SoundAlike™: A Fast Shortcut to Voice Conversion

Fathy Yassa, Sandeep Mohan, Meir Friedlander and Gerard Chollet

Mobile Voice Conference, San Jose, CA

April 2016
SoundAlike™ Outline of Presentation

- Objective
- Market Motivation
- General algorithm overview
- MFCC Voice features and GMM Modeling
- Clustering algorithm
- Reference Databases
- SoundAlike™ Testing and Classification
- SoundAlike™ Test demos
- Conclusions
SoundAlike™ Objectives

• Select the closest match to a target voice from a library of TTS voices
  – Same Language
  – Across Languages

• A target speaker needs to read random text in any language. The problem here is to choose a TTS with the nearest voice quality to the target out of the available TTS libraries.

• The chosen TTS for the target, can now be used to read random text in a voice sounding like the target.
The Opportunity: Personalized Voice Is The Next Wave

- Corporate customer messaging market is growing rapidly; Corporates look to increase level-of-engagement at all touch points
  - Financial institutions: investment manager voice for statements, personal product proposals
  - Service Providers: Use familiar account manager voice for customer retention, cross services up-sells, new product promotion
  - Health care: personalized messages from doctors, etc.

- Consumers with mobile and connected devices will be overwhelmed by generic synthesized, unfamiliar, undistinguishable voices

The Past
- Engagement Devices: Paper mail, phone, Keyboard/mouse web, radio, TV
- Engagement Options: Traditional messaging, advertisements and publishing
- Communication: Traditional messaging, advertisements and publishing

The Present
- Engagement Devices: Personal mobile touch devices, High Definition visual devices
- Engagement Options: Social media, mobile applications; pushy
- Communication: Personalized interaction with multiple devices using all senses

The Future
- Engagement Devices: Always connected, always engaged: hand-eye and mouth-ear
- Engagement Options
- Communication
SMI Focus Is On the Largest Near-term Use Cases

Enterprise Customer Engagement
- Personal familiar voice (e.g. bank manager)
- New products and services promotions
- Same voice across languages
- Reduces churn, increases loyalty, increases up-sales
- Unlimited scalability: daily communication
- Familiar voice reads emails and text messaging
- Relevant natural voice for personal voice-messages

Consumer Text-to-Personal Voice Apps
- Facebook wall genuine originator’s voice read-to-me
- Text messages, Twitter, etc.
- Mobile x-languages conversational application
- Specific voice emails and text messaging reading
- Automatic natural voice personal voice-messages

Automatic Voice Advertisements
- Mobile / digital radio
- Talking web pages
- Automatically & immediate with any voice
- Fits local & small businesses
- Like vocal Google adware
- Higher level of engagement
- Affordable high quality
- Better conversion rates
- Easy integration with existing ad-networks

Messages & Ads Receiving Use Cases
- While driving
- At the Gym
- Jogging / Biking
- Outdoor / Hiking
- Medicare wearables
- Eyes and hands are busy with other views
SoundAlike™ Technology: How it Works

• Build a library of TTS voices with enough variability in voice qualities

• Cluster the available TTS voices, grouping the voices with similar qualities in each cluster

• For random input voice having no TTS of its own, test the voice sample against the available clusters, and identify the cluster to which it belongs

• Identify the TTS library component in the matching cluster that closely matches the input voice

• Use the nearest TTS to represent the test voice and generate the required speech
Library Features: MFCC

- Windowed Frames
- N-Point FFT
- FFT Magnitude Coeffs.
- Mel-Spaced Filter Bank values
- Log Filter Bank Values
- L-Order Cepstral Analysis

20ms window
10ms Overlap
256 point FFT
24 Mel-Scaled Filters
42 Coefficients
20 Cepstral
1 Energy
20 Delta
1 delta energy
Modeling TTS Database member Voices: UBM-GMM

- Features from MFCC training on UBM-GMM model
- Universal Background Model trained with 1024 mixtures.
- Target model trained by MAP adaptation
- UBM ensures more reliability, robustness
Algorithm Learning Methods - K-means Clustering

- Features: Mean super vectors of every speaker from UBM-GMM.
- Criteria for selection of k:
  - Elbow method selects a range for k
  - Overlap Criteria
- Building class models
  - Combining feature vectors to represent a cluster
Algorithm Learning Methods: Elbow Curve for APPEN

- Curve of mean of within cluster sum of distances to the centroid to number of clusters.

- Point where elbow hits gives best k value.

- Determined by calculating the slope of the curve.
Algorithm Learning Methods: i-vectors with PLDA

- Provides a way to obtain low dimensional fixed length representation of speech
- Factor Analysis: learn a new low dimension space from large data
- Speech projected onto the subspace and coordinates are the i-vectors
- Compute UBM, sufficient stats to derive the total variability space matrix $T$
- Perform LDA for dimension reduction and obtain the i-vectors
- GPLDA to compute parameters for scoring
SoundAlike™ TTS Library Training & Prep.
We used several databases to test the SoundAlike™

- TIMIT (630 voices/10 sentences)
- RSR 2015 (300 voices/200+ sentences)
- Appen (200 voices/400 sentences)
Algorithm Learning Methods - K-means Clustering - TIMIT

Projected axis 1

Projected axis 2

Cluster Assignments and Centroids

Clusters: 1, 2, 3

Centroids:

Clustering result for TIMIT database
Algorithm Learning Methods: Elbow Curve for APPEN

- Curve of mean of within cluster sum of distances to the centroid to number of clusters.
- Point where elbow hits gives best k value.
- Determined by calculating the slope of the curve.
Algorithm Learning Methods: Overlap Criteria for APPEN

- Calculate the number of overlap groups for every case around the elbow point.

- The minimum number of overlapping gives the best k value.

- In case of a tie, look at the percentage of overlap in those cases and the lower percentage is chosen to be the best k value.

- Number of overlap cluster for k = 5 is 3

- Number of overlap cluster for k = 6 is 3

- **Number of overlap cluster for k = 7 is 2**
  - This is the best choice
Clustering Results

- Training Database: Appen
- Number of voices: 200
- Number of sentences: 100
- Number of Gaussian mixtures: 1024
- Number of classes: 7
- Total variability dimension: 100

<table>
<thead>
<tr>
<th>Cluster Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice Count</td>
<td>1 M 36 F</td>
<td>23 M 3 F</td>
<td>0 M 18 F</td>
<td>0 M 40 F</td>
<td>29 M 0 F</td>
<td>48 M 0 F</td>
<td>0 M 2 F</td>
</tr>
</tbody>
</table>

Cluster 6:
Cluster 3:
Cluster 2:
SoundAlike™ Test Results: Gordon target voice

- Gordon
- Best Match Voice (Cluster 6)
- Next best voice from same cluster
- Voice from different cluster (Cluster 5)

Classification Test: 5 clips per voice.

- Extract the incoming voice’s i-vector using trained model.
- 2 pass test: Compare i-vectors of the test voice with all the clusters. Find the best cluster.
- Next, within the cluster, identify the best voice.
- Scoring criteria: Log-likelihood
- Minimum sentence to classify is 1.
SounAlike™ Test Results: Karla target voice

- Karla

- Best Match Voice (Cluster 4)

- Next best voice from same cluster

- Voice from different cluster (Cluster 1)

- Classification Test: 5 clips per voice.

- Extract the incoming voice’s i-vector using trained model.

- 2 pass test: Compare i-vectors of the test voice with all the clusters. Find the best cluster.

- Next, within the cluster, identify the best voice.

- Scoring criteria: Log-likelihood

- Minimum sentence to classify is 1.
Conclusions

• A quick approach to finding a suitable TTS to synthesize a voice that’s nearest to a target voice.

• Same Target voice can possibly be represented by multiple TTS from different clusters depending on voice qualities affected by emotions.

• SoundAlike™ is text independent and language independent.
  – Which means that for a short sample we have an immediate TTS solution.
  – Across languages we have a fast TTS voice representation.

• Next we will be running formal evaluations on multiple libraries.
You can email me at
Fathy@speechmorphing.com
Cell: 831-419-5008
Thank You