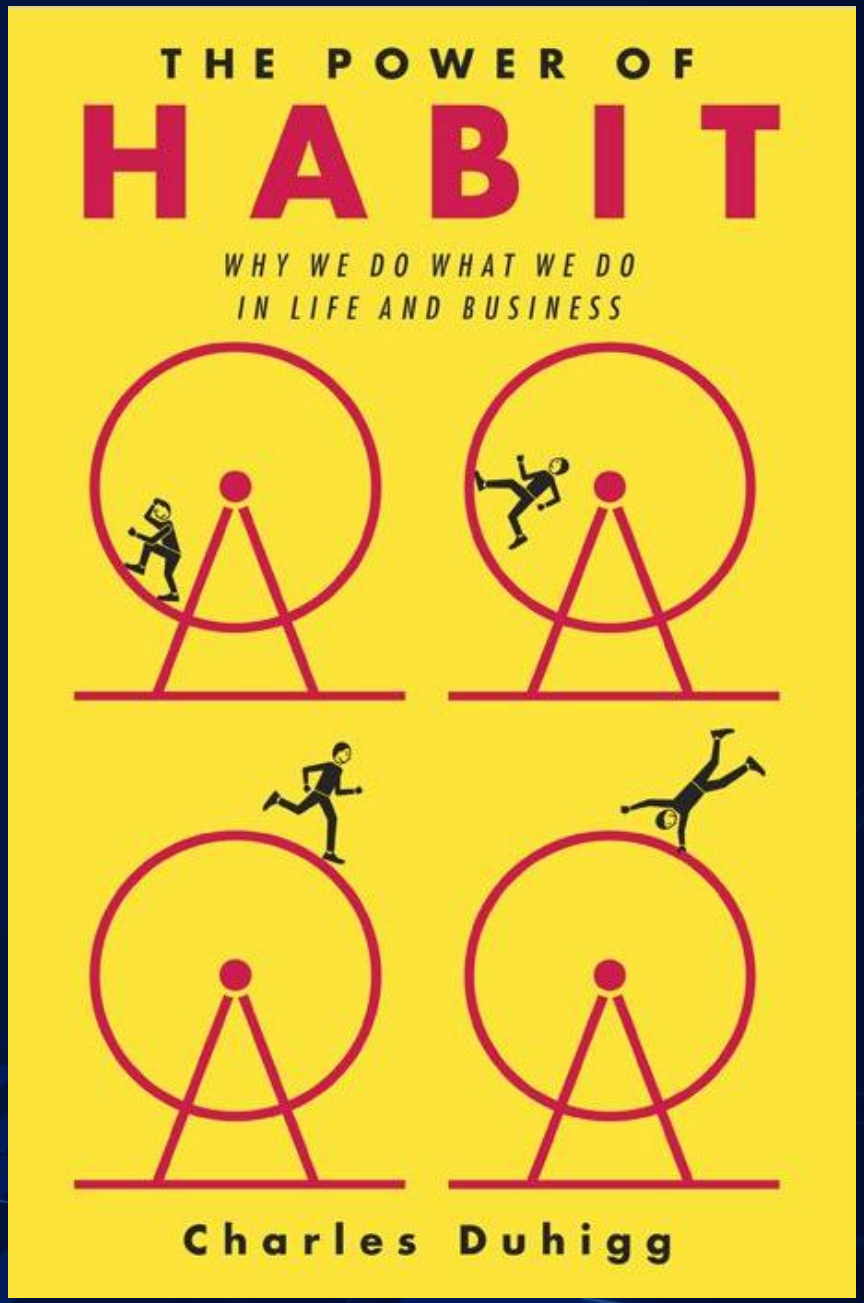
Charles Duhigg, *The Power of Habit*

Feedback with Quick Affirmation



Charles Duhigg, *The Power of Habit*

- Habits form due to positive feedback loop
- Key to have
  - Cue
  - Routine
  - Reward
  - Craving



Thinking Fast vs. Slow

2 x 5 = ?

Capital of France?

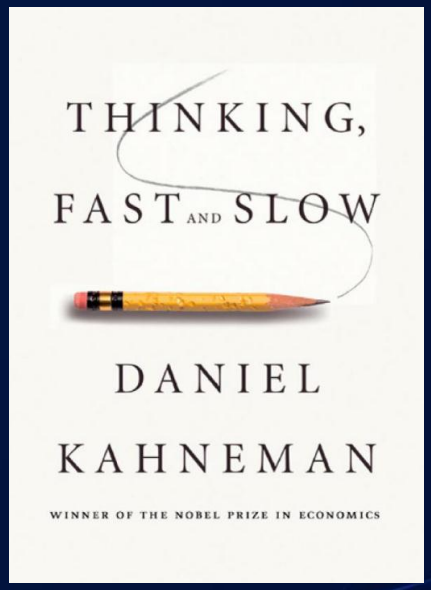
9 x 27 = ?

Capital of Estonia?

Microsoft Research

What is Natural?

- Brain has fast mode and slow mode
- When something is natural, reaction from fast brain
- Instinctive, without deliberation, and feels *natural*



No Technology is Inherently Natural

- Speech and writing require years of learning
- Community and society learn to adapt technology
- Natural user experience requires adapting what's already learned
- New learning require strong enough rewards and smooth ramp up

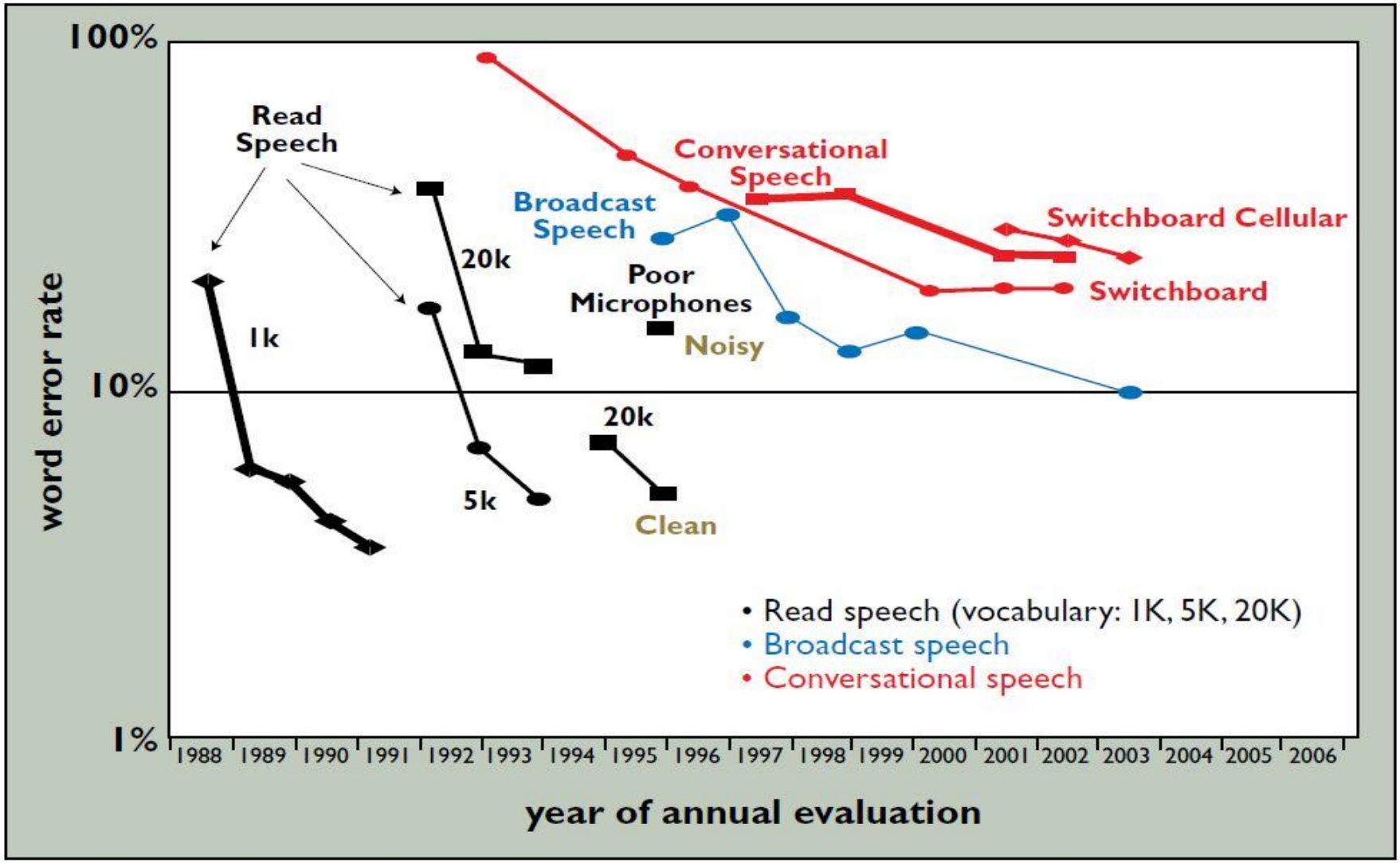
Talk Outline

- Introduction
- Factors for Success
  - Reliability
  - Delivered value
  - Frequency of use
- Paths from Research to Product
  - Kinect Based Speech Recognition
  - Speech as a 1st Class Data Type
  - Bridging the Language Barrier
- Opportunities for Research
- Summary

Microsoft Research

Factors for Success

- Reliability



Commercial Deployments | Collect Call | LV Call Center Automation | Dictation

Xuedong Huang, Li Deng, CACM 2004

Microsoft Research

Bell Labs Voice Call Transactions

- VRCP (Voice Recognition Call Processing)
  - 1 B calls per year (1992)
- Voice Prompter
  - 900 M calls/year (1992)
- SDN/NRA
  - 250 M calls/year (1996)
- Universal Card
  - 50 M calls/year (1995)
- MovieFone
  - 40 M calls/year (1999)
- Talking Call Waiting
  - ~110 M calls/year (2000)

Total: ≥ 2 billion calls/year

VRCP: Task Description

- First major deployment of voice-enabled telecom services
- Recognition of five call types to charge phone calls
  - Collect, calling card, person-to-person, third number, and operator
  - Fully automation of all follow-up services
  - Key phrases are often embedded in callers' requests
- Initial field trial in Haywood, CA for data collection
  - Only 75% accuracy which was much below expectation
  - 95% is the minimum accuracy for deployment
  - One quarter of the speech examples contain extraneous speech
- A key patent (Lee, Rabiner, and Wilpon) made it possible
  - Automatic training of keyword and non-keyword models
  - A grammar network allows keywords preceded and followed by optional background and non-keyword speech
  - 98% accuracy was obtained within 3 months after the initial trial

VRCP: Fully Deployment

- System development
  - Hardware boards were designed to handle the specific task
  - System integration into the AT&T network starting in 1992
- System deployment
  - Fully deployed in the 48 continental states and still being used
  - Known as 0+ service (dialing 0 followed by 10 numbers)
- System Impact
  - Handle over 1B call transactions a year (30M+ per day)
  - Stand as the most widely used voice-enabled services as of today
  - Lead to many successful automated speech applications
- Societal perception
  - General public: no noticeable difference
  - Union workers: system labeled as an evil empire

Factors for Success

- Reliability
- Delivered value
- Frequency of use

A Tale of Two Applications

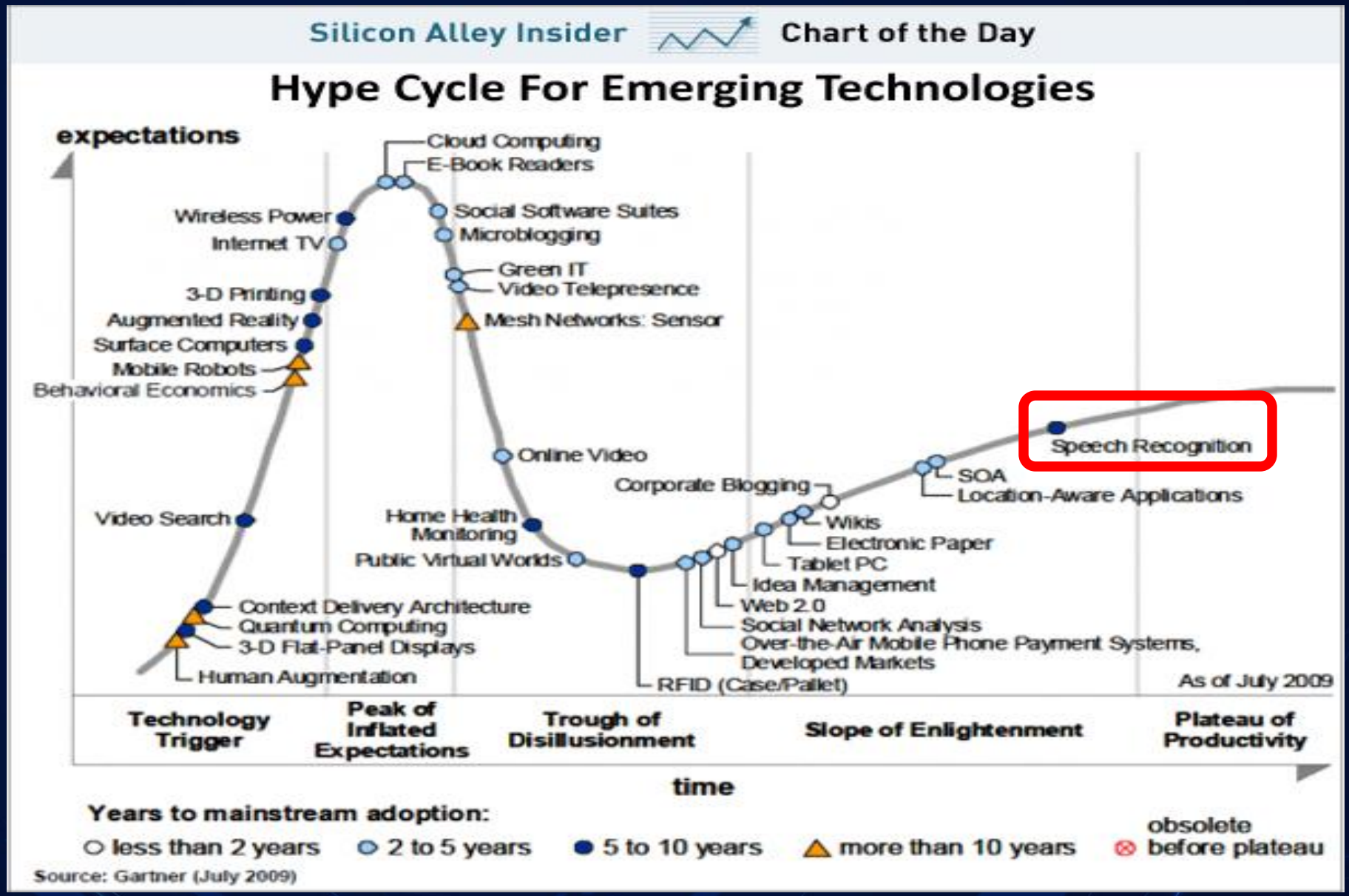
Brokerage	Bank
Check stock quote.	Check balance. Pay bill.
Tens of millions of customers	Tens of millions of customers
Larger grammar.	More complicated dialog.
Several times a day	Once a week

Technology Adoption Lifecycle

Innovator Early Adopter Early Majority Late Majority Laggards

Geoffrey Moore, *Crossing the Chasm*

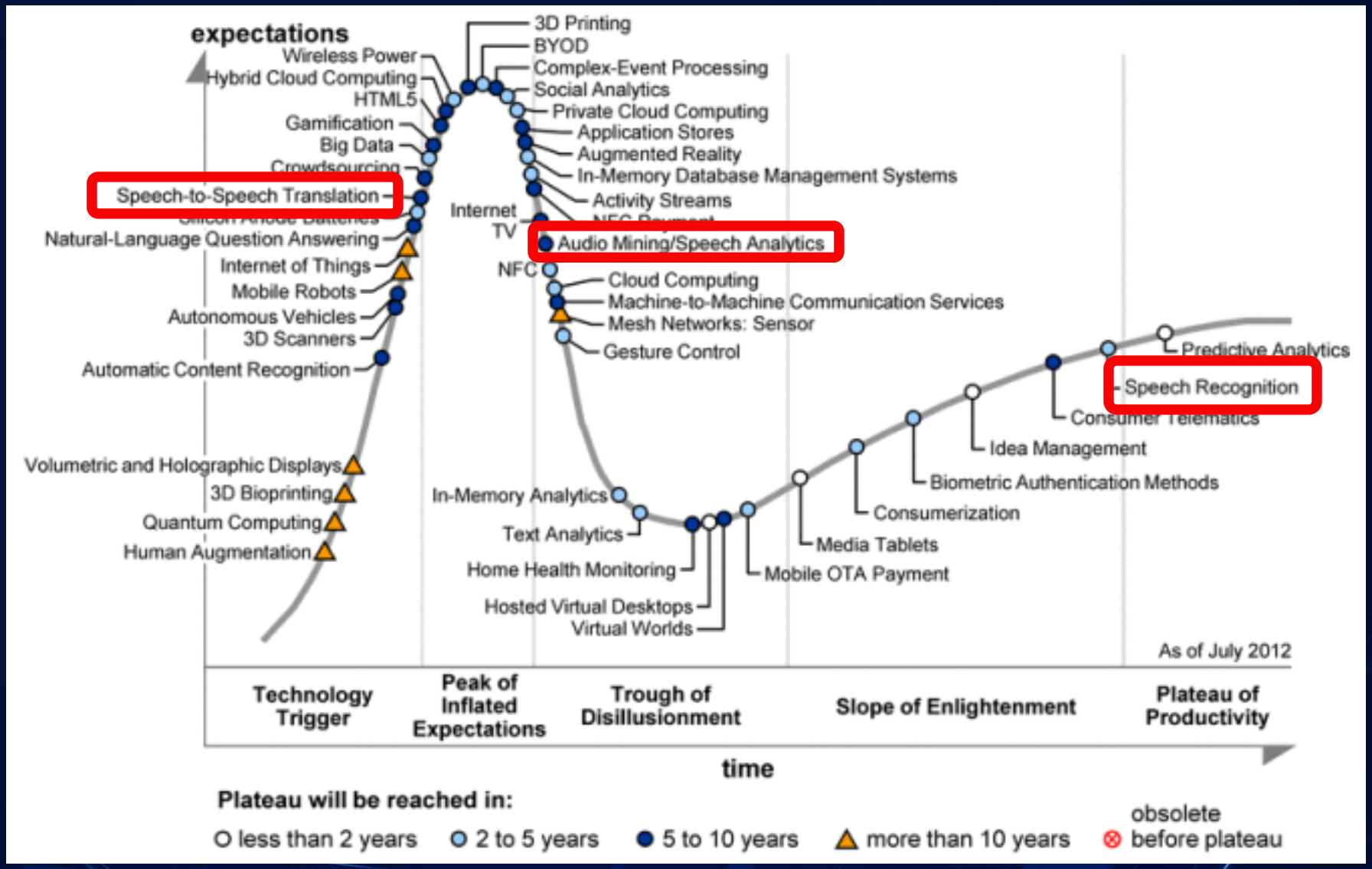
Microsoft Research



http://en.wikipedia.org/wiki/Hype_cycle

The final height of the plateau varies according to whether the technology is broadly applicable or benefits only a niche market.

Microsoft Research



http://en.wikipedia.org/wiki/Hype_cycle

The final height of the plateau varies according to whether the technology is broadly applicable or benefits only a niche market.

Microsoft Research

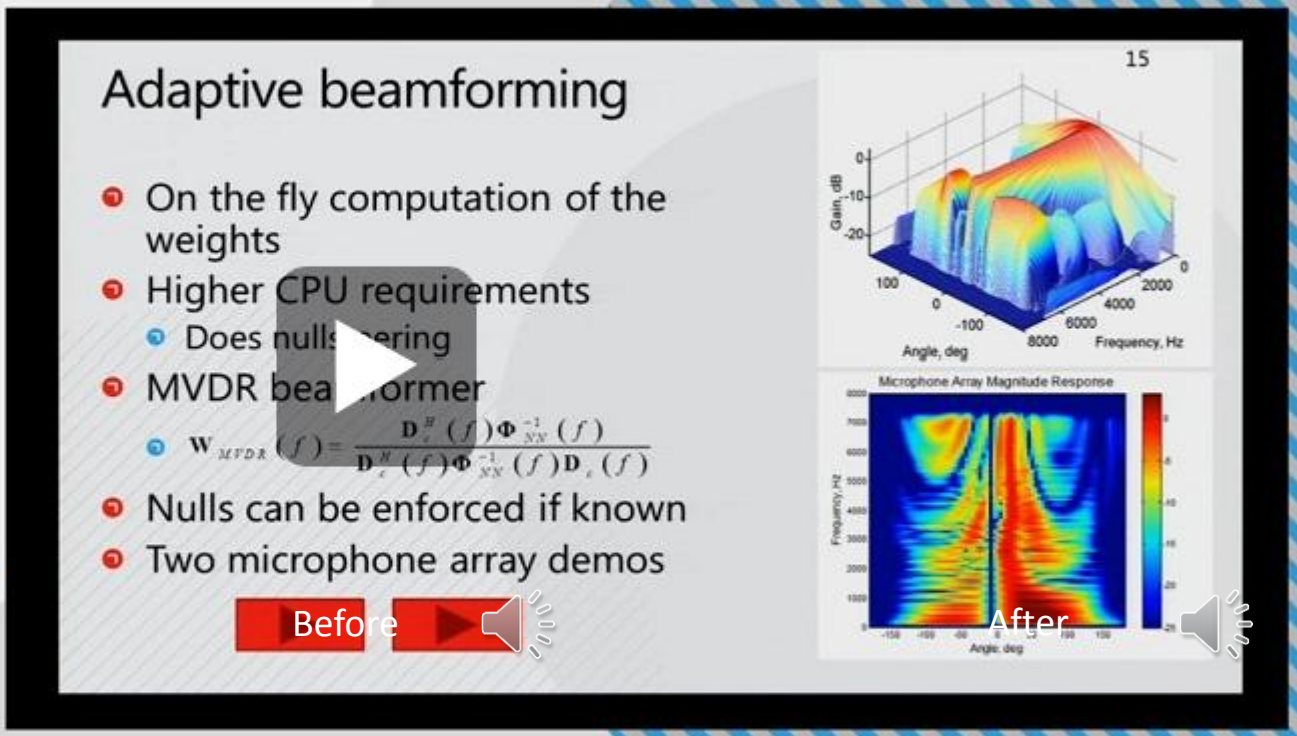
Talk Outline

- Introduction
- Factors for Success
  - Reliability
  - Delivered value
  - Frequency of use
- Paths from Research to Product
  - Kinect Based Speech Recognition
  - Speech as a 1st Class Data Type
  - Bridging the Language Barrier
- Opportunities for Research
- Summary

Paths from Research to Product

- Kinect Speech Recognition
- Speech as 1st Class Data Type
- Bridging the Language Barrier

Noise Robustness for Reliability



http://channel9.msdn.com/events/MIX/MIX11/RES01

59 minutes, 59 seconds

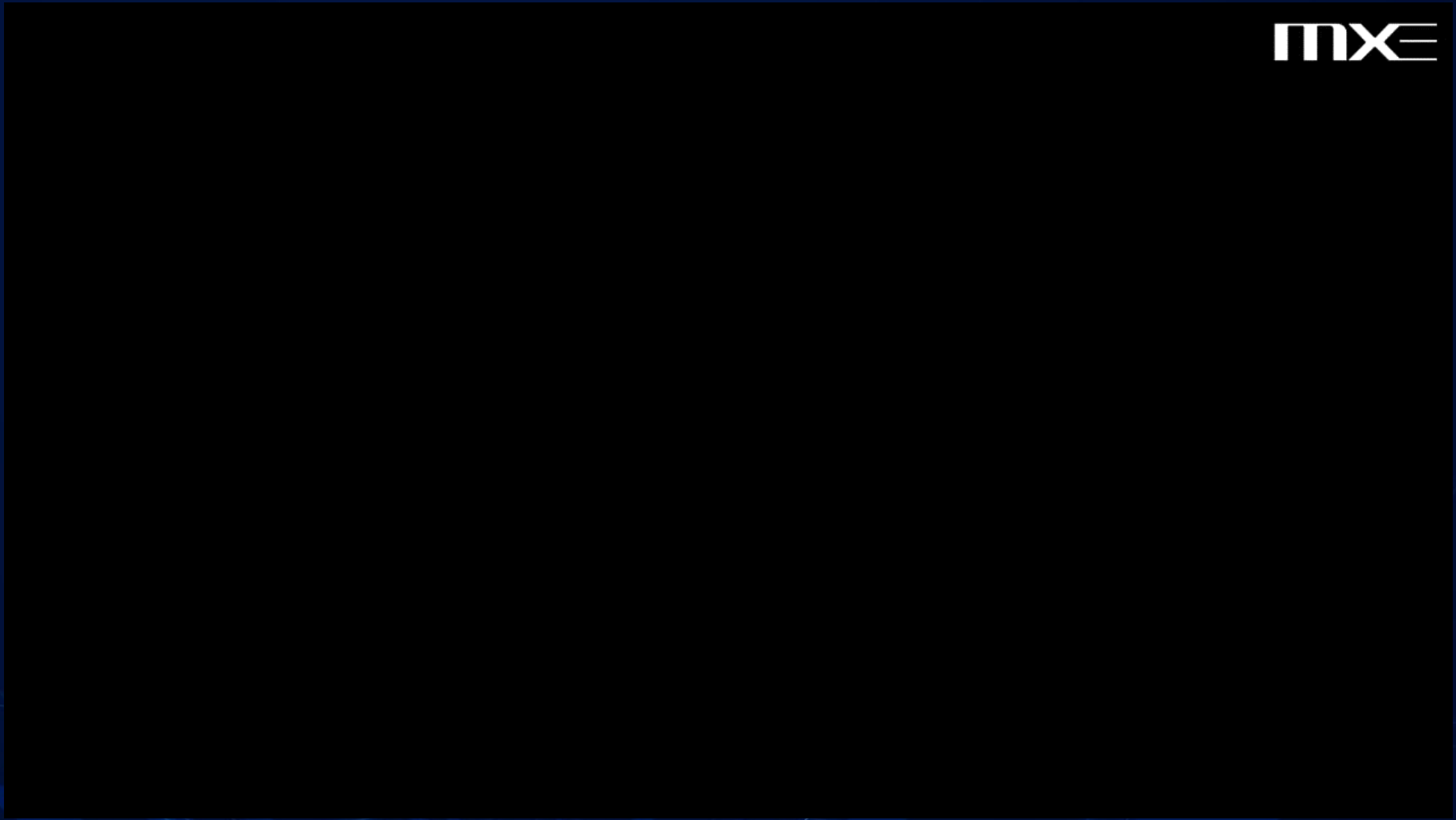
FIFA 2K13, Kinect Speech Recognition

Speech Commands in Games

- Madden Football 2013
- NBA 2013
- Skyrim
- Dance Central 3

- Across multiple languages
- Millions of tokens collected from users

Kinect Speech Recognition for UI



Kinect Speech Recognition for UI

- Search for content and apps
- Multimodal
- Challenges:
  - Scaling across languages
  - Fine tuning grammars and thresholds
  - Designing responsive interfaces

Talk Outline

- Introduction
- Factors for Success
  - Reliability
  - Delivered value
  - Frequency of use
- Paths from Research to Product
  - Kinect Based Speech Recognition
  - Speech as a 1st Class Data Type
  - Bridging the Language Barrier
- Opportunities for Research
- Summary

Microsoft Research

Mission for MSRA Speech Group

- 2000: “Improve computing experience for Chinese users using speech technologies, and then extend to Asia and beyond.”

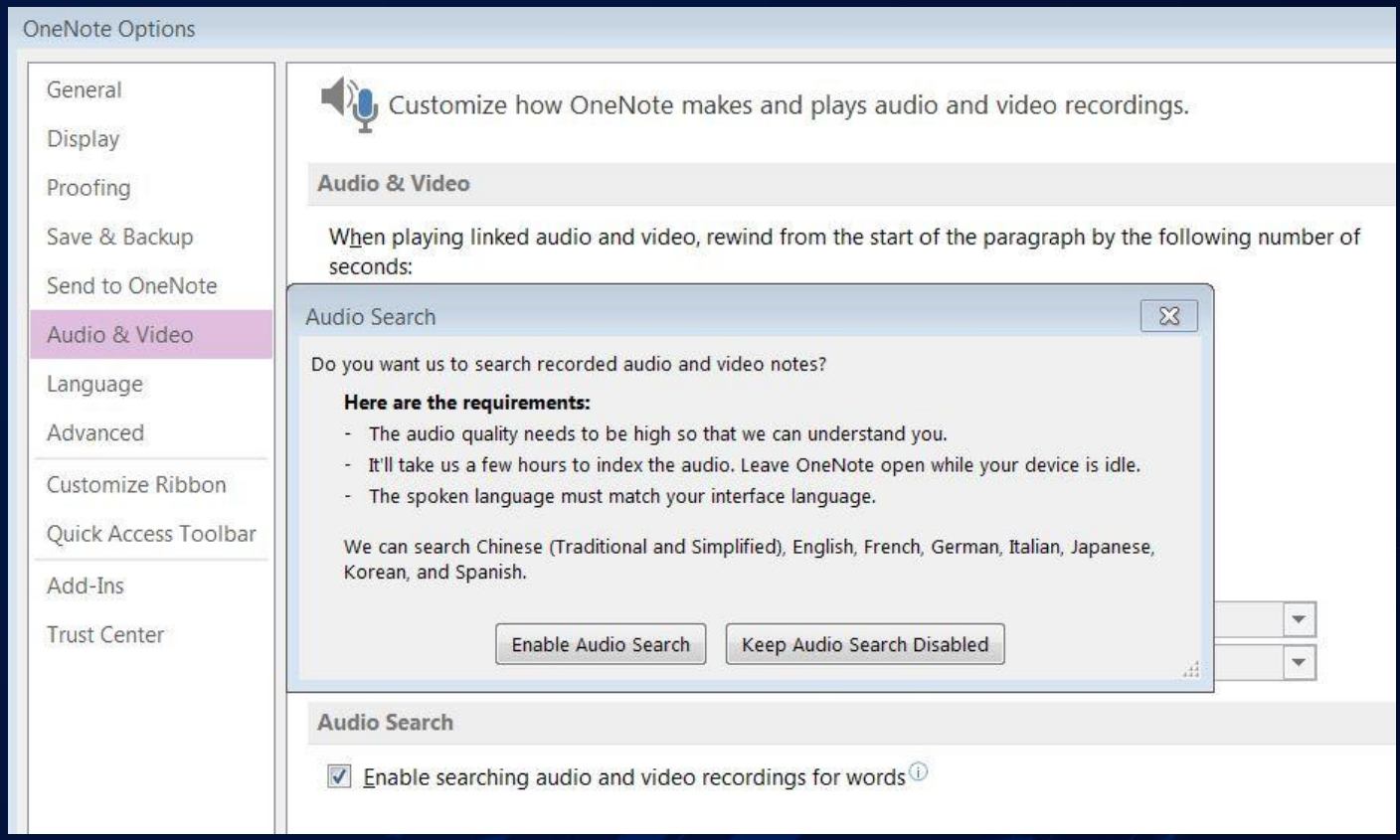
- 2003: “Enrich human-computer and human-human communications with speech technologies.”

Speech as 1st Class Data Type

- OneNote Audio Indexing
- Exchange Voicemail Transcription
- MAVIS Audio Indexing

OneNote Audio Indexing

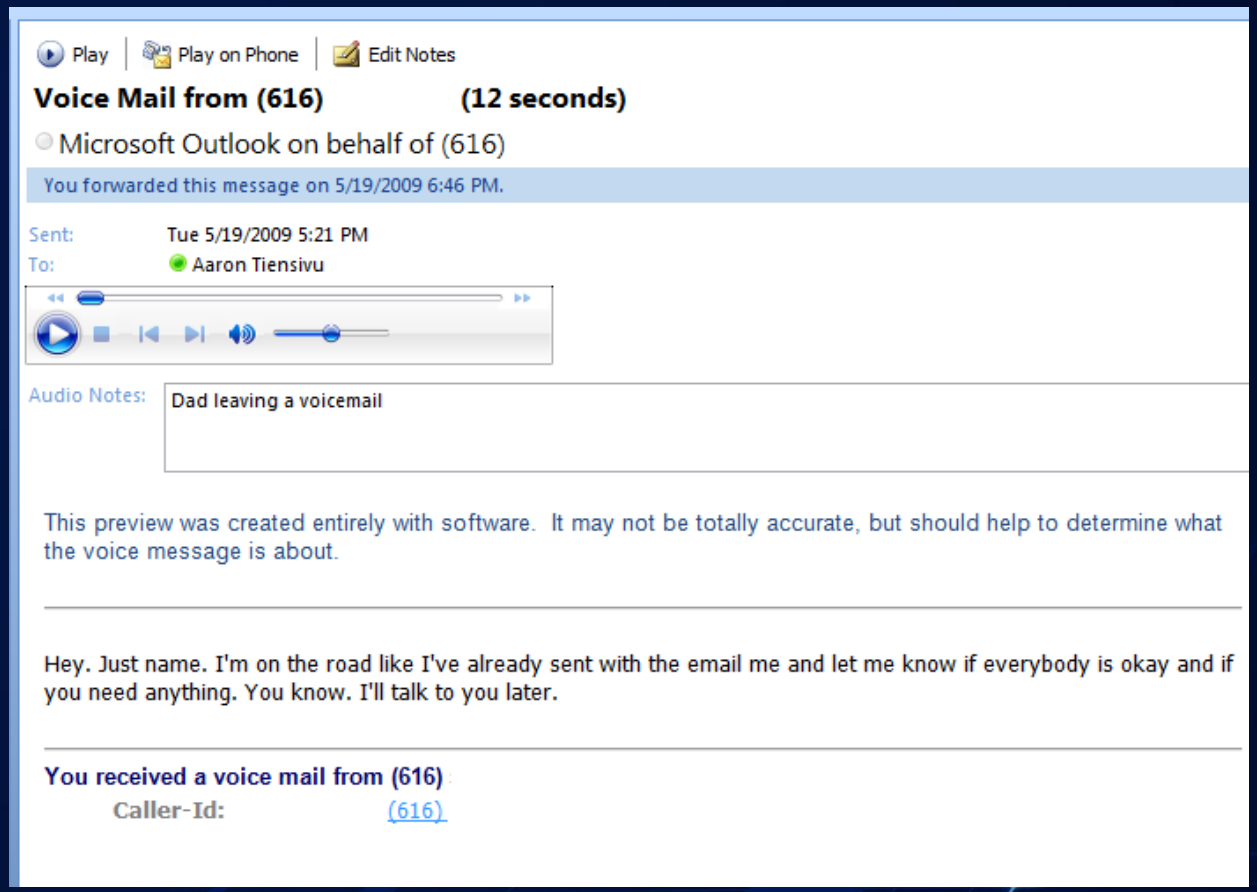
- Introduced in OneNote 2007
- Challenges:
  - Microphone frequently unsuitable
  - Lack of transcription



Microsoft Research

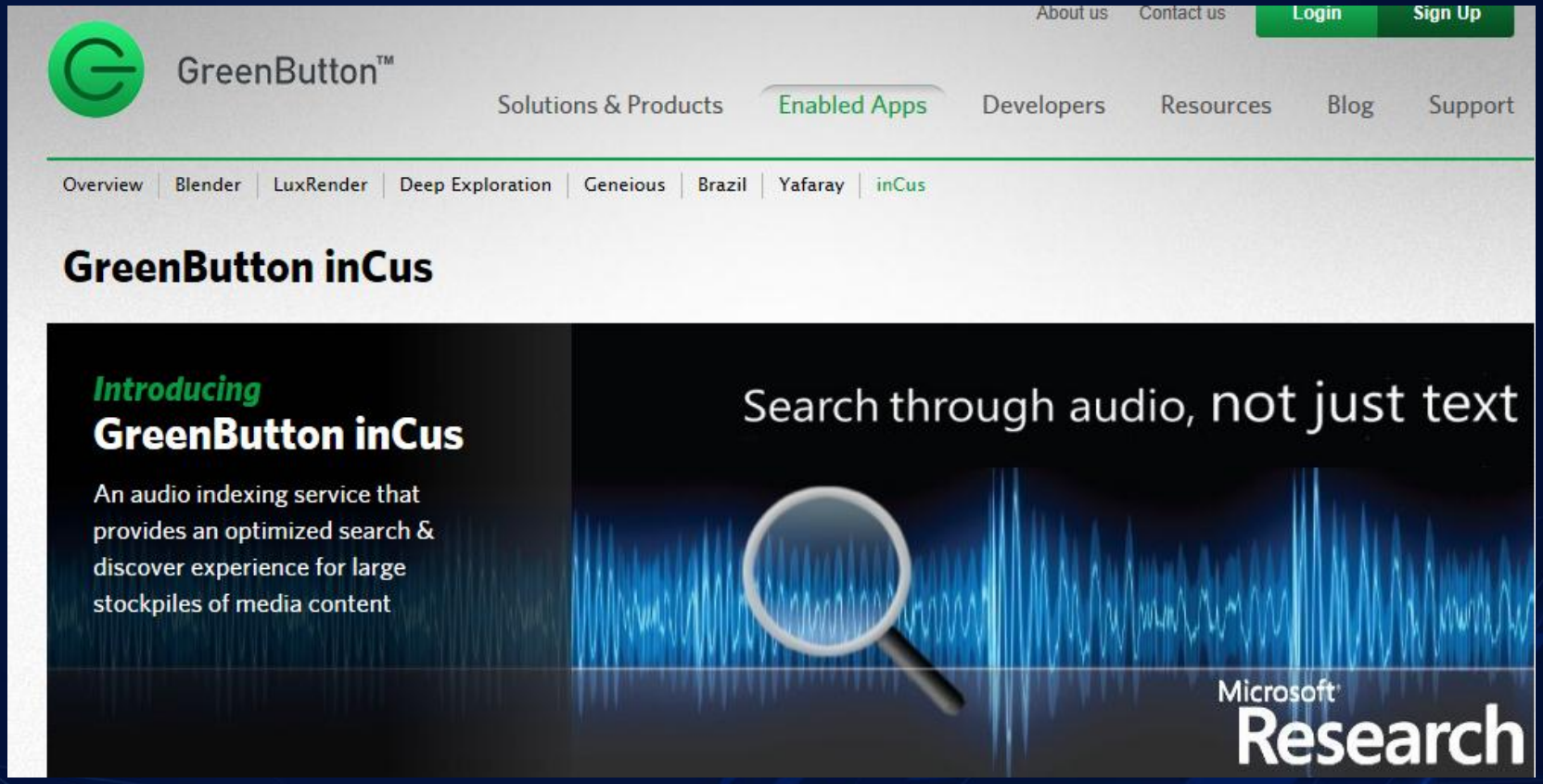
Exchange Voicemail Transcription

- Introduced in Exchange 2010
- Covers tier 1 languages



Microsoft Research

MAVIS (Microsoft Research Audio Video Indexing Service)



http://research.microsoft.com/mavis

Date	Milestone
Jan 2007 **Project inception**	Enable searching of State of Washington Digital Archives content digitized as a result of analog tapes going bad. Started with 100 hours of their content, they now have over 25,000 hours indexed and searchable on their site.
2008 **Move to a cloud service**	Given the computational complexity of speech recognition and the strategy to move to cloud services MAVIS was move to Azure and became an Azure service.
2008 - 2011 **Field trial expansion**	MAVIS is used on a trial basis at customer sites in Government, Education, Medical and Corporate domains.
Feb 2011 - **Launch of ScienceCinema**	US department of energy Office Of Science and Technology Launches the ScienceCinema site with over 1500 hours of Scientific video content to be searchable using MAVIS
April 2012 - **Launch of MAVIS as a commercial service**	Due to successful deployments and increased demand MAVIS is launched as a commercial service through a Microsoft partner.

Recent Progress on MAVIS

- Over 100,000 hours of video indexed to date
- Rapid adoption of new technology: DNN deployed in June 2012
  - 10 – 20% WER reduction
  - 30% faster processing time

MicrosoftVideoWeb
Powered by MAVIS and Microsoft Research Asia

Type a query to search audio and text ...

Microsoft Research

All | PDC build 2011 MIX TechReady MGX Engineering Forum | AcademyMobile Channel 9 | MSN Video

Search through Videos from selected Microsoft sites

We search through the audio, not just text. Search through the sound tracks of about 5000 hours of Microsoft videos including PDC, MIX and Channel9. Click on search-result text snippets to navigate directly to keyword matches in the video. Type a query into the search box above, or try one of these example queries:

cloud computing *energy efficient* *azure storage* *virtualization management* *natural user interface* *readyboost* *volcano*

Powered by MAVIS and Microsoft Research.

Terms of Use Privacy Statement Contact Us

Talk Outline

- Introduction
- Factors for Success
  - Reliability
  - Delivered value
  - Frequency of use
- Paths from Research to Product
  - Kinect Based Speech Recognition
  - Speech as a 1st Class Data Type
  - Bridging the Language Barrier
- Opportunities for Research
- Summary

Bridging the Language Barrier

- Engkoo
- Bing Dictionary
- Bing Translator
- Speech to Speech Translation

The Growing Language Gap

必应 bing™

网页 | 图片 | 视频 | 资讯 | 地图 | 词典

beginning vs. inception

书签 | 工具 | 帮助 | English

Engkoo 技术提供

忙里偷闲，背着老板巧充电

beginning 同 inception 对比

beginning US: [bɪ'gɪnɪŋ]

词形变化 beginnings

n.
1. 初,当初;开始,端绪,发端;出发点
2. 本原,起源
3. 早期阶段
4. 起头部分

v.
1. "begin"的现在分词

网络释义:
1. 起点
http://bbs.englishcn.com/archiver/tid-7761.html
2. 引入

类别: 全部 来源: 全部 难度: 全部

1. But she said it was an example of a "grass-roots" movement **beginning** on the social networking site.
但她认为,这一**始**于社交网站的“草根运动”是一个很好的例子。
http://www.chinadaily.com.cn/language_tips/news/2010-01/11/content_9301...

inception US: [ɪn'sepʃ(ə)n]

词形变化 (无相关结果)

n.
1. 开始,发端
2. (英国剑桥大学)硕士[博士]学位的取得

网络释义:
1. 盗梦空间
http://article.yeeyan.org/view/192532/158189
2. 植入
http://gb.cri.cn/27564/2010/09/02/4945s2977560.htm
3. 最初
http://voa.hjenglish.com/doc/1592609

类别: 全部 来源: 全部 难度: 全部

1. Like the Bond films, **Inception** was shot in various locations aro[illegible] including Morocco, France, Japan and Canada.
点击获取释义
像邦德的电影一样,**Inception**是在全球各地拍摄了各种场景,其中包括摩洛哥,法国,日本和加拿大.
http://bbs.ebigear.com/viewthread.php?tid=128975&extra=&ordertype=2

必应 bing™

网页 | 图片 | 视频 | 资讯 | 地图 | 词典

不管你信不信,反正我是信了。

首页 | 不管你信不信,反正...

不管你信不信,反正我是信了。

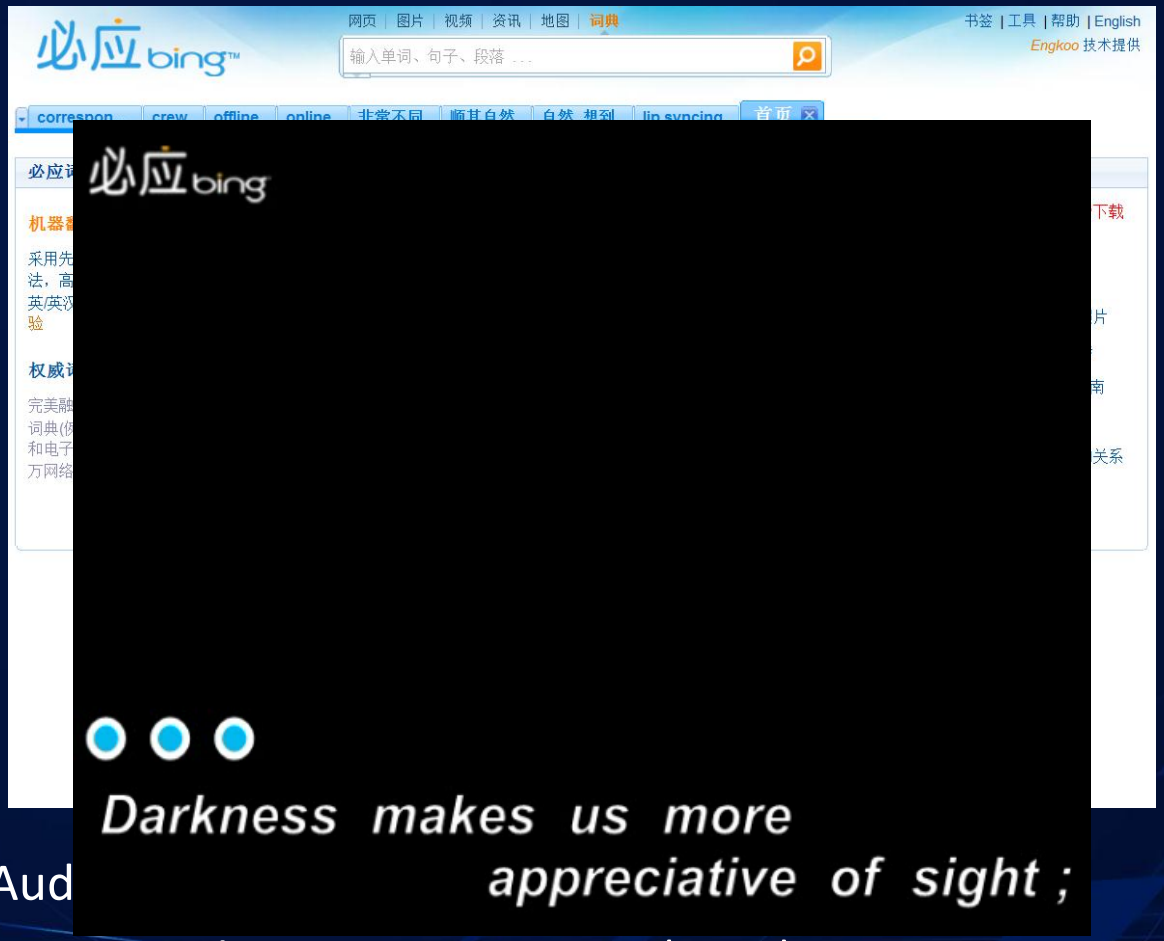
报告问题或瑕疵

计算机翻译:

Whether you believe it or not, anyway, I believe.

不管你信不信,反正我是信了.

Talking Head on Bing Dictionary



- Text-to-Aud
- Computer Assisted Language Learning (CALL)
- More engaging user experience

Microsoft Research

Lijuan Wang, et al, "Computerized Audio-Visual Language Learning", IEEE Computer, 2012

Bing Translator on Windows Phone

Microsoft Research

Good Market Response

- Over 1 million downloads to date
- Rating of 4+ stars

Follow User Feedback

“This app is amazing. Great for translating text using the camera, but the voice translator is especially cool, particularly if you don't know how a word is spelled.”

Speech to Speech Translation

<table>
<tr><th></th><th>Speech Recognition</th><th>Machine Translation</th><th>Speech Synthesis</th></tr>
<tr><td rowspan="4">Technical Challenges</td><td colspan="2">Error Rate</td><td rowspan="3">Cross-lingual Personalization (un-transcribed)</td></tr>
<tr><td colspan="2">Domain Variability</td></tr>
<tr><td colspan="2">Spoken Language Artifacts</td></tr>
<tr><td colspan="3">End-to-end Quality</td></tr>
<tr><td>Scenario Challenges</td><td colspan="2">Data Quality and Training Data</td><td>More engaging TTS</td></tr>
<tr><td rowspan="3">Our Technology</td><td>Deep Neural Network (DNN)</td><td>Fast Domain Adaptation</td><td>Fast Personalization</td></tr>
<tr><td colspan="2">Joint Optimization</td><td rowspan="2">Talking Head</td></tr>
<tr><td colspan="2">Built-in data collection</td></tr>
</table>

Microsoft Research

Cross-language and Personalized TTS

- **Challenge**
  - use monolingual speech data to train personalized TTS in a new language
- **Applications**
  - S2S; Computer Assisted Language Learning (CALL)
- **Problems**
  - Phonetics (segmental) + prosody (supra-segmental)
- **Solution**
  - Same Trajectory Tiling algorithm

Available Speech Data

English

Mandarin

Craig's public lectures, un-transcribed (1 hr)

Reference speaker read speech (2 hr)

Train Personalized TTS across Language

Vocal Tract Length Normalization

Warped trajectory

Trajectory Tiling

Tiled Craig's Mandarin sentences

HMM Training

HMM-based TTS

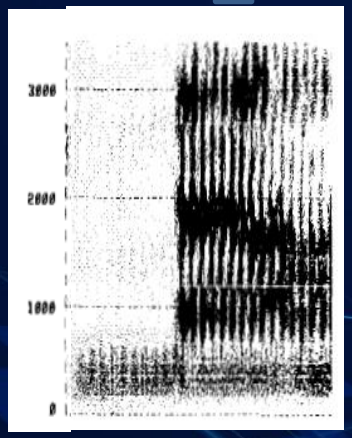
Yao Qian, Frank Soong, and Zhi-jie Yan, "A Unified Trajectory Tiling Approach to High Quality Speech Rendering", IEEE Transactions on ASLP, 2012

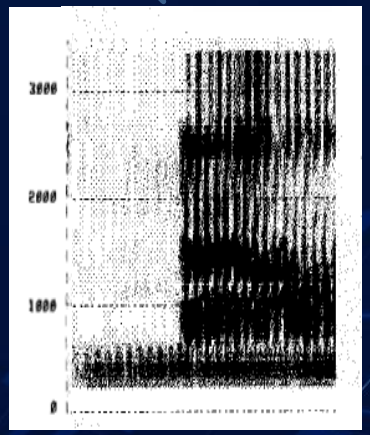
Microsoft Research

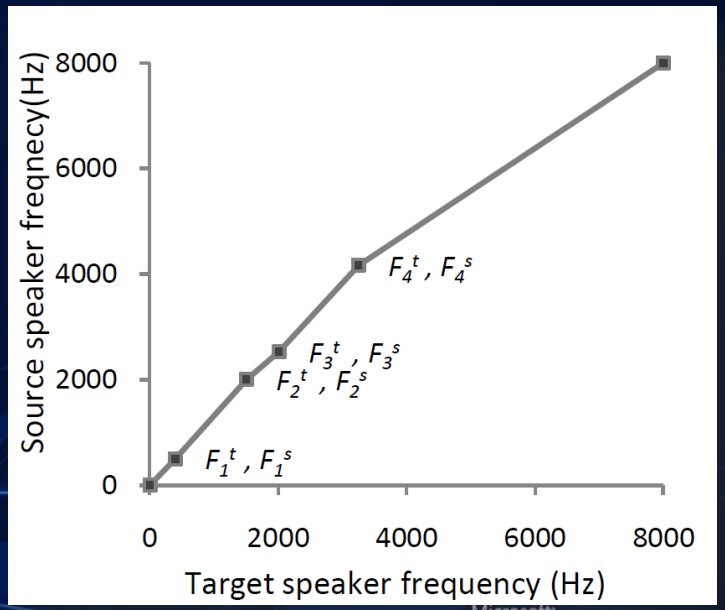
Vocal Tract Length Normalization (VTLN)

- Equalize speaker difference by VTLN
- Warp source speaker's spectrum unto target speaker's
- Warping function: Vowel formant frequency mapping

Frequency Warping







Microsoft Research

Personalized TTS with Talking Head

Lijuan Wang, Yao Qian, Frank Soong *et al*.

Microsoft Research

Live Speech to Speech Translation

Microsoft Research

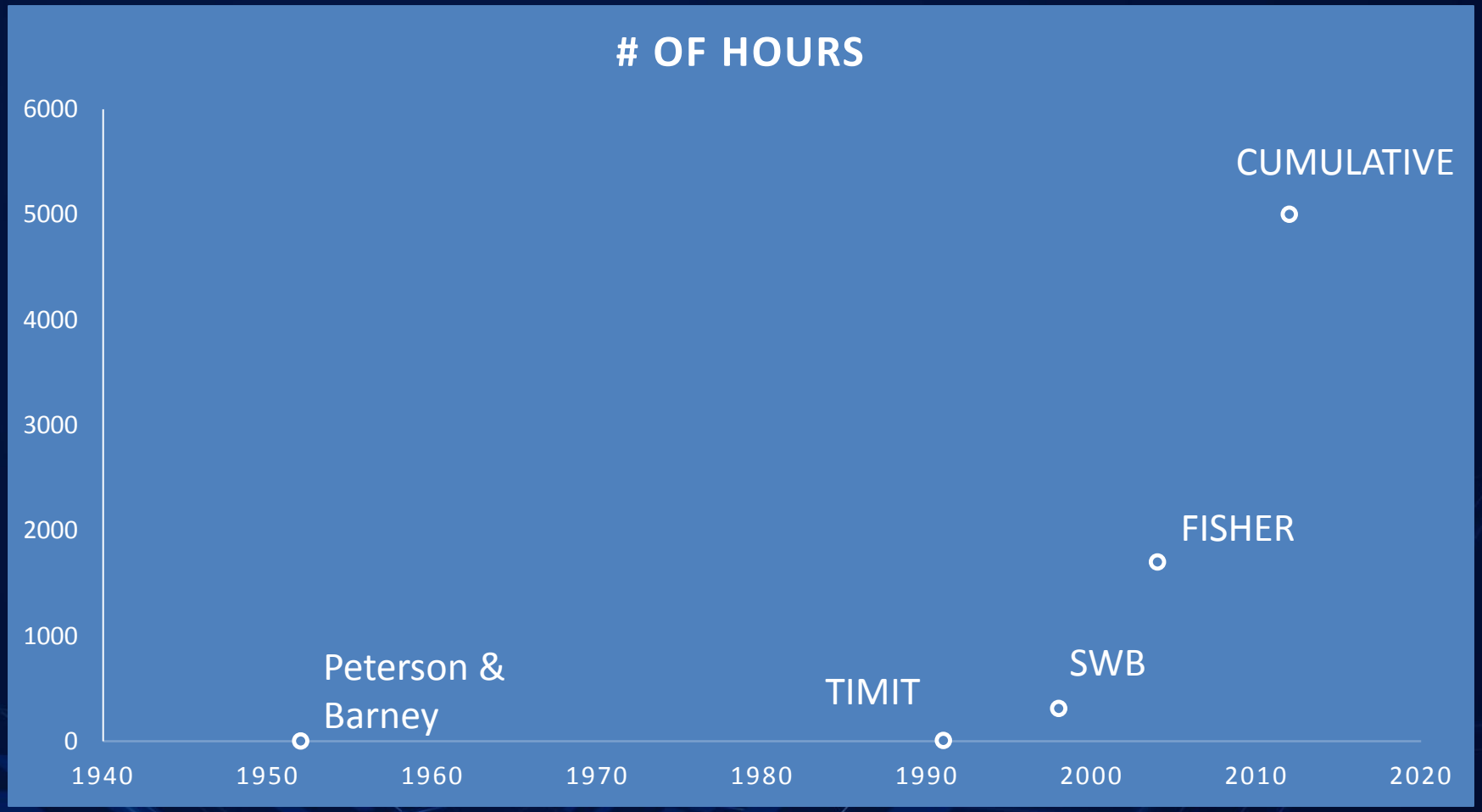
Talk Outline

- Introduction
- Factors for Success
  - Reliability
  - Delivered value
  - Frequency of Paths from Research to Product
  - Kinect Based Speech Recognition
  - Speech as a 1st Class Data Type
  - Bridging the Language Barrier
- Opportunities for Research
- Summary

Opportunities for Research

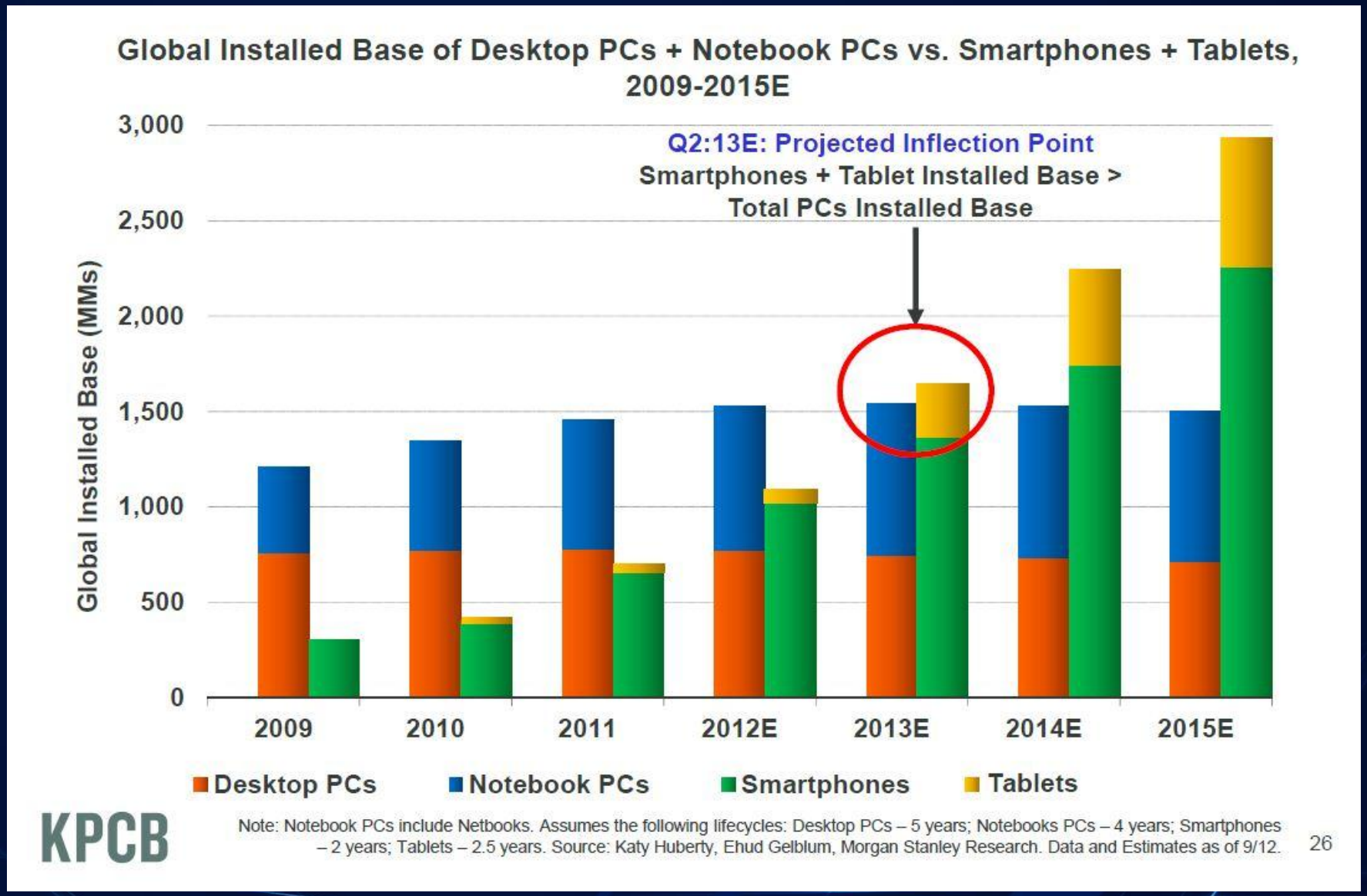
- Scalability of Model and Data
  - Unsupervised training and weakly supervised training
  - Training data scaling from 100's to 1000's of hours
  - Scaling across domain and languages

Growth of Data in Research



LDC:Mark Liberman

Microsoft Research



Mary Meeker Internet Report

Microsoft Research

Opportunities for Research

- Scalability of Model and Data
  - Unsupervised training and weakly supervised training
  - Training data scaling from 100's to 1000's of hours
  - Scaling across domain and languages
- Leverage cross discipline knowledge
  - Natural language processing
  - Knowledge database

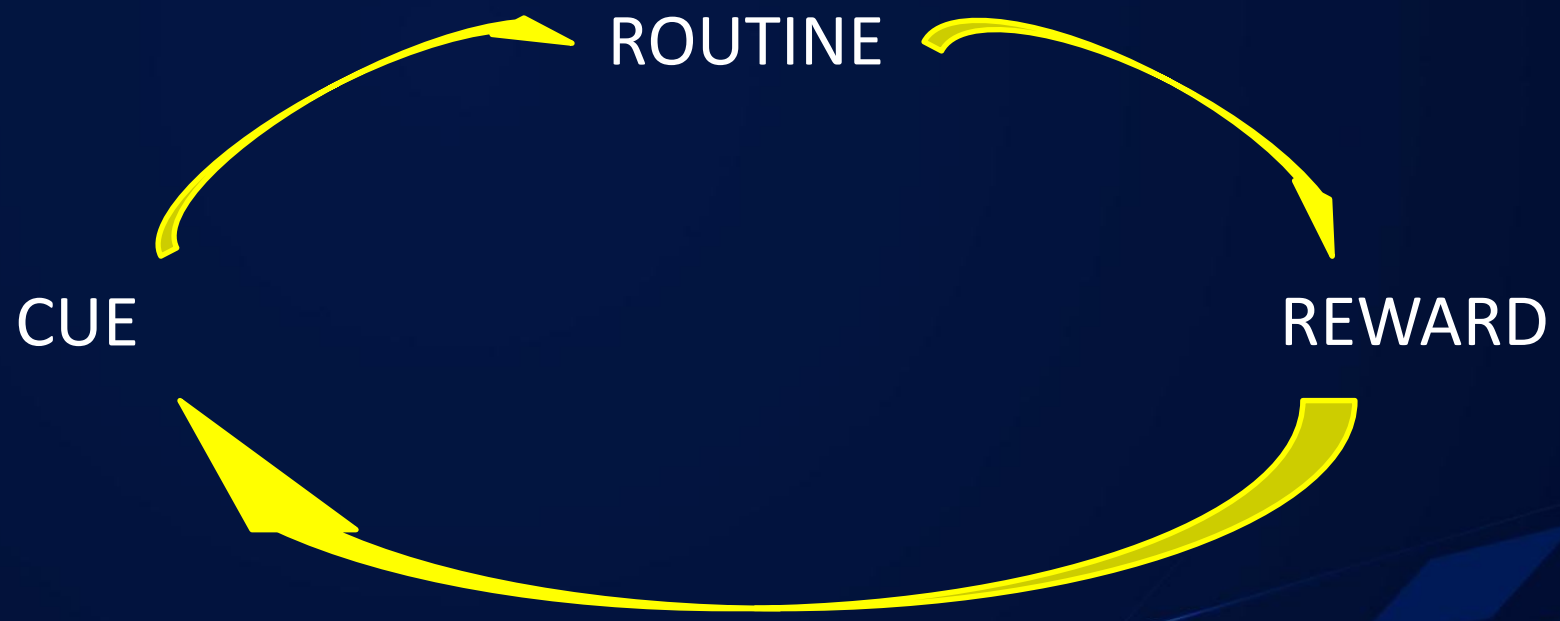
IW84U

Microsoft Research

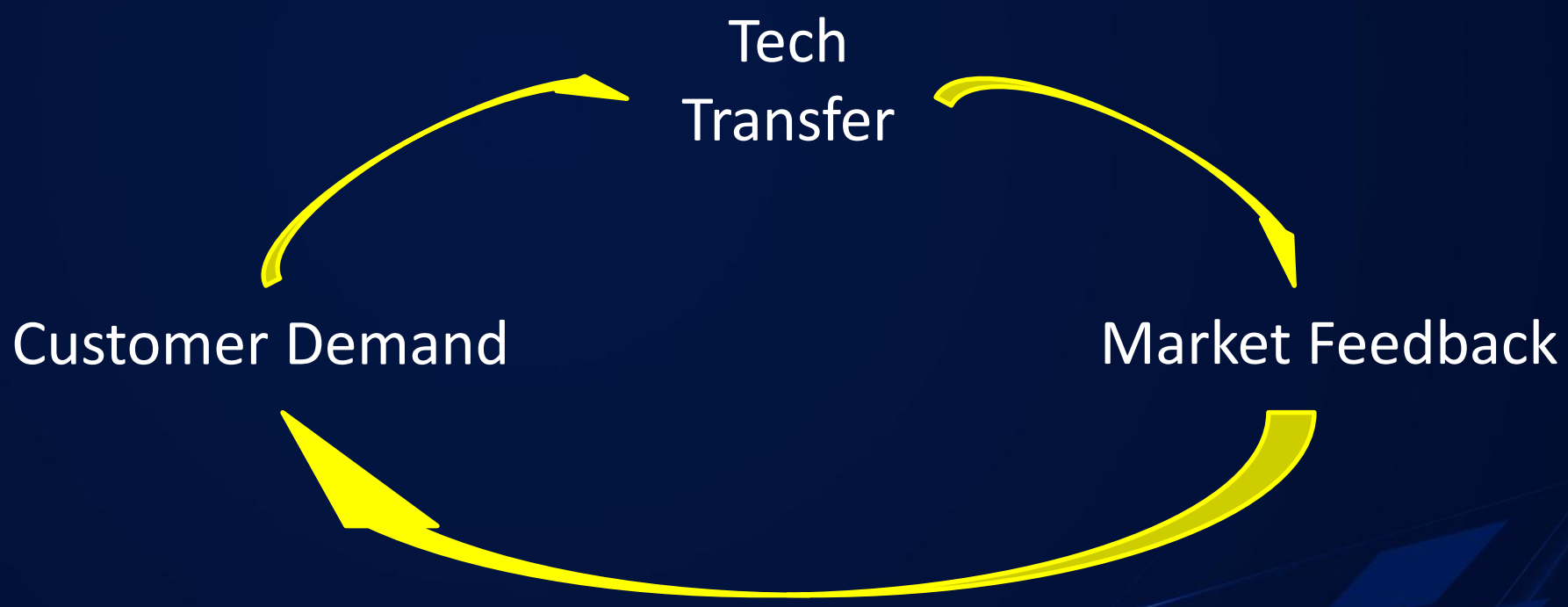
Summary

- From research to product
  - Shipping is only the first step
  - Scaling to large number of users much harder

Developing Habits



Habits of Shipping



Microsoft Research

Conclusion

- Habit formation requires reliability and rewards
- Takes patience, persistence, perceptiveness
- Exciting opportunities for speech technologies in the future!

Thank You!

eric.chang@microsoft.com

Thanks to:
Alex Acero, Behrooz Chitsaz, Li Deng, Qiang Huo, Chin-Hui Lee, Mark Liberman, Yao Qian, Matt Scott, Frank Seide, Frank Soong, Ivan Tashev, Dong Yu, Lijuan Wang , Chris Wendt

Microsoft Research