Speech Internationalization

Voice Search in Many Languages

Martin Jansche

Software Engineer, Google New York

mjansche@google.com

Joint work with Linne Ha, Pedro Moreno, and many, many others

In a nutshell...

Speech and Language technology is directly *about* language. The need to deal with different languages should be readily apparent.

In addition, speech technology deals with aspects of (spoken) languages and internationalization that can often be ignored in traditional localization projects.

Speech Technology

- Speech-to-Text
  a/k/a Automatic Speech Recognition, ASR, "reco"

- Text-to-Speech
  a/k/a TTS, Speech Synthesis

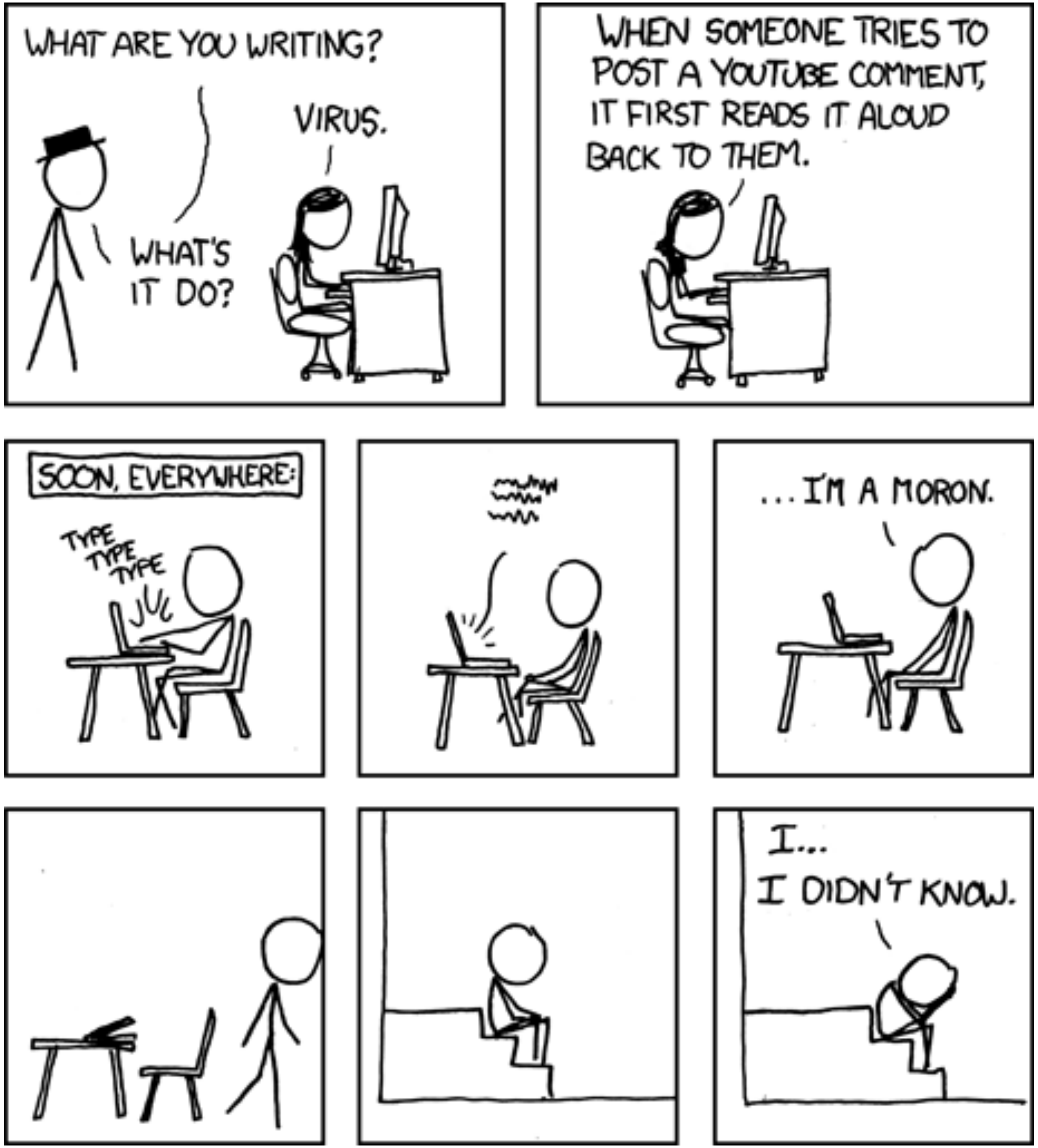
- Dialogue Systems

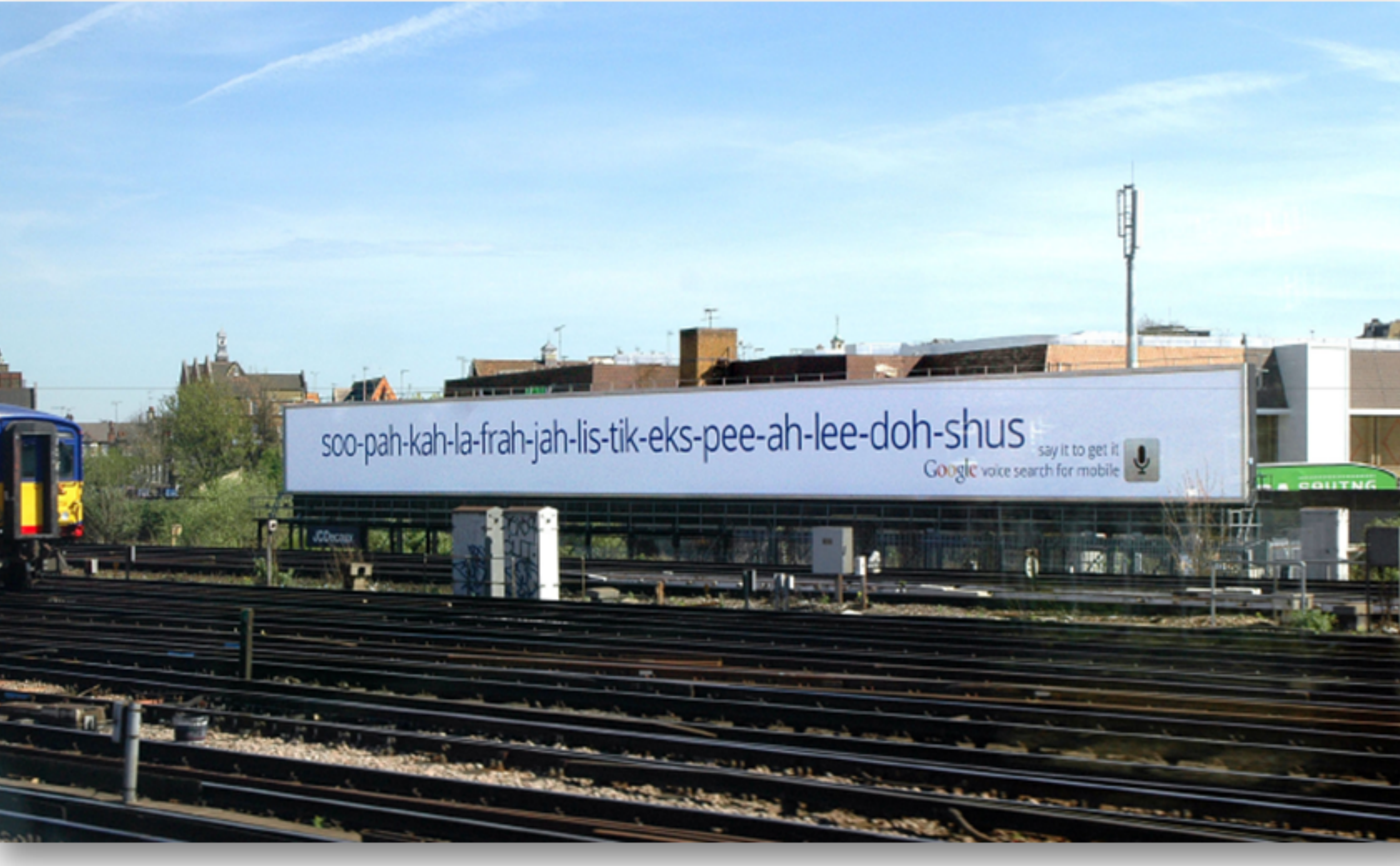
- Lots of other applications, including:
  speech coding and compression, speaker verification,
  speech-to-speech translation, audio indexing, ...

Speech Products from Google

- GOOG-411 (2007-2010)
- YouTube Audio Comment Preview (2008)
- Google Audio Indexing (2008, Google Labs)
- Google Search by Voice for mobile (since 2008)
- Voicemail transcription for Google Voice (since 2009)
- Android Speech Recognition API (since 2009)
- Android Voice Input Method (since 2009)
- YouTube auto captions (since 2009)
- Voice Actions for Android (since 2010)
- Chrome Voice Search (since 2011)

GOOG-411


WHAT ARE YOU WRITING?
VIRUS.
WHAT'S IT DO?
WHEN SOMEONE TRIES TO POST A YOUTUBE COMMENT, IT FIRST READS IT ALOUD BACK TO THEM.
SOON, EVERYWHERE:
TYPE TYPE TYPE
...I'M A MORON.
I...
I DIDN'T KNOW.


soo-pah-kah-la-frah-jah-lis-tik-eks-pee-ah-lee-doh-shus
say it to get it
Google voice search for mobile

Focus of this talk

Speech Recognition

Google Voice Search

Internationalization experience, process, tools, challenges

Languages: Afrikaans - Zulu (EFIGS, CJK, BRIC, ...)

English (Australia, Canada, India, New Zealand, South Africa, UK, US, Pig Latin, generic), French, Italian, German, Spanish (Argentina, ..., Spain, US, ..., Venezuela), Dutch, Mandarin (Mainland China), Mandarin (Taiwan), Cantonese/Yue (Hong Kong), Japanese, Korean, Portuguese (Brazil), Russian, Polish, Czech, Turkish, Indonesian, Malay, Afrikaans, Zulu, Latin

"Language" vs. "dialect" redux

Speech Recognition l10n/i18n

The Hot 300 & The Cold 6,000

There are ~300 languages with more than 1M speakers each. Cumulatively, they cover >95% of the world's population. A these: the most important languages of India; virtually all varieties of Chinese; 10 of the 11 official languages of South Africa; many of the smaller languages of Europe.

Significant commercial interest behind the top 10, 40, 300 languages. Technology developed for these languages will benefit virtually everyone, and set the stage for the long tail.

There are >6,000 languages in the long tail. See yesterday's Keynote for examples of efforts in this space.

Speech Internationalization

For each language, a separate engineering effort to bring up a speech recognition model.

Analogous: Design fonts for all/many of the world's scripts.
Compile proposals for adding many new scripts to Unicode.

Need a plan and a process.

If you don't know if you're doing internationalization,
you're not doing internationalization.
— Jeff Sorensen

Main Ingredients

Enforce/create uniformity.
Languages are much more alike than different.
Speech recognizers follow a common blueprint.
Control data acquisition.
Most aspects of data operations managed within the team.
Built specialized tools for all aspects of data acquisition.
Do not sacrifice simplicity for diminishing returns in quality.
Overall systems complexity increases with each language;
don't complicate the system to benefit a single language.

On Uniformity

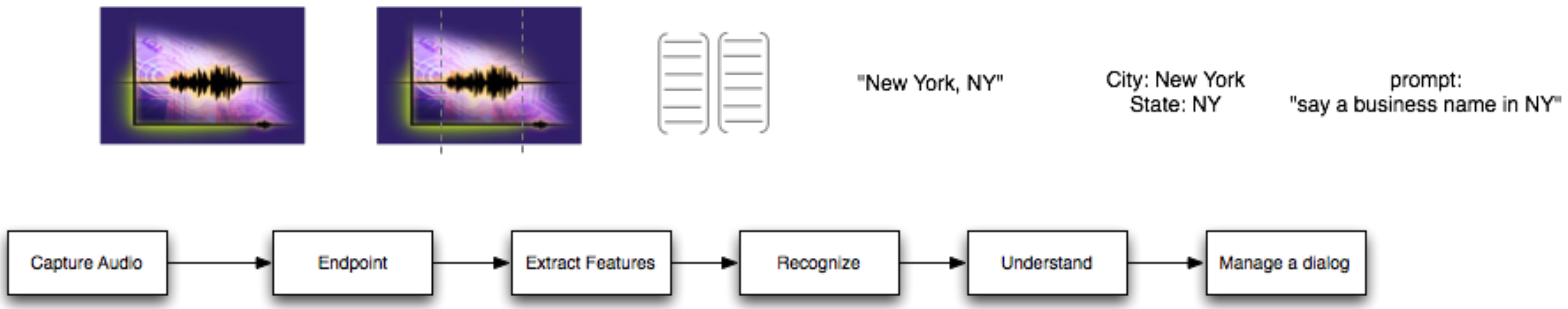
Languages are much more alike than different. Especially as far as the current capabilities of speech technology are concerned.

Differences between closely related languages take on an exaggerated importance in the minds of their speakers, reminiscent of Freud's "narcissism of small differences."

Fitting all languages into our Speech Recognizer blueprint minimizes development effort. This still leaves plenty of room to express language-specific variation.

Speech Recognition Basics

Speech Recognition



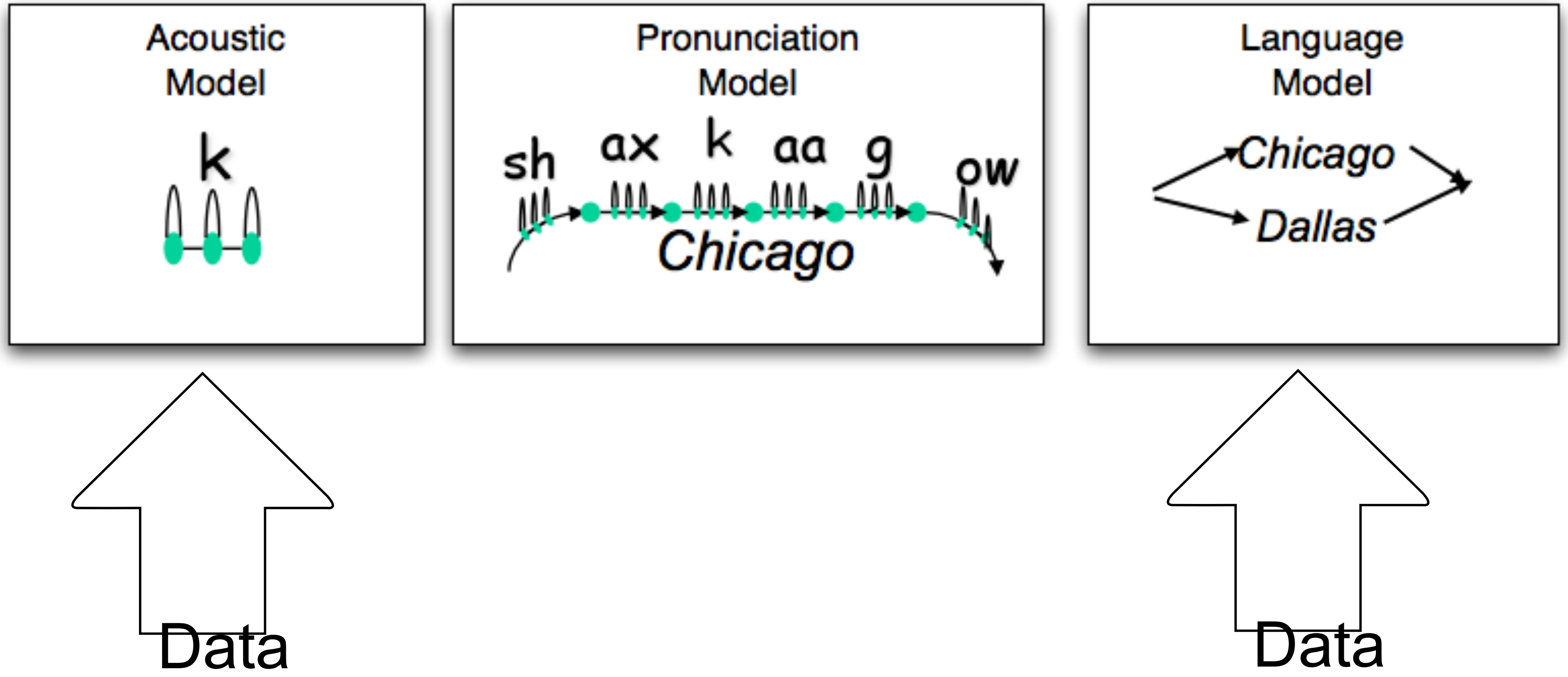
"New York, NY"

City: New York
State: NY

prompt:
"say a business name in NY"

Capture Audio → Endpoint → Extract Features → Recognize → Understand → Manage a dialog

Speech Recognition - Recognize



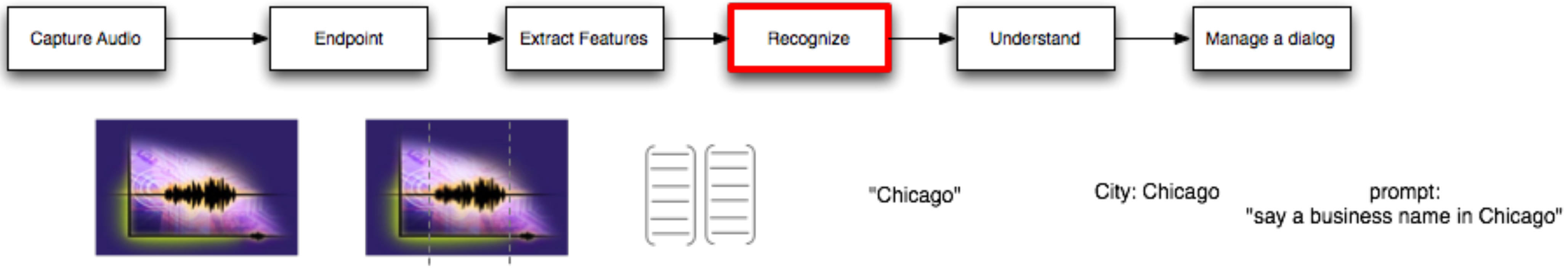
Acoustic Model

Pronunciation Model

Language Model

Data

Data



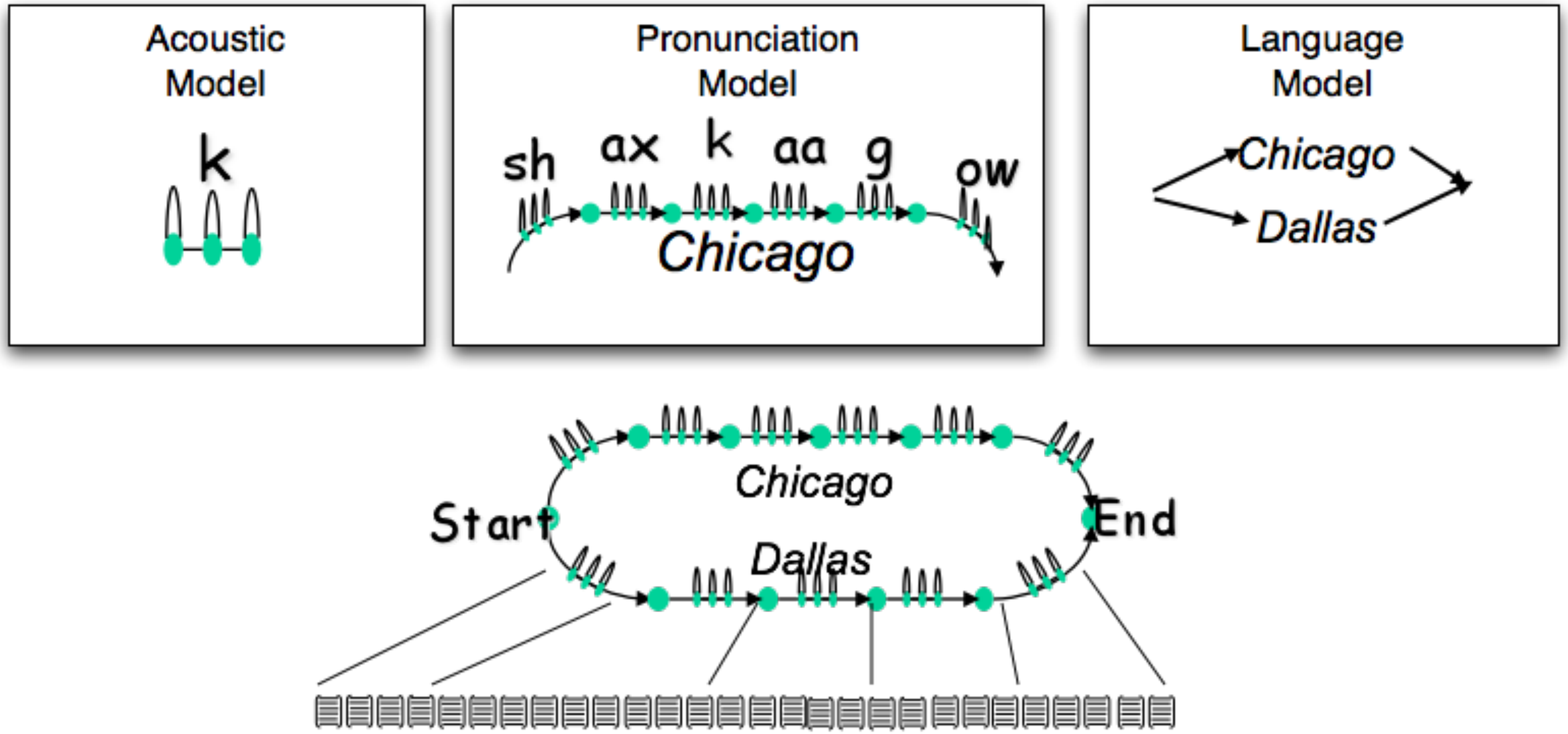
Capture Audio → Endpoint → Extract Features → Recognize → Understand → Manage a dialog

"Chicago"

City: Chicago

prompt:
"say a business name in Chicago"

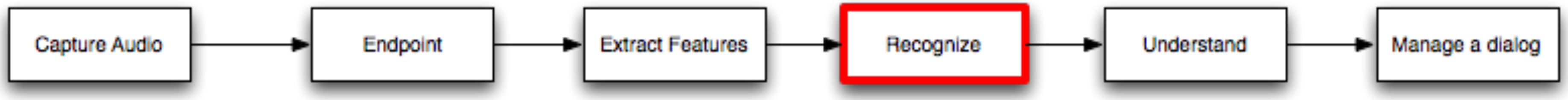
Speech Recognition - Recognize



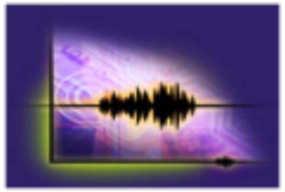
Acoustic Model

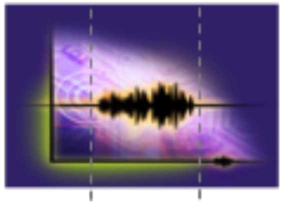
Pronunciation Model

Language Model



Capture Audio → Endpoint → Extract Features → Recognize → Understand → Manage a dialog





"Chicago"

City: Chicago

prompt:
"say a business name in Chicago"

Speech Recognition, mathematically

Our "best guess" what words were uttered, given an utterance.

Find the word string *w* which maximizes Pr(*w* | *audio*), where:

Pr(*words, audio*) :=
Pr(*audio* | *phones*) x Pr(*phones* | *words*) x Pr(*w*□ords)

Acoustic Model ***Pron. Lexicon*** ***Language Model***

The main components of a recognizer

Acoustic model (AM): relates acoustics to contextual phones
Pronunciation lexicon (L): relates phones/phonemes to words
Language model (LM) or Grammar: probability of word strings

Parameter estimation, or "training"

Need to estimate the parameters of these probabilistic models.

Train Acoustic Model on speech utterances and their textual transcripts.

Train Pronunciation Model on pairs of words and their phonetic transcriptions.

Train Language Model on a large text corpus.

The unreasonable effectiveness of data

Halevy, Norvig & Pereira, *IEEE Intelligent Systems*, 2009.

Speech Recognition models are built from fairly general probabilistic components. Most of their power comes from data.

Speech Recognition thrives on and thirsts for data. At the moment, more data generally means better systems.

For example: US English Voice Search trained on thousands of hours of speech, billions of word *n*-grams.

Shalkwyk *et al.*, "Google Search by Voice: A Case Study,"
in *Advances in Speech Recognition*, 2010.

Acoustic Model

Acoustic Data Acquisition: Process

Linne Ha's crowd-sourcing project ("Word of Mouth").

The target is to collect 250,000 utterances from volunteers for each language/dialect project.

Linne's description: "In each country, we found small groups of people who were avid fans of Google products and were part of a large social network, either in local communities or online. We gave them phones and asked them to get voice samples from their friends and family. Everyone was required to sign a consent form and all voice samples were anonymized."

Acoustic Data Acquisition: Tools

DataHound is our client-server application for gathering spoken utterances from volunteers.

Hughes et al., "Building transcribed speech corpora quickly and cheaply for many languages," *Interspeech*, 2010.

Open-source alternative:
De Vries et al., "Woefzela - An open-source platform for ASR data collection in the developing world", *Interspeech*, 2011.

Both are simple Android apps that display randomized prompts and record volunteers reading those prompts. DataHound fetches prompts from and uploads data to a server. Woefzela works entirely offline.

Google
Word of Mouth

Pronunciation Lexicon

Pronunciation Lexicon: Challenges

The Voice Search vocabulary contains frequent unusual items:

- Alphanumeric strings: Nokia E71
- Numbers
- URLs: www.cancertreatmentcentersofamerica.net
- Acronyms and abbreviations: IL, PA, UFO
- Names of international places, persons, organizations, etc.: Eyjafjallajökull
- Foreign words from English, former colonial languages, prestigious nearby languages, regional linguae francae, etc.
- Non-standard words: CLNG

We need to know how these are pronounced.
This is hard, even when native word pronunciations are "easy".

Pronunciation Lexicon: Approaches

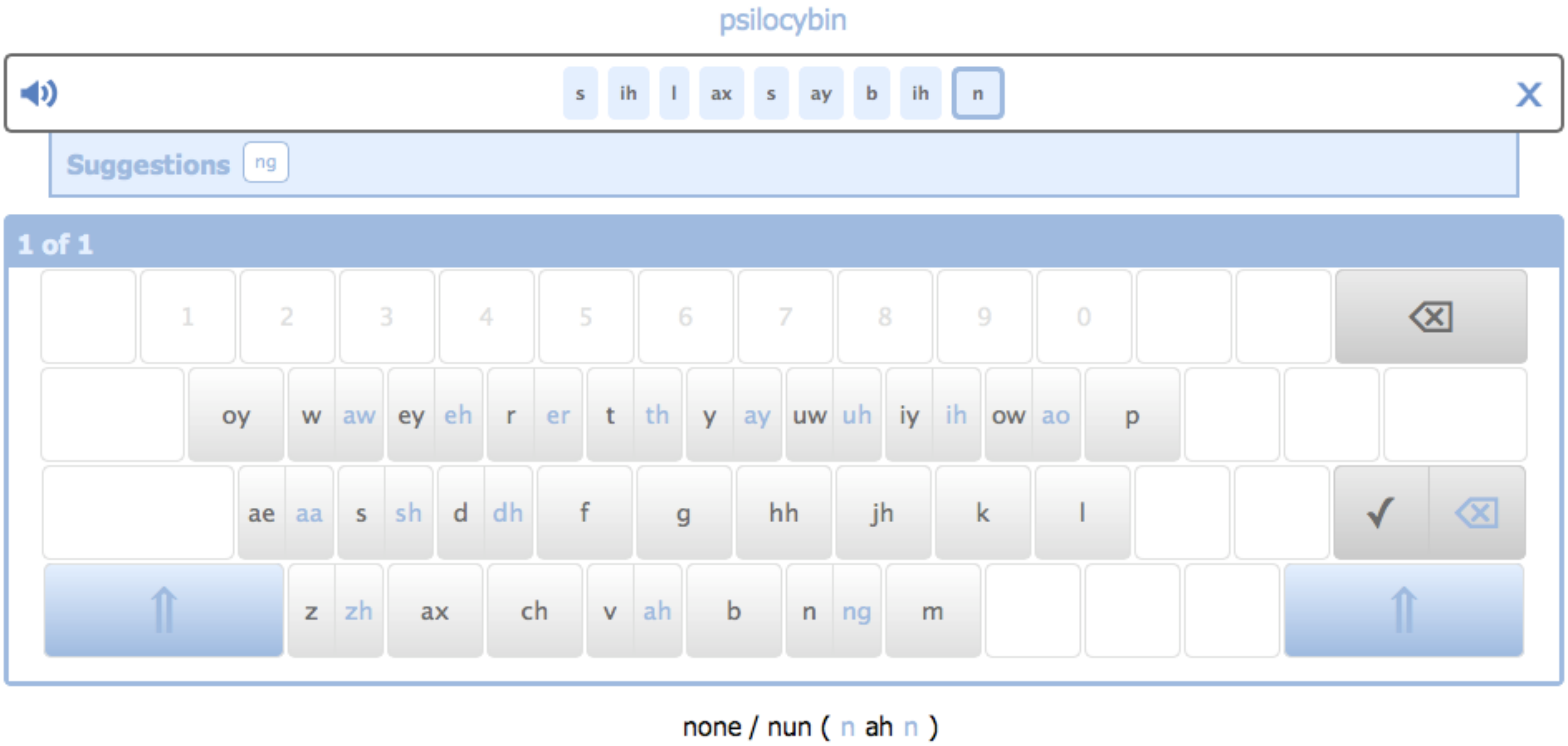
ICU Transforms
Well suited for highly regular orthographies, including Spanish, Indonesian & Malay, Russian, Polish, Czech, ...
We have contributed a few such transforms to CLDR 1.9.

Pronunciation dictionaries
Syllable/character dictionaries for Chinese varieties.
"Word piece" dictionaries for Japanese.

Stochastic pronunciation models
Well suited for the messier orthographies, including English (first and foremost), French, German, Dutch, ...
Require pronunciation dictionaries for training.

Pronunciation Lexicon: Tools

Ainsley *et al.*, "A web-based tool for developing multilingual pronunciation lexicons," *Interspeech*, 2011.



psilocybin

s ih l ax s ay b ih n

Suggestions ng

1 of 1

none / nun (n ah n)

1 2 3 4 5 6 7 8 9 0
oy w aw ey eh r er t th y ay uw uh iy ih ow ao p
ae aa s sh d dh f g hh jh k l
✓
z zh ax ch v ah b n ng m
X

Language Model

Language Model: Challenges

Large amounts of text are relatively easy to come by. But even this is problematic on the other side of the Digital Divide.

Written text is not well matched to what people say.

Text normalization mediates between "written domain" and "spoken domain". For example, "125" can be pronounced like:
one two five
(one) hundred (and) twenty-five
one twenty-five

"$20 books on amazon.com" in the written domain is spoken like "twenty dollar books on amazon dot com".

Text Normalization Challenges

In Japanese, *Nintendo* can be written
in Kanji: 任天堂
in Katakana: ニンテンドー
in Hiragana: にんてんどう
in Romaji: Nintendo *or* Ｎｉｎｔｅｎｄｏ
If a user says "nintendo", which transcript should Voice Search return? Or does it even matter? If we search on google.co.jp for any of these, the top result is always the Nintendo Japan site.

Take this into account for deciding if recognition for Voice Search was successful: Did the user find what they said?

Voice Search Evaluation Criterion

In addition to traditional speech recognition evaluation criteria like sentence accuracy or word error rate, measure how often a spoken query triggered the expected search results.

In other words, evaluate with a Search Engine in the loop.

All modern search engines are powerful text normalizers that can deal with synonyms and variants; missing or extraneous diacritics, hyphens, etc.; common misspellings; etc.

Language Stories

Englishes

Yes, there are many...

Biggest challenge: pronunciation dictionary. English spelling is messy, and that's just for native words. In addition, English as a former colonial language and a global language has influenced and been influenced by many other languages.

Unexpected trivia: Indian English has acquired at least one voiced aspirated (marginal) phoneme, for words like *ghost*.

French, German, Dutch, ...

Biggest challenge: Significant number of non-native (often English) words in common vocabulary.

This goes beyond popular domain names (yahoo, facebook, bing, youtube, spotify, hi5, ...) and other proper names; includes words for everyday concepts (e.g. "handy" meaning cellular/mobile phone in German) whose pronunciation is not predictable using native pronunciation conventions.

French and Spanish, but not Italian or German

Recurring challenge: Recovering diacritics, especially in the text of anonymized queries that form part of the training data for Language Models.

The website of the Académie Française is at academie-francaise.fr, which can be found by typing e.g. [academie francaise] into Google, Bing, etc.

Problem: "academie" and "francaise" are nonstandard spellings that suggest different pronunciations than "académie" or "française".

Solution: Restore diacritics before LM training.

Orthography Recovery

Generate different diacritizations of "francaise", including "francaise", "française", "francaisé", "françaisé", etc.

Look up the probabilities of the differently diacritized utterances in an existing French Language Model (e.g. from Google Translate).

Very similar to (but much easier than) the problem of Arabic diacritic recovery (cf. Eldawy's talk from last year's IUC).

Not done for German (people almost always indicate umlauts somehow) or Italian (accent marks don't produce useful distinctions for us).

French

Problem of the week: Sound change.

French is said to have a palatal nasal, namely the central consonant in *agneau* /aɲo/ 'lamb'. Distinct from a /n/ + /j/ sequence: *cagner* /kaɲe/ vs. *cannier* /kanje/ is a minimal pair.

Our language experts do not have this distinction!

A palatal lateral was lost earlier in Standard French, so that *ail* is now pronounced /aj/ (from earlier /aʎ/, cf. Catalan *all* /aʎ/, Italian *aglio* /aʎʎo/, Latin *allium*).

Contemporary Parisian French seems to have completed the elimination of the palatal consonants (not counting glides).

Afrikaans

Afrikaans: One of the most common words in the language, the indefinite article /ə/, is written 'n.

Or: ŉ (U+0149 LATIN SMALL LETTER N PRECEDED BY APOSTROPHE)

NFC is not sufficient.

NFKC turns U+0149 into U+02BC (MODIFIER LETTER APOSTROPHE) plus n.

Still need to turn U+02BC into U+0027 (APOSTROPHE).

isiZulu

Phonological inventory contains 6 clicks and 2 tones.
Clicks are not especially problematic for the AM.
Tones are never indicated in the normal orthography, so we have no choice but to ignore them, absent a full pronunciation dictionary.

Biggest challenge: Sufficient Zulu data for building a Language Model.

Chinese varieties

Mandarin for Mainland China (cmn-Hans-CN): Arguably more widely spoken as a second language than as a first language, huge variation, interference from first language etc.

Mandarin for Taiwan (cmn-Hant-TW): Very distinct from Taiwanese (nan; but nan-Hant or nan-Latn?). But of course influenced by it...

Cantonese/Yue for Hong Kong (yue-Hant-HK): Necessary for Hong Kong, desirable for Southern China, Chinese expat communities around the world (cf. SF MUNI announcements).

Chinese varieties: Cantonese

Cantonese has 6-10 lexical tones. These correspond to ~3 effective distinctions per syllable.

Biggest problem is heavy mixing of English words: ~1/3 of vocabulary is English. We know the (US/UK) pronunciation of these words, but those are in a different phonology. How should we combine English and Cantonese phoneme inventories?

Sung, Jansche, Moreno, "Deploying Google Search by Voice in Cantonese", *Interspeech*, 2011.

Spanish varieties

Four systems: Spain (es-ES), Mexico (es-MX), Argentina (es-AR), rest of Latin America (es-419).

Determined by linguistic and population factors.

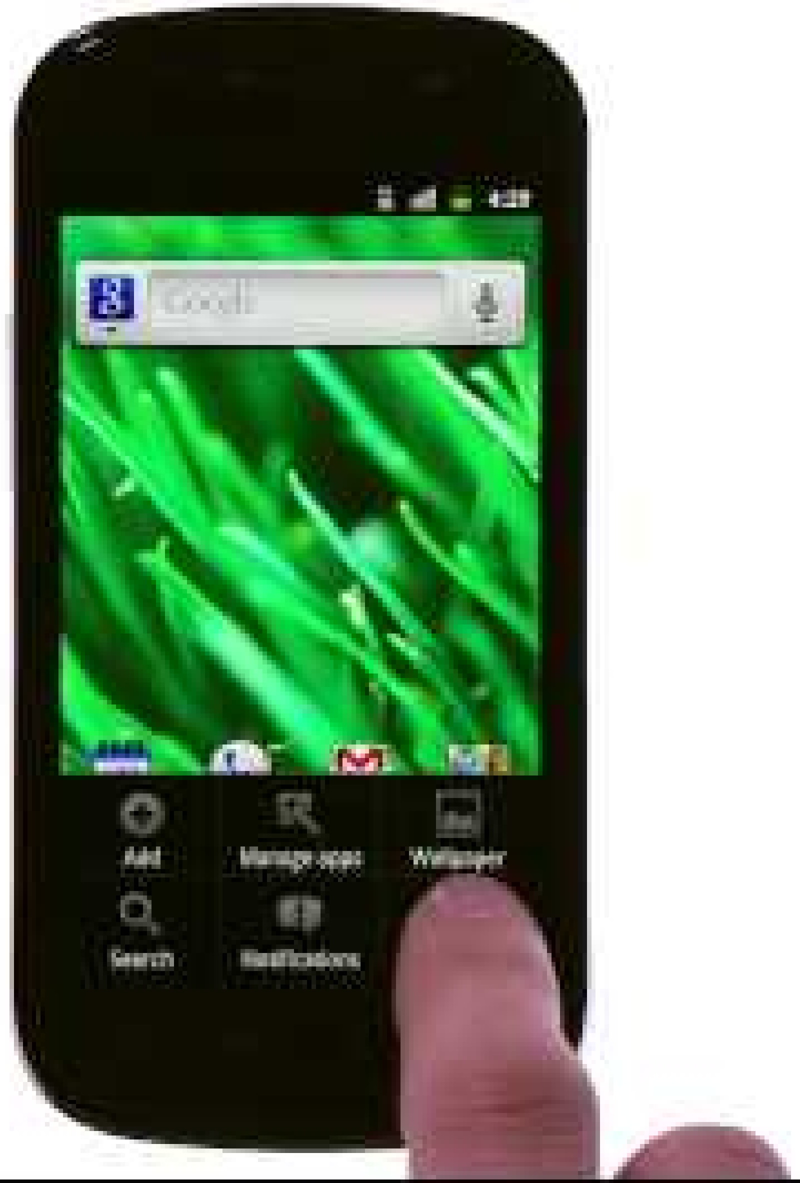
Lexicon handled uniformly across all four systems, despite known phonological and lexical differences.

Latin

A chimera:

- Italian Acoustic Model.
- Classical Latin pronunciation rules targeting modern Italian phonology: consonant length distinction, but no vowel length distinction, no /h/ phoneme, "qu" pronounced /kw/, etc.
- Latin LM trained on classical texts.

Used by Google Translate.


Google
Add
Manage apps
Wallpaper
Search
Notifications

Pig Latin

Really, a Pig-Latin-to-English translator.

Only need an alternative pronunciation lexicon:
"pig" normally pronounced /p ih g/, but pronounced /ih g p ey/ in Pig Latin.

Implemented as a simple post-lexical transformation of ordinary English pronunciations.

(There are 3 major dialects of Pig Latin. We implement the "plain" variety, not the "w" or "h" variety.)

Random War Stories

Acoustic data collection: noise and shenanigans.

Language models and text normalization: The usual kind of encoding conversion/confusion: UTF-8, UTF-16, Latin-2 mistaken for Latin-1 and converted to UTF-8, etc. Locale-dependent case conversion for Turkish.

Tools: Web applications don't automatically work offline.

Android uses Java (<<7) Locale, so it will send "in" for Indonesian (we use "id"). Locale strings captured in the wild are often fanciful ("ja-JA", "en-").

Need quality control at all levels.

Language Selection/Preferences

Do not forget about spoken languages.

Spoken "Chinese" does not primarily fall into "simplified" vs. "traditional". You can have a long conversation and never make a decision about "simplified" vs. "traditional".

Language/country pair is useful, but may not be enough.

"ar-EG" is more informative than "ar", but does not nail down the spoken language. "zh-Hans-CN" leaves the spoken language even more open.

Developers: No unresolved ISO 639-3 macrolanguages.

Be specific about spoken forms of Arabic and Chinese.

UX designers: Make spoken language selection intuitive.

Speech Recognition and You

Android example

```
Intent intent = new Intent(
   RecognizerIntent.ACTION_RECOGNIZE_SPEECH);

intent.putExtra(RecognizerIntent.EXTRA_LANGUAGE_MODEL,
            RecognizerIntent.LANGUAGE_MODEL_FREE_FORM);

intent.putExtra(RecognizerIntent.EXTRA_LANGUAGE,
            "cmn-Hans-CN");

startActivityForResult(intent, myRequestCode);
```

Chrome HTML5 example

```
<input type="text" x-webkit-speech/>
```

Speech Reco for Your Language

Ingredients when starting from absolute zero:

- Large set of textual prompts in the target language.
- Recordings of read prompts, from at least 100 different speakers, ideally 500.
- Pronunciation lexicon, with good coverage for prompt set.
- Description of how numbers and other special tokens are pronounced.
- Large monolingual text corpus.

Implications for Preservation

Speech Technology arguably not of topmost concern in the long tail of languages.

In the meaty middle, need dozens (ideally hundreds) of hours of speech with transcripts. Using read speech plus quality control can save transcription effort. Substantial data collection effort, only useful if completed at scale and quality.

Easy targets: Under-represented languages of the developed world, e.g. Welsh, Irish, Romansh, the languages of Italy and Spain, etc.

Conclusions

Spoken input, especially on mobile devices, is here to stay. Already "good enough" in many scenarios, with many improvements yet to come.

Speech internationalization is a major effort and requires as much or more planning than typical localization efforts.

Thank You!

Questions?