From research to products

3–5–2017

PHONEXIA

Helps clients to extract automatically maximum of valuable information from spoken speech. Turns speech to knowledge.

Based in 2006 as spin-off of Brno University of Technology

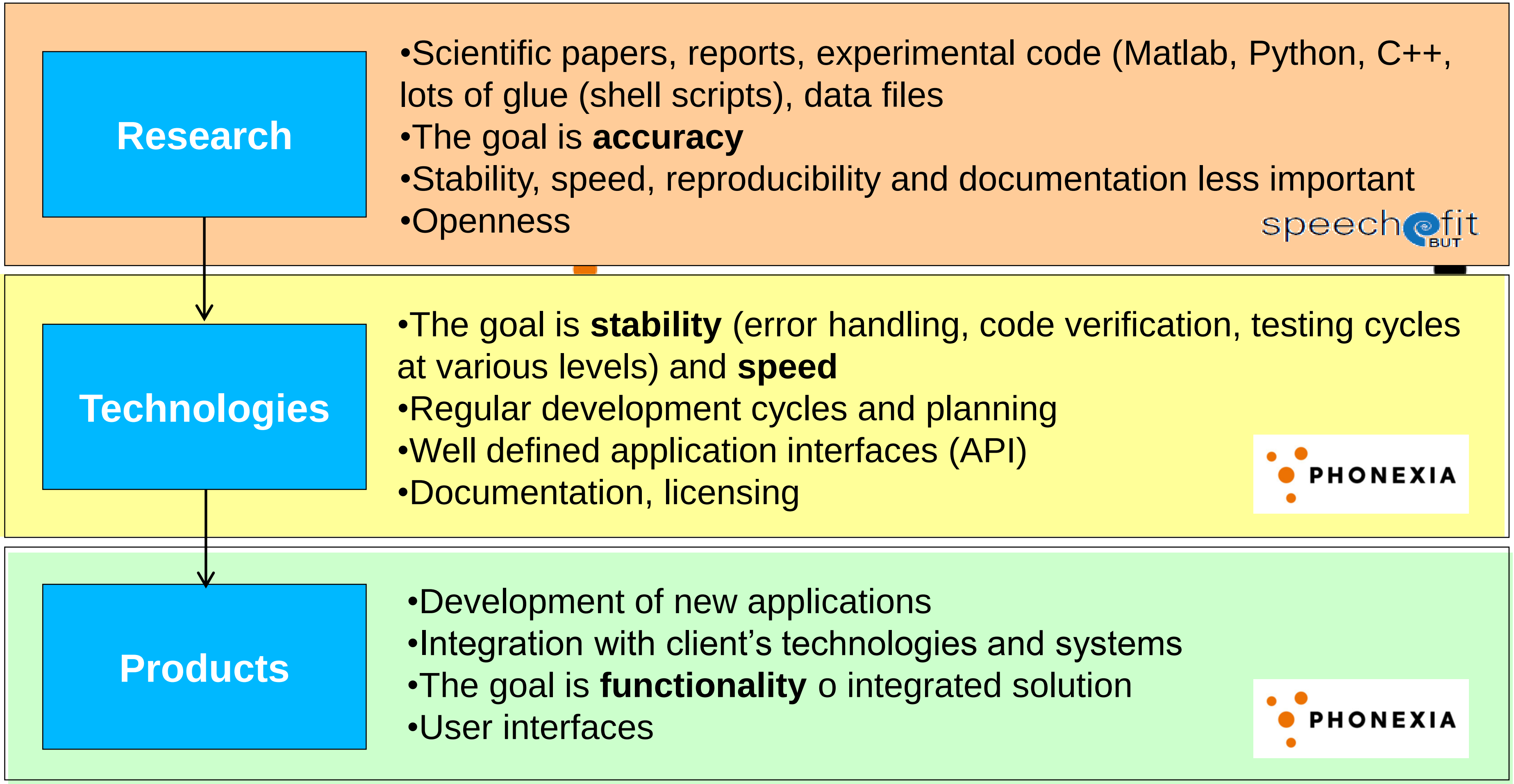
Seat and main office in Brno

Customers worldwide - governmental agencies, call centers, banks, telco operators, broadcast service companies

The main focus on platform and developer tools

Profitable, no external funding

From research to products



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

speech@fit BUT

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

PHONEXIA

Products

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

PHONEXIA

What is in speech?

Speaker

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

Environment

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

Content

Language, dialect
Keywords, phrases
Speech transcription
Topic
Data mining

Equipment

Device (phone/mike/...)
Transmit channels
(landline/cell phone/Skype)
Codecs (gsm/mp3/…)
Speech quality

Voice interfaces – potential or thread?

Technologies

Voice activity detection

Language identification

Gender recognition

Speaker identification

Diarization

Keyword spotting

Speech transcription

Dialog analysis

Emotion recognition

Sentiment analysis

Use cases

Call centers - quality control / business inteligence

News agencies - search for some information

Intelligence agencies - search for some information in a haystack, forensic expertise

Banks - fraud detection / voice as a password

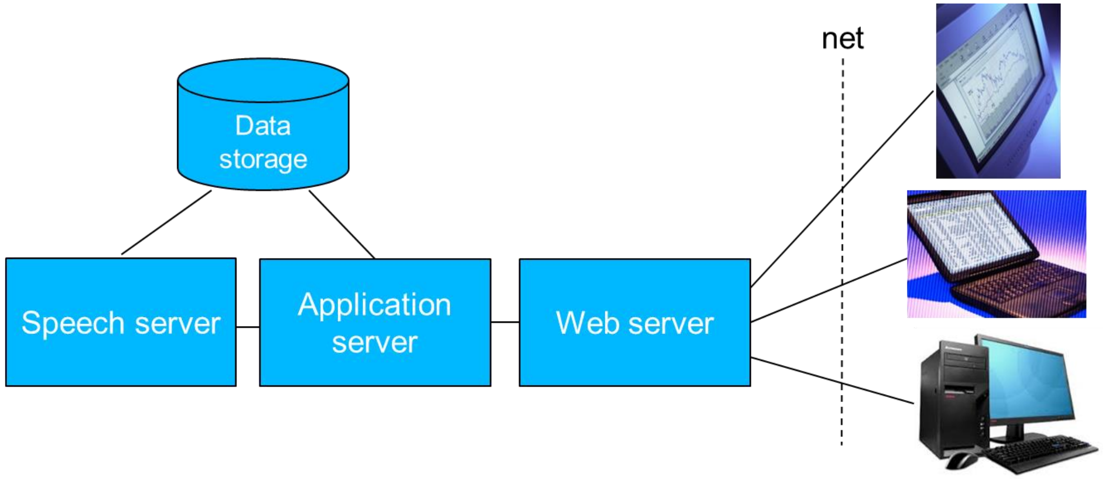
Speech platform

1 **We are delivering tools for developers**

2 One interface for all technologies – REST + HTTP/RTP streams

3 Simple installation, integration and scalability

4 Partnering with developers and integrators all around world

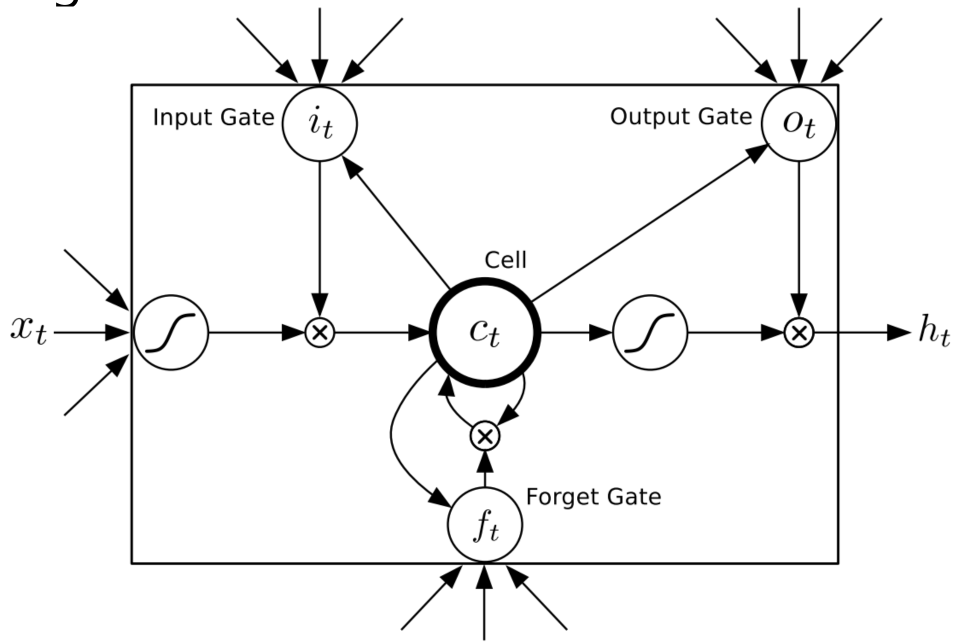


Hot tasks

Neural networks

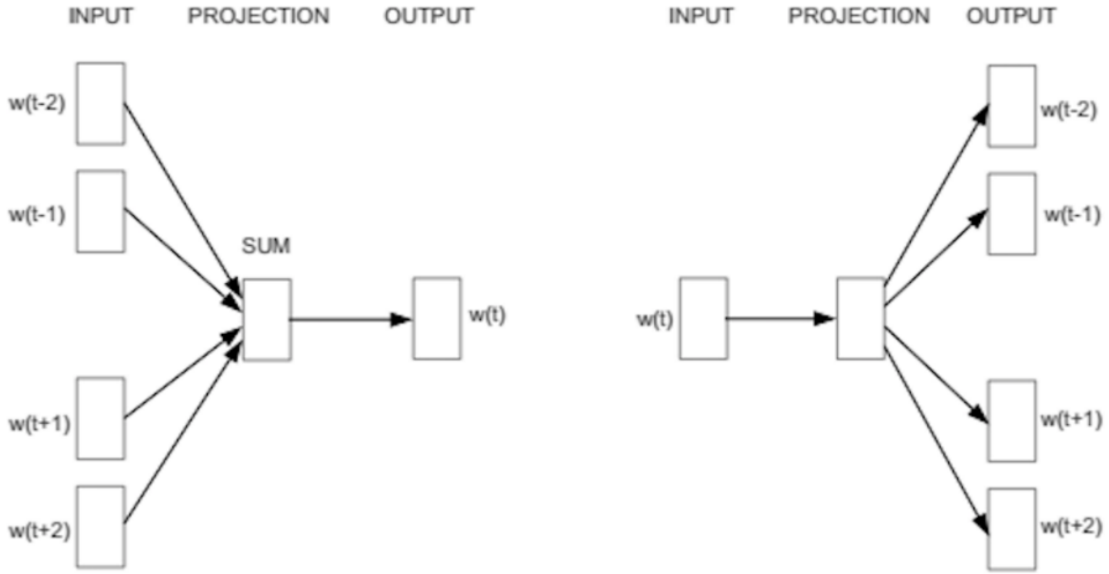
1 **Current systems are deep neural network based**

2 Recurrent neural network are more powerful but gradient vanish too quickly -> move to Long Short Term Memories (LSTM) or BLSTM



Vector representation of words – from words to meaning

1. **Word2vec algorithm proposed by Tomaš Mikolov**
2. Word in some context is mapped to short N-dimensional vector
3. Similar words are mapped to the same place in the N-dimensional space



Punctuation and capitalization

ben is asked to wait for amy but he does not wait he continues to run so amy's request is changed now ben is asked to help amy ben stops and amy is helped

⇓

Ben is asked to wait for Amy, but he does not wait. He continues to run. So Amy's request is changed. Now Ben is asked to help Amy. Ben stops and Amy is helped.

1. A neural network is trained to predict upper/lower case and punctuation. The input is a vector representation of words.
2. Similar technique can be trained to detect POS tags, name entities, other key information etc.

Sentiment analysis

-0.981	very poor service totally unprofessional
-0.971	cost too much money to do
-0.962	i wish it didn't cost so much
-0.940	and we're really really disappointed

0.946	my agent was very very helpful and informative
0.995	excellent service good job thank you
0.996	everyone was extremely happy and very nice thank you
0.996	had great courtesy fast service with a smile

Again, neural network can be trained for this task. The input is a vector representation of words.

Training data preparation and cleaning

1 **Each speech recognizer needs word transcribed speech recordings for training, 100h and more, 1h ~ 15h of human time**

2 Transcription can be done by crowd-sourcing, but good quality control is necessary

3 The goal is to develop speech recognizers in just days

4 A confidence score based on comparison of a sentence model and an arbitrary word sequence model can be used.

Dynamic decoder for lower memory consumption and embedding

1 **A decoder is a key part of each speech recognizer**

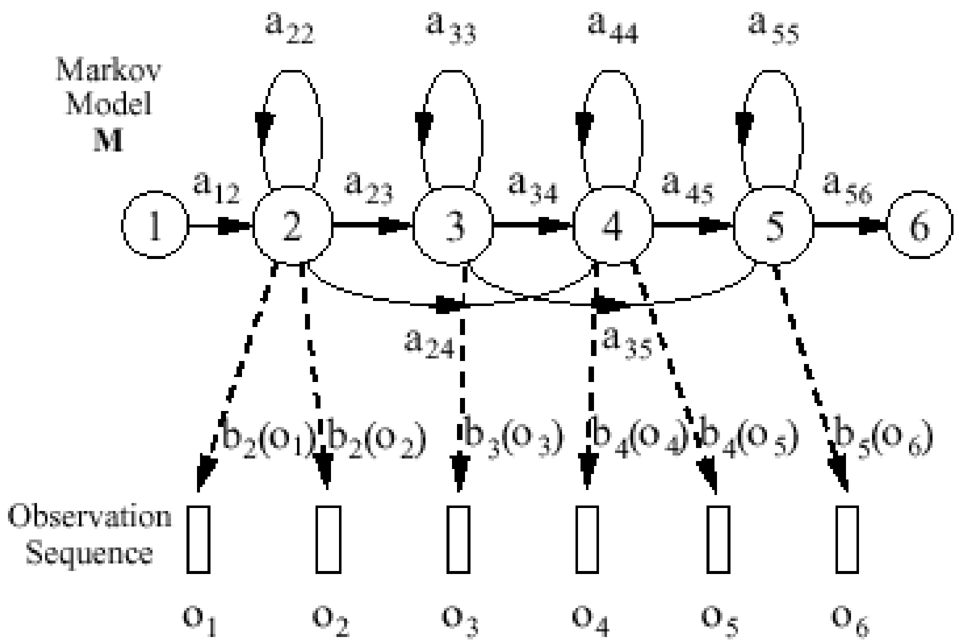
2 Static decoder uses a precompiled recognition network but it is very memory consuming

$$M = H * C * L * G$$

3 Dynamic decoder does the composition on-fly, so it needs only a little of memory. The whole recognizer can fit to some embedded devices.

One state Hidden Markov Models

1) **Classical concept uses more states per phoneme in HMM**
2) Neural networks and long temporal context modelling makes is possible to use just 1 state per phoneme
3) 1 state phoneme models increases the recognition speed 3x

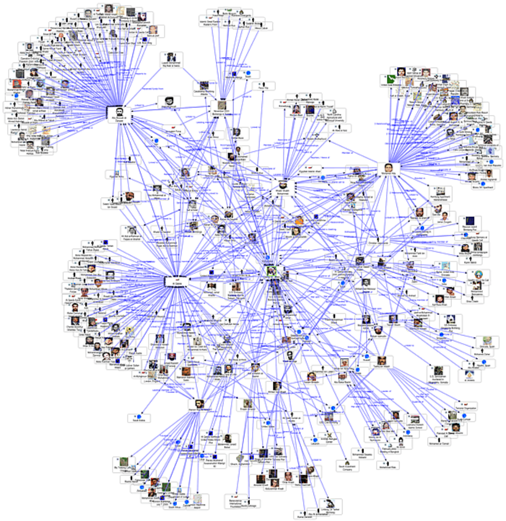


Knowing more about speech source

1) **The recognition result is given by audio quality**
2) Speech can be degraded noise, reverberation, or any loss compression
3) The audio can pass many codecs
4) Detection of audio codecs and bit rates is key to predict the quality of output data
5) It is important for any forensic expertise
6) It is important to prevent spoofing attacks to voice biometry

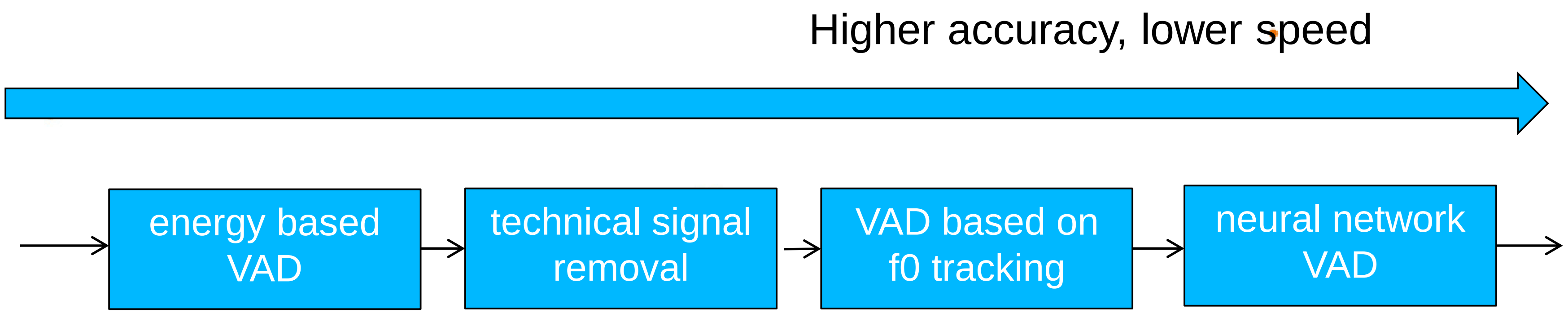
Social network / link analysis

1) There is a lot of metadata in speech recordings
2) Making relations among metadata inside recording or across recordings makes particular pieces of information easier to find
3) The current searching capabilities can be improved by few orders
4) The same happened with text when Google came
5) There are well known algorithms from text



Some technologies

Voice activity detector



Higher accuracy, lower speed

energy based VAD

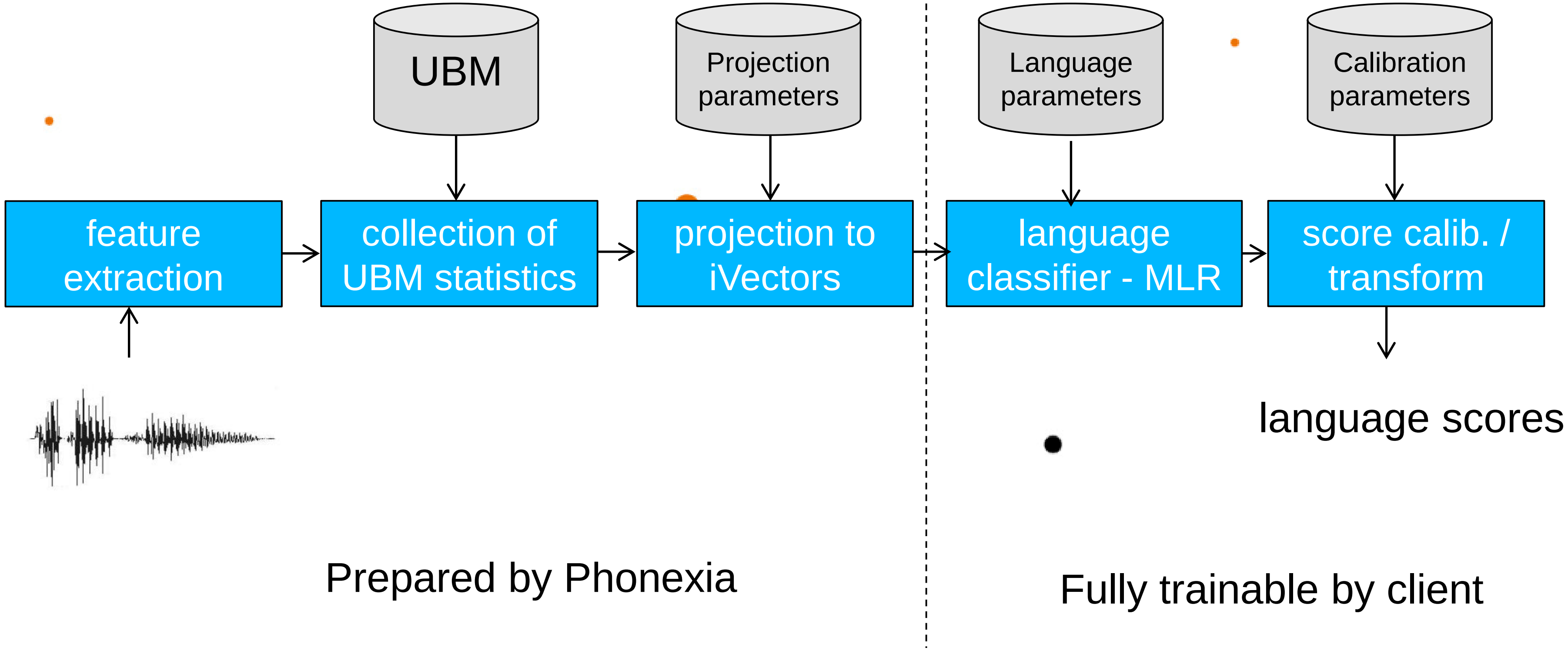
technical signal removal

VAD based on f0 tracking

neural network VAD

- 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

Language identification



UBM

Projection parameters

Language parameters

Calibration parameters

feature extraction

collection of UBM statistics

projection to iVectors

language classifier - MLR

score calib. / transform

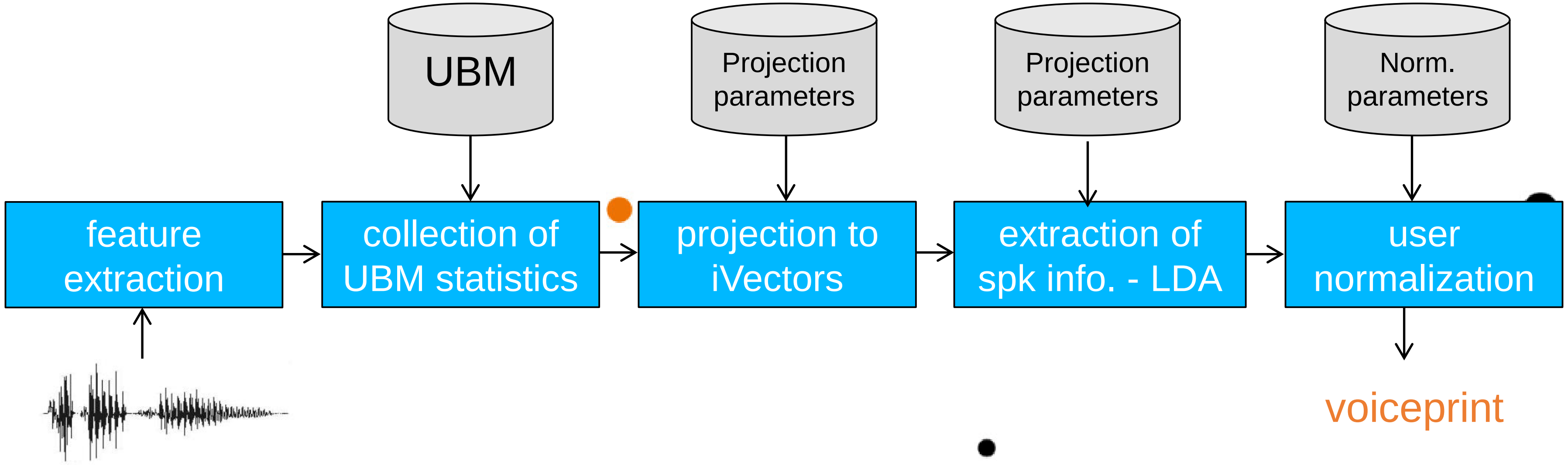
language scores

Prepared by Phonexia

Fully trainable by client

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

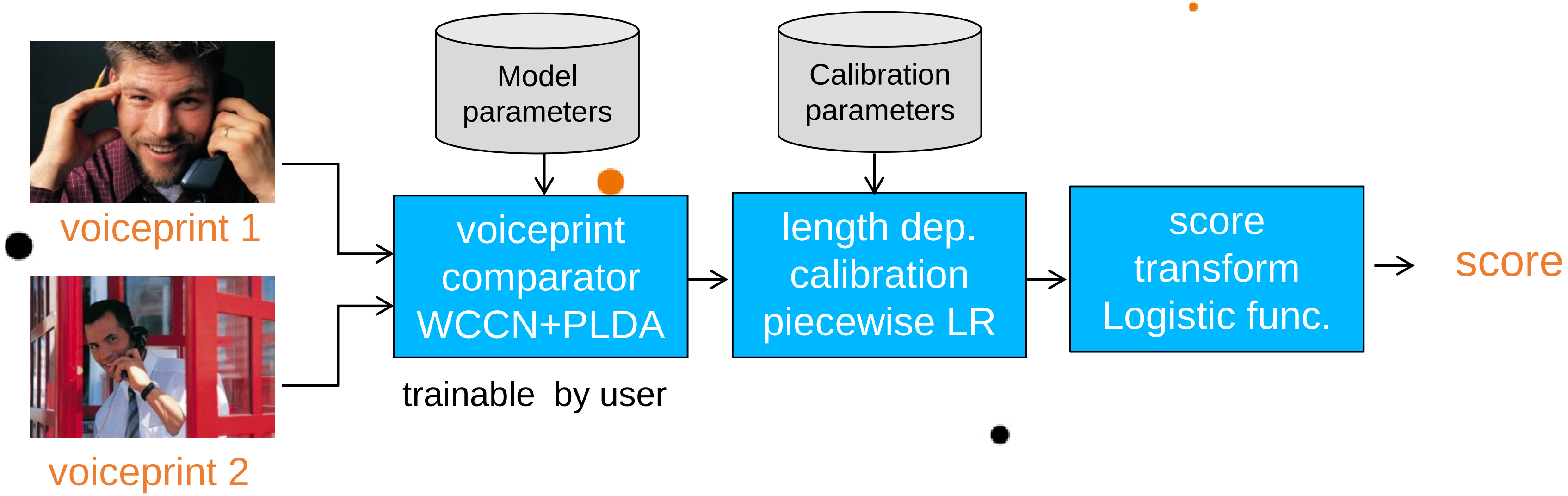
Speaker identification – voice print extraction



prepared by Phonexia

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

Speaker identification – voice print comparison



voiceprint 1

voiceprint 2

Model parameters

Calibration parameters

voiceprint comparator WCCN+PLDA

trainable by user

length dep. calibration piecewise LR

score transform Logistic func.

score

- Voiceprint comparer 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

Thanks for your attention

Petr Schwarz
CTO

T +420 733 532 891
E petr.schwarz@phonexia.com

phonexia.com

PHONEXIA