Speaker Recognition: Enrollment of Users into the Voice Recommendation System

Vamsi Potluru
Rick Ruiz
Gene Chipman

July, 2015

Outline

1. Biometrics

Problems

A Painless way to create/activate TV profiles

Enable personalized features

- Home startup screen or best channel on startup.
- Personalized results (search, recommendations, saved, last 9)
- Favorites and UI customization
- Social media profiles (like for Tweet this)
- Parental controls (may overlap with authentication)

Enable authenticated features

- Enable home security features
- Enable transactions (credit card, PPV purchase)

Approaches — Explicit and Implicit Profile Creation

Implicit profiles

- Automatically adjust results based on user
- This requires a soft touch to not spook users.
- Cannot be used for authentication purposes.

Explicit profiles

- Can be very simple to create.
- Biometrics will enable automatic login to an existing profile.

Why Biometrics?

What Biometrics can do

- Match an utterance to one voice user in a household.
- Learn number of (voice) users in a household.
- Verify that an utterance is from a specified user.

Challenges of Biometrics

- Accuracy dependent on duration of speech sample.
- Up to 1 minute of speech needed to learn users.
- $5 - 10$ seconds of speech needed for recognizing users.

Trained vs Clustering

Trained — supervised learning

- Requires samples where the speaker is known.
- Nuance SDK
  - Fixed phrase ("My voice is my password")
  - Free speech (lower accuracy)

Clustering — unsupervised learning

- Can use samples of unknown speakers.
  - Can determine number of speakers in a household.
  - Can be applied to existing utterance data.
- Lower accuracy — more samples required.
- Incorporate VREX metadata (time of day, action and so on)
  - Improved accuracy.

Implicit Profile Creation

Background clustering on utterance data

- Determine number of users in household.
- Collect enough data for moderate accuracy.

Two possible ways to ID users

- Using just clustering.
- By training Nuance free speech on clustered data for each user (Nuance may be better).

Explicit Profile Creation

- Background clustering on utterance data.
  - Determine number of users in household.
  - Collect enough data for marginal accuracy.
- Prompt for profile creation based on a threshold.
- Profile creation with a short dialogue session.
  - Profile responses come from a known user.
  - Use Nuance SDK with free speech training.
  - Dynamically adjust session length for accuracy.
  - Ask interesting and useful questions during session.
- Automatically log in users on usage or prompt user to login.

Research Steps — Green is completed

- Pull existing VREX data for “known” users.
- Experiment with clustering accuracy — to follow.
- Experiment with Nuance free speech training.
- Experiment with customer service voice call logs.
- Add VREX metadata to clustering.
- Compare clustering versus Nuance for implicit case.
- Build profile creation UX for explicit case.
- Adjust algorithms and thresholds for both cases.
- Report on data requirements and accuracy for both cases.

Datasets for Speaker Clustering

Office dataset

We hand (mouse/ear?)-labeled bunch of utterances from 6 speakers in the office — Herminia, Tom, Robert, Grace and others. The duration of speech ranged from over a minute for Tom (maximum) to around 20 seconds for Grace.

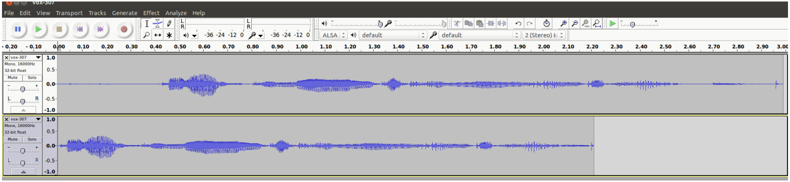
Production dataset

We extract utterances from 8 speakers who use the mobile app on Android to get good labeled data. This is a much bigger dataset with each speaker having at least one minute of data.

Speaker clustering algorithm

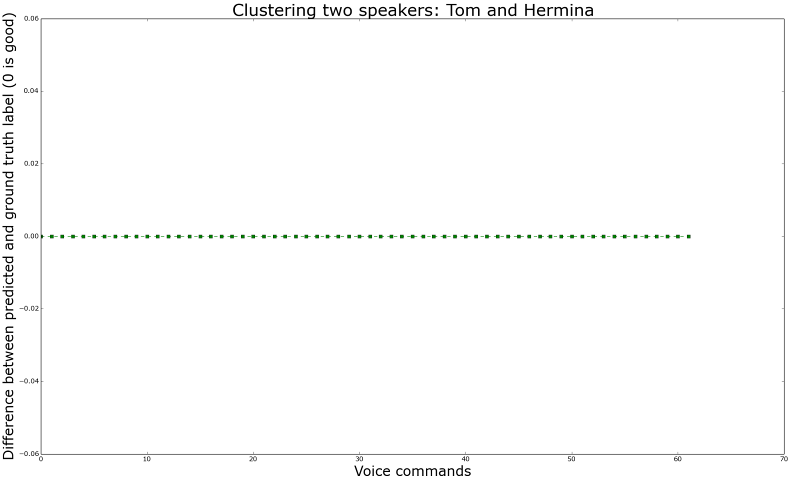
- MFCC — Frames, Spectrograms, Mel filterbank (correlation), logarithm (ear), DCT (uncorrelated)
- Extract MFCC's + energy $(19 + 1)$, delta-MFCC's (20) and delta-delta-MFCC's (20): 60-dimensional vectors.
- Assign random labels to the files.
- Learn GMM's with 5 components.
- Use the learnt GMM's to predict the class labels and go back to last step until convergence.

We can also learn the number of speakers (clusters).



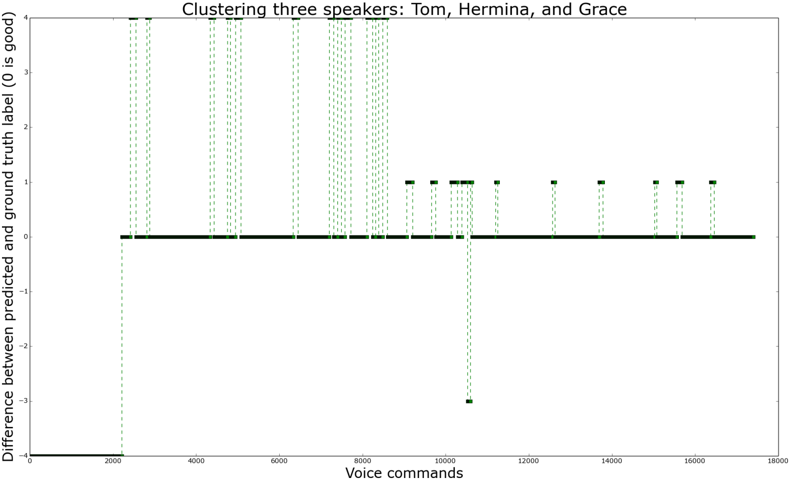
Original file(Grace) and the corresponding silence removed file.

Two speakers: Hermina and Tom



Clustering results using roughly one minute of speech (30 commands) each.

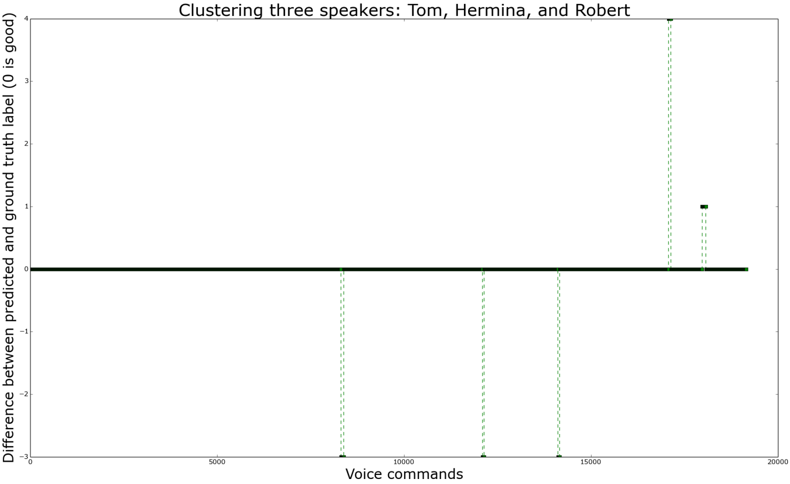
Three speakers: Hermina and Tom and Grace



Clustering results using roughly 20 seconds of speech from Grace. The units are in frames

Three speakers: Hermina and Tom and Robert

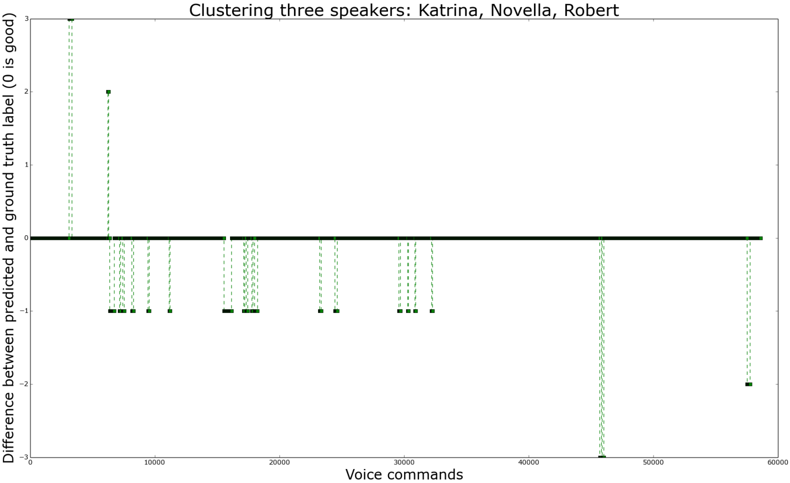
Vamsi Potluru
Rick Ruiz
Gene Chipman



Clustering results using roughly one minute of speech (30 commands) each.

Three speakers: Production system

Vamsi Potluru
Rick Ruiz
Gene Chipman



Clustering results couple of minutes of speech.

Metrics

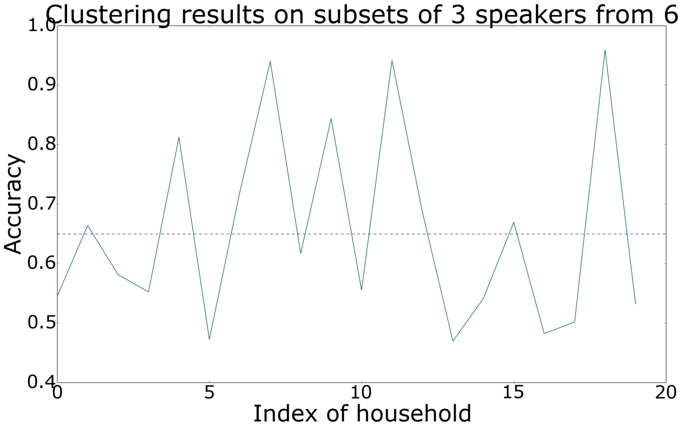
	kat	novella	rob
kat	5724	2632	280
novella	0	28344	237
rob	233	67	21112

Table : confusion matrix (frames/time) for the three speakers: kat, novella and rob.

	kat	novella	rob
kat	22	23	2
novella	0	224	1
rob	1	1	107

Table : confusion matrix (file-level) for the three speakers: kat, novella and rob.

Subsets of Three speakers: Clustering accuracy



We evaluated clustering accuracy on all possible combinations of 3 users out of 6 users.

Questions?