Automated Conversational Analysis for Predictive Analytics

Eric BOLO
Nicolas SEICHEPINE

Batvoice Technologies

At batvoice we analyze and predict the outcome of sales and service calls to optimize customer relations

Black hole in the customer journey

Online navigation

Mail exchanges

Call

Purchase

Contract cancellation

Cookies, tracking

CRM log & mails

?

What do we do / Use cases

- Entreparticuliers: broker deals between individual property owners and prospective buyers
- Activity split between digital and phone

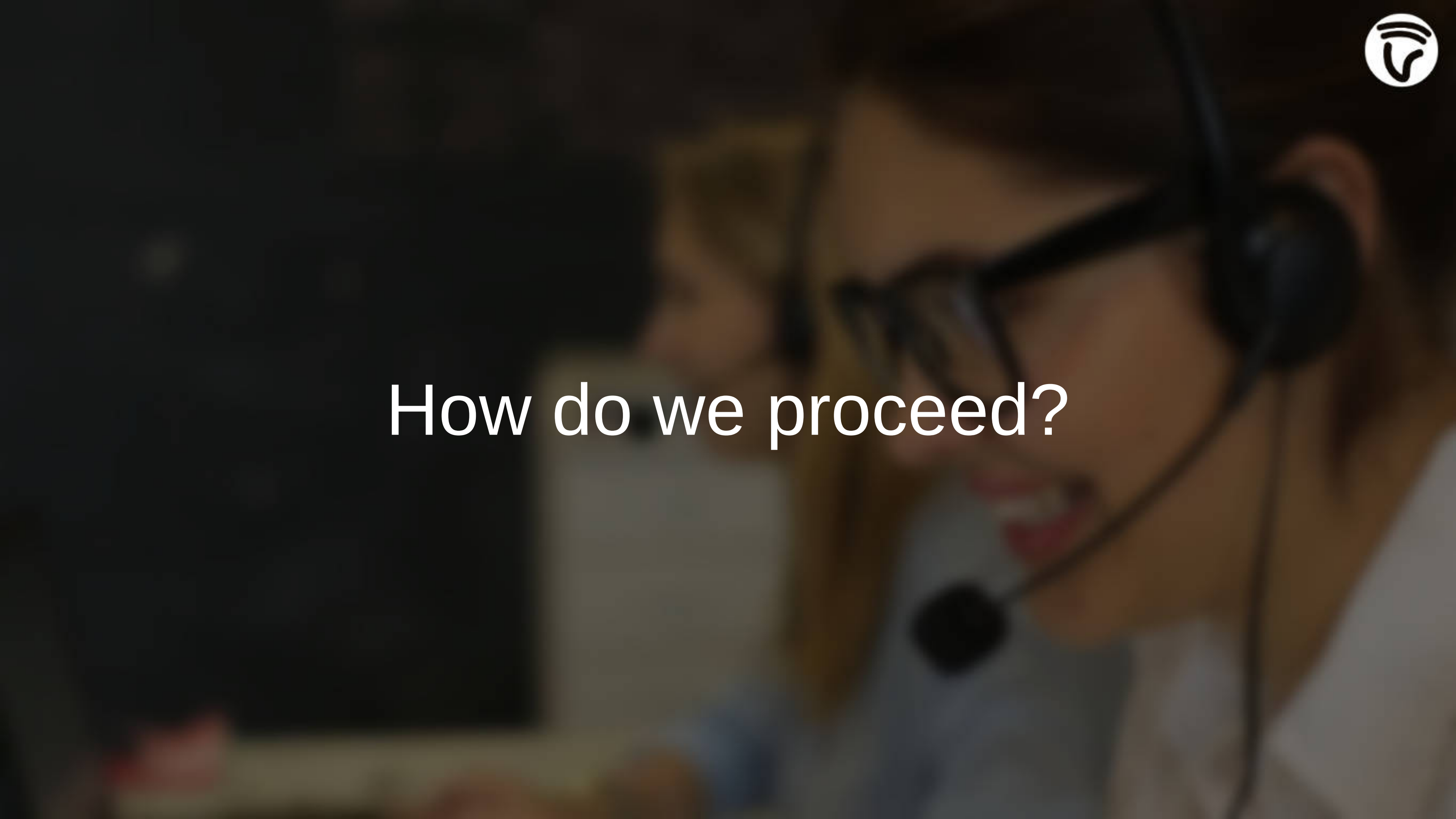
What do we do / Use cases

- Sales call: SMS callback (highly qualified calls), conversion rate ~10%

- Calls usually short (< 5 min), but can last up to an hour

What do we do / Use cases

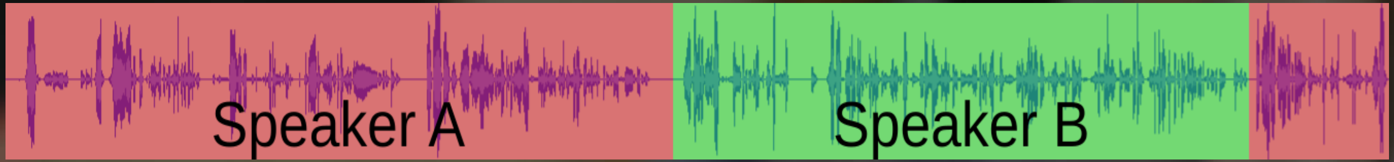
- Challenge: predict the likelihood of conversion @ the beginning of the call, and notify the sales agent
- From then: give up (no dice!), keep going, change strategy



How do we proceed?

How do we proceed / Diarization

- Some measures are local (e.g. tone, text content)
- They have to be *attributed* to one speaker
- Hence the need for *diarization*



How do we proceed / Diarization

- In our case, we *usually* have only two speakers
  - This makes a huge difference!

- Various algorithms exist to perform diarization

- They all amount to building features representative of the voice at a given time…

How do we proceed / Diarization

Signal

↓

MFCC coefficients

↓

MFCC centroids

↓

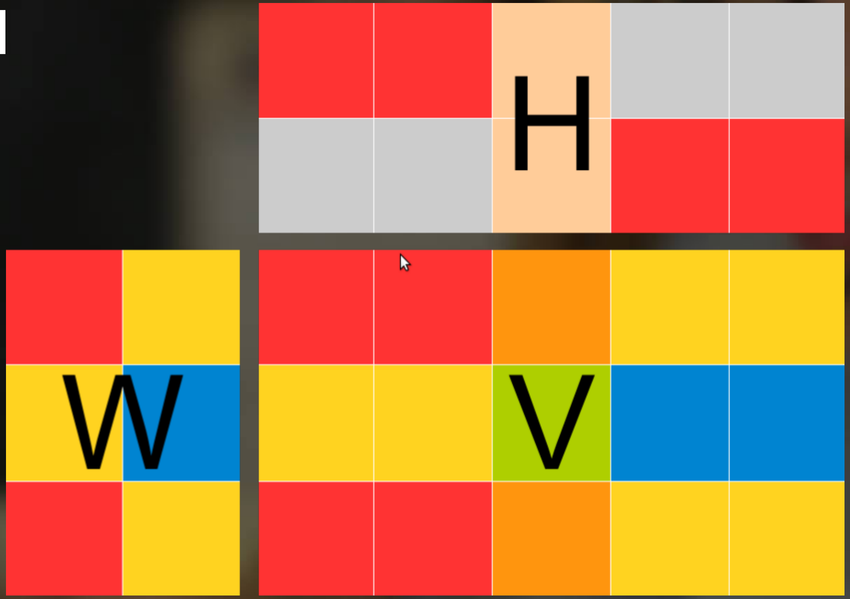
MFCC labels

↓

Vector of labels counts

How do we proceed / Diarization

- … Then group features into homogeneous segments
- Related to a given speaker
- Which is done (here) using nonnegative matrix factorization (NMF): $V \approx WH$

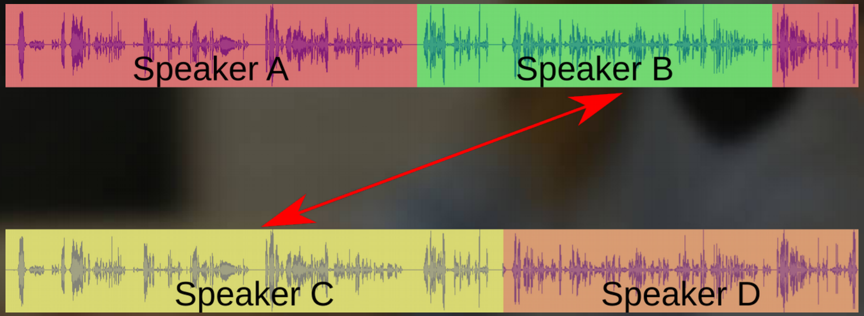


- Separating two speakers is not enough

- Interaction/social parameters need to be associated with the role of a given speaker
  - Client, agent?

- This might be complex (discourse related)

- Cross-identification is possible in real world schemes



How do we proceed / Low level features

- Tone is generally considered to be related with physiological parameters
  - Vocal tract, glottal source
- These parameters directly affect audio production
- Hence are encoded in low level audio features
  - Energy frequency repartition, variations
  - See (e.g.) GEMAPS
- We can train emotion predictors with these features, or keep them for higher level predictions

How do we proceed / Low level features

- Problem: the number of extracted low-level features directly depends on audio length
- This is impractical for classical learning algorithm, which require fixed-length entries
- It’s unrealistic to make “global” predictions from such local features

How do we proceed / Low level features

- Therefore, low-level features are transformed using statistical functionals
  - Percentile, ranges, slopes, …
  - Fixed length
  - Also accounts for tendencies

- Aggregation is always made speaker by speaker!
  - But also at a sentence level

- Interaction can be measured in numerous ways:
  - How long each one speaks?
  - How are speaking turns distributed?
  - How does prosody vary across speaking turns?
  - How much time does each one take to react?
  - Can we measure influence/ascendency?

- Numerical data are also aggregated to avoid dependency on the number of turns

How do we proceed / Social features

- Gender is known to affect interactions outcomes
- Gender affects physiological parameters involved in voice production
- Predictors can be trained, that take audio recording as input
  - Solved problem

How do we proceed / Social features

- Age is known to affect interactions outcomes
- Age affects physiological parameters involved in voice production
  - Vocal tract volume increments
- Predictors can be trained, that take audio recording as input
  - Still in development

How do we proceed / Text features

- Imagine you are listening to a conversation in an unknown language:
  - Interaction gives a lot of clues
  - Content is still useful
- Speech to text algorithms made a lot of progress
  - See EESEN
- Raw text is hard to use directly
  - Hence the need for sentiment analysis
  - And vector representation

How do we proceed / Learning

- Stack all the features…
- Add available meta-parameters…
- And you are left with tabular data with scalar labels
  - Currently more than 700 features
  - More available, should we use them?

- Numerous algorithms & tools are available:
  - Logistic regression, random forests, SVM, neural nets
- Features selection algorithm (e.g. SFFS) partly solve the difficulties related to features number
- As always, you might have to tune some hyperparameters

- Regarding predictions and number of features, the more data the better

- Computational power is limited
  - Feature extraction take some time
  - Learning algorithm also, especially when learning is made iteratively (hyperparameters, feature selection)

- Orders of magnitude of a standard problem:
  - 10x real time (prediction)
  - 10k hours audio (learning)
  - 100k files (learning)

How do we proceed / Infrastructure

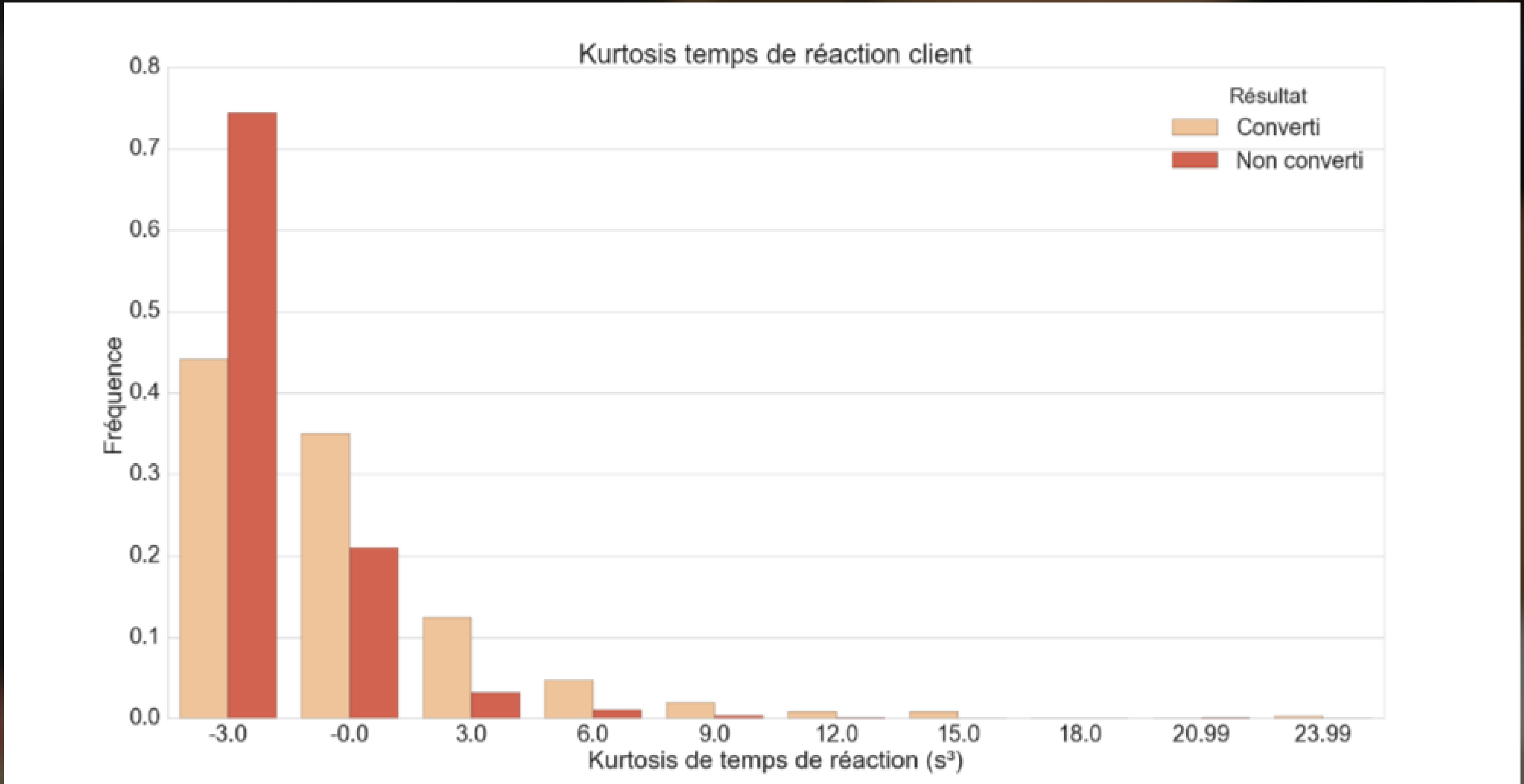
- Need for scalable computing power
- In our case: Amazon cloud
- Automated docker deployment/use
- Spark parallelization

Wave file $\xrightarrow{\text{Upload}}$ Amazon S3 $\xrightarrow{\text{Triggers lambda}}$ Queue service $\xrightarrow{\text{Distributes}}$ Cluster $\xrightarrow{\text{Computes}}$ CSV file $\xrightarrow{\text{Writes}}$ Amazon S3

How do we proceed / Example

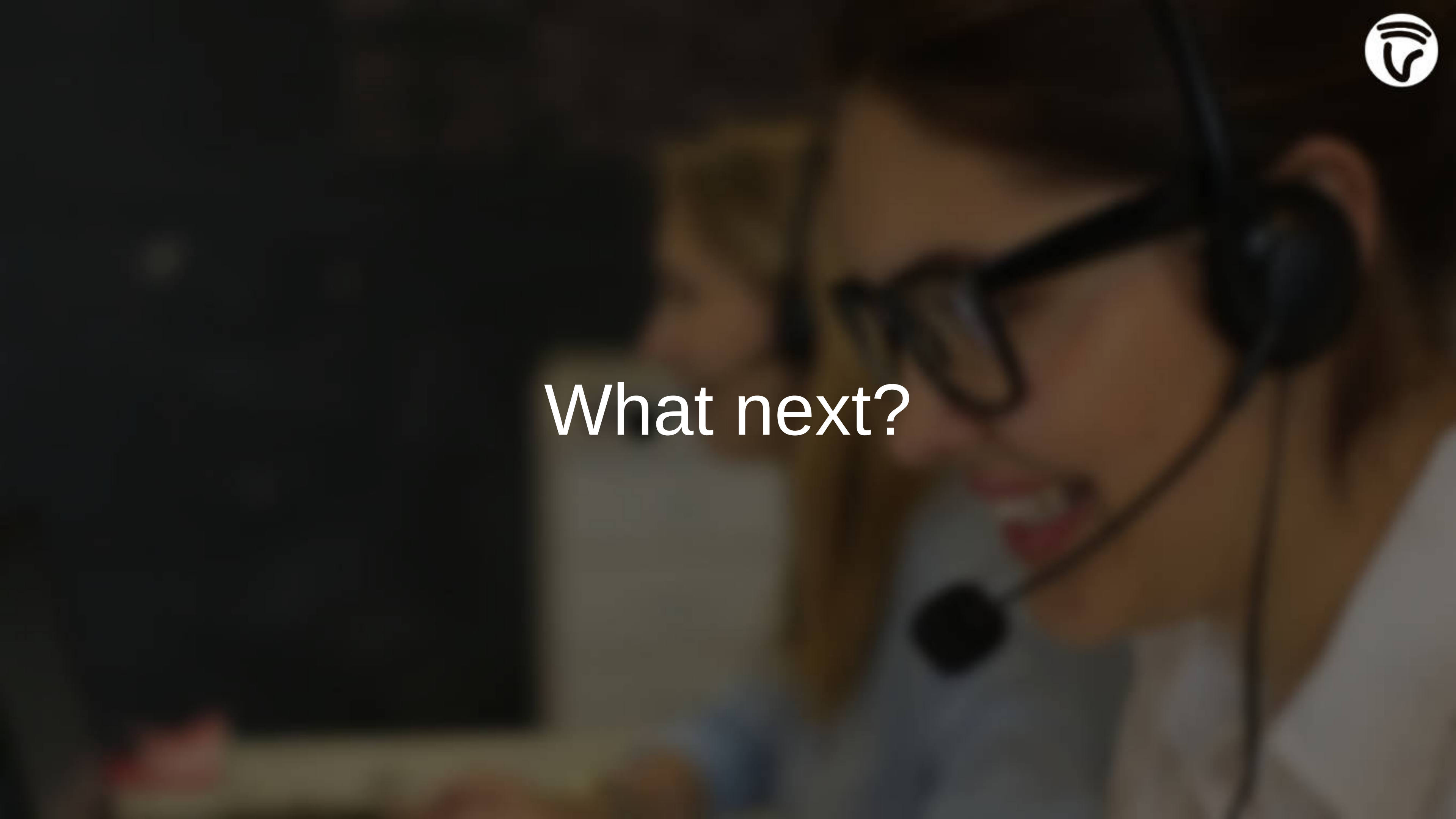
- A specific use case: Entreparticuliers.com
- Binary output
- ~200k files, median ~4 minutes with heavy tail
- Could we shorten calls by rapidly predicting output?

- Get insight on variables
  - And what happens within calls



- Specific results (random forests, no semantic):

Audio processed (s)	60	120	180	240	300
Accuracy (%)	**75**	**75**	72	69	65



What next?

- Adding the semantic dimensions

- Partnership with a specialized speech-to-text (STT) company

- Over the summer: in-house end-to-end STT using the EESEN framework
  - Phonetic model: deep Bi-Directional Recurring Neural Networks + CTC loss
  - Linguistic model: Weighted Finite State Transducers (WFSTs)

- Pros of in-house solution: can adapt the language model (and even the phonetic model given annotated speech) to each new case

- Classical ML models cannot handle sequential data
- Basic idea: split the conversation into consecutive speaking turns
- Two possibilities: convolutional and recurrent networks
- Data extraction:
  - Paralinguistic: turn-level feature summaries
  - Semantic: word embeddings (word2vec, doc2vec)

What next / Other use cases

- Automatic, personalized offers
- Predict other types of outcome, such as churn (high-stake issue for subscription-based services; telecom, magazines, etc.)

- Speaker Diarization: A Review of Recent Research
  - Xavier Anguera, Simon Bozonnet, Nicholas Evans
- EESEN: End-to-End Speech Recognition using Deep RNN Models and WFST-based Decoding
  - Yajie Miao, Mohammad Gowayyed & Florian Metze
- The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing
  - Florian Eyben & al.
- https://aws.amazon.com/blogs/compute/better-together-amazon-ecs-and-aws-lambda/

Conclusion

- Many thanks to:
  - Entreparticuliers.com
  - Smart School

Conclusion

- Don’t be shy!
- Contact: contact@batvoice.com