The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text

Christopher Cieri, David Miller, Kevin Walker
{ccieri,damiller,walkerk}@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium and Department of Linguistics
3600 Market Street, Philadelphia, PA 19104 U.S.A.

www.ldc.upenn.edu
Background

• Corpus users and authors increasingly interested in:
  – greater volumes of data in more languages
  – with more sophisticated annotation
  – for use in an expanding number of disciplines
  – requiring standards, tools and best practices

• LDC addressing needs by
  – specific projects in data collection, annotation and publications
  – incorporating annotation, research and tool development

• Need to increase the quantity, quality and diversity of language resources
  – more intensive collaboration between researchers and data providers
  – yielding more data creators, researchers with better appreciation for data creation and data creators with better appreciation of data uses

• Requires more intensive resources planning (roadmaps)

• Need greater cooperation among international data centers which is compatible with local mandates.

• LDC open to cooperation with individuals and data centers around the world.
EARS Program

- Effective Affordable, Reusable Speech-to-Text
  - DARPA common task project driven by annual go/no-go criteria
  - to achieve 5 fold increase in speed, accuracy
  - generate readable transcripts adapted for downstream processing

- Case study in resource planning where demand exceeds supply
  - exploited existing resources: Switchboard, TDT, new TIDES collections
  - required difficult decisions RE
    - priority of different research areas, languages (effort for English > Arabic > Chinese) and volumes of data for training and testing
  - raw data collection required to supply STT & MDE, training and test corpora
  - focus on simple annotations that humans perform consistently in high volume

- LDC provides
  - broadcast news, conversational telephone speech, meetings
  - time aligned transcripts, annotation for metadata extraction (MDE)
  - training, development test and evaluation data
  - English, Mandarin and Arabic
English CTS Goals

• Just one of many EARS data goals

• Volume
  – 2000 hours
  – each subject makes 1-3 calls
  – maximum call length is 10 minutes

• Assigned topics
  – 40 original
  – 60 implemented in November

• Demographic Goals – balanced within 10% absolute
  – Sex: m/f
  – Age: 16-29, 30-49, 50+
  – Region: North, Midland, South, West, Canada, Other (?)
  – also monitor handset, education, occupation in collection

• High Quality, Time-Aligned Transcripts for all speech
Human Subjects

• All LDC telephone studies
  – follow US regulations on treatment of human subjects
  – audited annually by an Internal Review Board (IRB)
  – managed by the University of Pennsylvania Office of Regulatory Affairs

• Main issues informed consent & risk vs. benefit
  – all participants informed that calls recorded for research, educational purposes
  – main benefits are societal
    » benefit to subjects is monetary compensation, free call
  – main risk is to anonymity
    » Subjects identified by 5 digit PIN

• New IRB protocol covers all speech collections
  – prompted or conversational
  – human-human or human-machine
  – face-to-face or telephone
Fishboard
Fishboard
Fishboard Performance
Collection

- Collection began 12/15/2002, continued for 1 year
- Platform in operation
  - 7 days per week
  - noon (EST) > midnight (PST)
- Call collection driven by:
  - availability schedules of participants
    » given by day and hour
    » robot operator called at least once in each available block
  - caller activity
    » in Fisher, callers had little motivation to initiate calls
    » Mixer offer incentives for call-ins and volume is much higher
    » platform functioned well in both cases
    » non-participation = de-selection
  - total platform activity (energy)
- Relatively small number of calls per subject increased requirement on recruiting
Recruitment

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers
Recruitment

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers
Recruitment

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers
Recruitment

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers

Make What You Say Count!

Earn some spare cash at the same time!

The Linguistic Data Consortium is recruiting participants for the Fisher telephone speech collection. If you are a fluent speaker of English, and are at least 16 years old, we need your help!

As a FISHER participant you will be asked to take part in 1 to 3 telephone calls talking to other participants for ten minutes. You will be paid $10 for each call that you complete, and you will also be entered in a lottery (three $1000 prizes).

All calls are recorded and are used for ongoing linguistic research. We need lots of new speakers, so sign up now! If you have any questions, please call - our friendly recruiters would love to talk to you.

1 800 380 7366
www.ldc.upenn.edu/Fisher
Recruitment

- referrals
- print media
- web ads
- groups
- radio
- posters, flyers
- 16,454 calls, 2742 total hours audio
• Gender balance
• 53% female
• 47% male

• Distribution by Age Group
  – 16-29  38%
  – 30-49  45%
  – 50+   17%
Yields

Distribution by Region

- North 24%
- Midland 26%
- South 19%
- West 17%
- Canada 1%
- Non-USA 3%
- Non-Native 10%
Audit

• All calls receive quick human audit
  – 160 seconds, 4 segments
  – Grade: A, C, F

• Auditors check for:
  – Language: Is it English? Is it understandable?
  – Speaker: Does speaker seem to belong to age, gender registered?
  – Channel: Do noise, echo, distortion levels interfere with comprehension
  – Call Content: Is discussion directed speech on assigned topic?
Quick Transcription

- Provides order of magnitude more training data by focusing on speed of transcription

- Specification
  - complete, verbatim
  - without punctuation, special symbols, talker/background noise
  - with limited interjections, non-lexemes
  - (( )) for unclear speech, – for truncated speech
  - annotators may insert other special symbols, punctuation if natural

- Rates
  - Segmentation: 3xRT > 0xRT (automatic or forced aligned)
  - Transcription: 5xRT
  - Post Processing 1xRT: QC on spelling, format, numbers

- Challenges:
  - spelled acronyms, numbers, spacing, proper names, disfluencies

- Compared favorably with carefully transcribed training data
  - all new EARS English and Arabic training data is QTr style
  - most English produced by WordWave under contract to BBNT.
  - LDC provides some English QTr and all Levantine Arabic
Conclusions

• Fisher 2003 used in EARS; released in 2004-2005 (?)

• Fisher 2004 underway
  – similar model
  – >1000 hours new collection
  – subjects allowed to make up to 20 calls

• Collection protocol used in MMSR
  – Multilingual, Multi-channel Speaker Recognition
  – Subjects complete 10+ six-minute calls on assigned topics
  – 400+ bilingual subjects speak in Arabic, Mandarin, Russian, Spanish
  – 200 subjects recorded on 9 different channels, sensors
  – 550 subjects completed 20+ calls
  – See the poster today at 5:00 in session 9-SE in the Laman room
A: and it's not popular at all and i think

A: it just because she's not likeable and people don't like her

B: right yeah that makes sense

A: uh-huh

B: yeah i think also like with reality tv people can imagine themselves more easily

B: being

B: a part of that

B: you know like