OVERVIEW

• How to move speech technology from research labs to the market?
• What are the current challenges is speech recognition research?

• Phonexia introduction
• Technology deployment use cases
• Technologies and what is behind
• Speech core and application interfaces
• Grand challenges
WHAT IS IN SPEECH?

**Speaker**
- Gender, age
- Speaker identity
- Emotion, speaker origin
- Education, relation
- When speaker speaks

**Content**
- Language, dialect
- Keywords, phrases
- Speech transcription
- Topic
- Data mining

**Environment**
- Where speakers speaks
- To whom speakers speaks (dialog, reading, public talk)
- Other sounds (music, vehicles, animals…)

**Equipment**
- Device (phone/mike/…)
- Transmit channels (landline/cell phone/Skype)
- Codecs (gsm/mp3/…)
- Speech quality
Goal:
help clients to extract automatically maximum of valuable information from spoken speech.

• Based in 2006 as spin-off of Brno University of Technology
• Seat and main office in Brno, Czech Republic, active worldwide
• Customers in more than 20 countries governmental agencies, call centers, banks, telco operators, broadcast service companies …
• Profitable, no external funding
FROM RESEARCH TO MARKET

Research

- Scientific papers, reports, experimental code (Matlab, Python, C++, lots of glue (shell scripts), data files
- The goal is **accuracy**
- Stability, speed, reproducibility and documentation less important
- Openness

Technologies

- The goal is **stability** (error handling, code verification, testing cycles at various levels) and **speed**
- Regular development cycles and planning
- Well defined application interfaces (API)
- Documentation, licensing

Products

- Development of new applications
- Integration with client’s technologies and systems
- The goal is **functionality** o integrated solution
- User interfaces
USE CASES

- Call centers
- Banks
- Intelligence agencies
CALL CENTERS

Two main application areas:

1) Quality control
2) Data mining from voice traffic
Supervisor is responsible for:

- Team leading
- Rating of calls
- Evaluation of operators
- Analysis of results
- Reporting

⇒ Only 3% of calls are analyzed by listening
⇒ 100% of calls are analyzed using speech technologies, new statistics
⇒ lower staff costs, lower operating costs
⇒ Higher satisfaction of customers
Technologies:

• **VAD + discourse analysis**
  To get important statistics about call progress (start time, speaker turns, speech speed, reaction times …)

• **Diarization**
  Separation of summed conversation to two channels

• **Keyword/phase detector**
  Detection of obligatory phrases, rough words, call script compliance …

• **Speech transcription + search**
  To search for important places in calls
DATA MINING FROM INCOMING CALLS

Use cases:

• Prevention of call center from overloading (for example large power outage)
• Added value information for business (big data)

Technology:

• Speech transcription
• Data mining tool
• Search engine
Use cases:

• Banks have call centers
  – Quality control
  – Data mining from incoming traffic

• Authentication of people using voice biometry
  – Using key phrase (text dependent speaker identification)
  – Authentication on background (text independent speaker identification)

• Identification of frauds
  – People with fake identities calls repeatedly to request loans
INTELLIGENCE AGENCIES

- Huge amount of information, can not be processed manually
  - public news, telecommunication networks, air communication, internet ...
- Search for a needle in haystack
- Combination of all technologies
  - language identification, gender identification, speaker identification, diarization, keyword spotting, speech transcription
  - data mining tools
  - correlation with other metadata
- Operational and forensic speaker identification
TECHNOLOGIES

- Voice activity detection
- Language identification
- Gender recognition
- Speaker identification
- Diarization
- Keyword spotting
- Speech transcription
- Dialog analysis
- Emotion recognition
• Energy based VAD – fast removal of low energy parts
• Technical signal removal and noise filtering - removal of tones, removal of flat spectra signal, removal of stationary signals, filtering of pulse noise
• VAD based on f0 tracking – removal of other non-speech signals
• neural network VAD – very accurate VAD based on phoneme recognition
VAD CHALLENGES

• Important area of research, not fully solved, VAD is a key part of other technologies and directly affects accuracy of these technologies
• Music/singing detector
• Detectors of non-speech speaker sounds (cough, laugh)
• Detectors of other environment sounds (transport vehicles, animals, electric tools, door slam)
• Technical signal detectors
• VAD for high noisy speech (SNR lower than 0 dB)
• VADs or distorted channels
• Non-parametric VADs
• Distant mike VADs
LANGUAGE IDENTIFICATION

- Automatic recognition of the language spoken.
- 50 languages + user can add new ones themselves
- Can be used also as dialect recognition
- iVector based technology, discriminative training, < 1kB language prints
- Acoustic channel independent

Usage:
- Crime is caused by small groups speaking specific languages very often
- Call record forwarding (to operator / other technologies / archive ...)
- Analysis of the audio archive
- Insertion of advertisement to media
- Language verification in broadcast signal distribution
LID SYSTEM ARCHITECTURE (IVECTOR BASED SYSTEM)

1. Feature extraction
2. Collection of UBM statistics
3. Projection to iVectors
4. Language classifier - MLR
5. Score calibration / transform

Prepared by Phonexia

Fully trainable by client

Language prints (iVectors) can be easily transferred over low capacity links
SPEAKER RECOGNITION

- Several scenarios: speaker verification, speaker search, speaker spotting, link/pattern analysis
- Text independent or text dependent mode
- iVectors based technology, < 1kB voiceprints
- Voiceprint extraction and scoring
- Millions of comparisons in fraction of seconds
- Diarization (speaker segmentation)
- User-based system training, user-based calibration
SYSTEM ARCHITECTURE
- VOICE PRINT EXTRACTION

- feature extraction
- collection of UBM statistics
- projection to iVectors
- extraction of spk info. - LDA
- user normalization

UBM
Projection parameters
Projection parameters
Norm. parameters

prepared by Phonexia
trainable by user

• iVector describes total variability inside speech record
• LDA removes non-speaker variability
• User normalization helps user to normalize to unseen channels (mean subtraction)
Voiceprint comparator returns log likelihood

Calibration ensures probabilistic interpretation of the score under different speech lengths

Score transform enables to selects log likelihood ratio or percentage score
LID AND SID CHALLENGES

- LID/SID on very short records (< 3s) while keeping training at user side
- How to ensure accuracy over large number of acoustic channels and languages (SID)
- Graphical tools for system training/calibration and evaluation at user side
- LID/SID on Voice over IP networks
- LID/SID form distant mikes
• Fully Bayesian approach with eigenvoice priors (Valente, Kenny)
• Initial number of speakers is higher than expected number of speakers
• A duration model is used to prevent fast jumps among speakers
• Target number of speaker can be chosen based on minimal speaker posterior probability, or pre-set by user
• Very accurate but slower and more memory consuming
DIARIZATION CHALLENGES

• Diarization is a technology that still needs a lot of research
• Very sensitive to initialization
• Very sensitive to non-speech sounds (laugh, cough, environment sounds), integration of a good VAD is necessary
• Very sensitive to changes in transmit channels and language
• Even with DER close to 1% there are recordings where current algorithms fail completely (often two women speaking with high pitch)
• Uses iterative approach, one iteration is equivalent to one run of SID, users expect much faster run than SID
• Diarization from distant mikes
• Beam-forming and diarization from microphone arrays
• How to accurately estimate number of speakers
Two approaches:

**KWS based on LVCSR**
- Very accurate
- Slower
- Expensive for development

**Acoustic KWS**
- Fast
- Less accurate
- Cheap development

- Speech transcription based on state-of-the-art acoustic and language models
- Posteriors from confusion network used as confidences
- About 100h of training data
- Simple neural network based acoustic model
- Simple language model (phone loop as background)
- About 20h of training data
- Phoneme-based calibration
SPEECH TRANSCRIPTION

- PLP + bottle-neck features, HLDA
- fast VTLN estimated using a set of GMMs
- GMM or NN based system
- Discriminative training
- Speaker adaptation
- 3-gram language model
- strings / lattices / confusion networks
Rather engineering then research challenges:

- Accuracy
- Speed
- Lower memory consumption
- How to train new system fully automatically
- How to run hundreds of recognizers in parallel
- How to do channel normalization and speaker adaptation for any length of speech utterance
It is much easier to sell speech transcription with a higher-level data mining tool

- There is too much text to read
- The text has too many errors (users will never be happy unless the text is 100% correct)

This can be overcome by integration with existing text-based data-mining tools:

- Categorization of recordings
- Indexing and search using complex queries
- Exploration of new topics
- Content analysis
- Reaction on trends

Integration is done on confusion networks (alternative hypothesis, increased probability to find specific information)
TOVEK TOOLS – CONTENT ANALYSIS
HOW TO MOVE OUR SOFTWARE TO USERS?

• Decision to write new speech core in 2007 – Brno Speech Core
• Focus on stability, speed and proper error handling
• Object oriented design, proper interfaces, no dependency among modules except through regular interfaces
• One C++ compiler, binary compatibility of libraries with others
• One code base for all technologies
BRNO SPEECH CORE

- More than 250 objects covering large range of speech algorithms (feature extraction, acoustic models, decoders, transforms, grammar compilers, …)
- More than million of source code lines
- Code versioning, automatic builds, test suits, licensing
- Still easy maintainable
  - extension of functionality inside objects
  - splitting of functionality to more objects
  - replacement of objects (fixed interfaces)
FAST PROTOTYPING

- Research to product transfer time is essential for commercial success
- Research done using standard toolkits – STK, TNET, KALDI, Python scripts, …

For production systems:
- New system can be implemented in few days
- Often no single line of C/C++ code is written
- Only objects for new algorithms are implemented
- Objects are connected through one configuration file to form data streams
[source:SFileWaveformSourceI]
...

[fconvertor:SWaveformFormatConvertorI]
input_format_str=lin16
output_format_str=float
nchannels=1
...

[melbanks:SMelBanksI]
sample_freq=8000
vector_size=200
preem_coef=0.97
nbanks=15
...

[posteriors:SNNNetPosteriorEstimatorI]
...

[decoder:SPhnDecoderI]
...

[output:STranscriptionNodeI]
...

[links]
source->fconvertor
fconvertor->melbanks
melbanks->posteriors
posteriors->decoder
decoder->output
APPLICATION INTERFACES

- Customers are used to work with specific programming tools and do not want to change their habits
- GUI/command line/SDKs
- C/C++ API – binary compatibility with many compilers
- Java API – a middle layer created using Java Native Interface (JNI)
- C# API – automatically generated using SWIG
- MRCPv2 network interface – for integration to telephone infrastructure (IVRs) - through UniMRCP open source project
- REST – server platform, simple network interface for each technology
- Supported OS – Windows/Linux, 32/64 bits, Android
USUAL DESIGN OF SYSTEMS WITH WEB BASED GUI

- Speech Server (REST), application server, web server, database
- Speech technology inside application server (TomCat, JBoss) through Java API
CLIENT FOR REST SERVER

Voice Analysis Tool

File name | File size | Channel | Date | Speaker score | Likelihood ratio | Gender | Speech length | Record length | Language | Language score
---|---|---|---|---|---|---|---|---|---|---
3003137.wav | 134.71 MB | A | 13.11.2013 10:07:45 | 0.00 | -14.10 | M (66.83) | 0:36:52 | 0:53:23 | Czech | 48.47
3003856.wav | 49.42 MB | A | 13.11.2013 10:07:45 | 0.00 | -11.87 | M (69.70) | 0:14:38 | 0:19:34 | Czech | 46.48
3004614.wav | 54.33 MB | A | 13.11.2013 10:07:46 | 0.00 | -13.72 | M (65.71) | 0:16:25 | 0:21:31 | Czech | 47.60

Speaker: Speaker2, Language pack: Default

vybudovat

premiéra

velmi

sahnout

připust

00:00:6, 00:00:9, 00:00:12, 00:00:15, 00:00:18, 00:00:21, 00:00:24, 00:00:27, 00:00:30, 00:00:33

00:03.801, 00:01.300 / 00:06.006
GRAND CHALLENGES

• Training data collection
• Guarantee of accuracy
• Reduction of hardware cost
TRAINING DATA COLLECTION

• We would like to offer cheap speech technology to anyone on this planet.
• Hundreds of languages and thousands of dialects
• The data is costly - about 30 000 EUR for existing language, more than 100 000 EUR for collection and annotation of new corpora
• The existing data often do not match the target dialect and acoustic channels

⇒ We can add only few languages to our offer per year

Can we find a smarter and cheaper way to collect data?
DATA COLLECTION PROJECT

1. Collection of data from public sources (broadcast)
2. Automatic detection of phone calls within broadcast (high variability in speakers, dialects and speaking style)
3. Language identification to verify language, speaker identification to ensure speaker variability
4. Annotation through crowd sourcing platform
5. Fully automatic process including training of ASR, unsupervised adaptation
   - Experience from data collection for language identification (LDC and NIST adopted this process, now mainstream in Language ID)
   - SpokenData.com ready for online annotation (ReplayWell company)
   - Cost of collection of 100h of data, annotation and system training could be reduced bellow 15 000 EUR per language
   - Interested? Please write to info@phonexia.com
GUARANTEE OF ACCURACY

• More and more customers buy speech solutions. But each new installation brings new risks.
• Speaker identification is not fully language independent
• Language identification is not dialect independent
• Speech transcription is not domain independent
• All technologies are not channel independent

How to estimate in advance if the installation is going to be successful? What will be the target accuracy of each technology?

⇒ Data collection project can help
⇒ Can be extended to World Map of Spoken Languages
HARDWARE COST

• Speech solution is not only software, but also computational hardware, data storage, physical planning of the HW etc.
• Computers and cooling consume electricity
• The additional cost can be about 50% of total project cost
• Most research is directed to reach maximal accuracies
• Any improvement in speed can have large effect on the success of your technology

⇒ Perfect optimization of software
⇒ Use of HW acceleration (GPU cards etc.)
THANKS!

Phonexia s.r.o.
info@phonexia.com