Mobile and Ubiquitous Computing on Smartphones

Chapter 9a: Voice Analytics

Emmanuel Agu

Speech Analytics

Voice Based Analytics

- Voice can be analyzed, lots of useful information extracted
  - Who is talking? (Speaker identification)
  - How many social interactions a person has a day
  - Emotion of person while speaking
  - Anxiety, depression, intoxication, of person, etc.
- For speech recognition, voice analytics used to:
  - Discard useless information (background noise, etc)
  - Identify and extract linguistic content

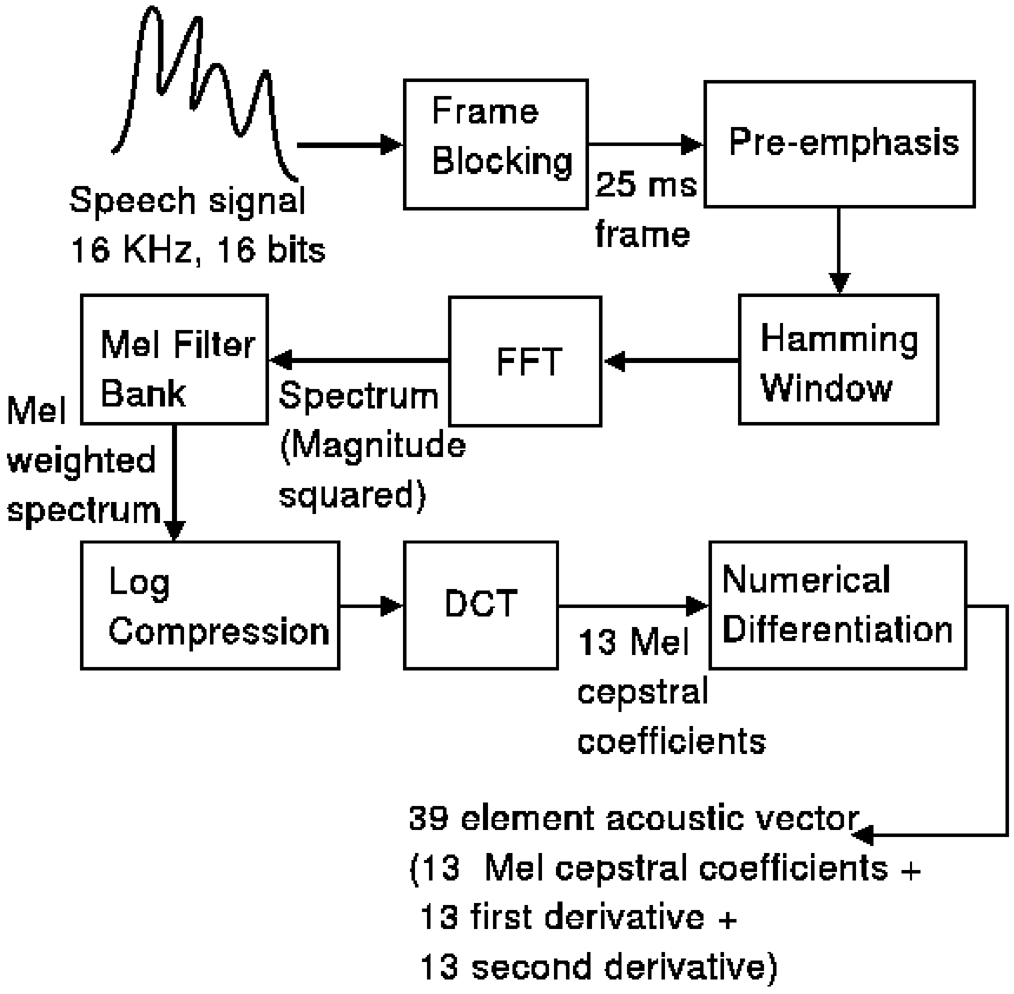
Mel Frequency Cepstral Coefficients (MFCCs)

- MFCCs widely used in speech and speaker recognition for representing envelope of **power spectrum of voice**
- **Power spectrum?** Amount of power at various frequencies
- Roughly
  - Male voice: low frequency (bass)
  - Female voice: high frequency (treble)
- Popular approach in Speech recognition
  - MFCC features + Hidden Markov Model (HMM) classifiers

MFCC Steps: Overview

1. Frame the signal into short frames.
2. For each frame calculate the periodogram estimate of the power spectrum.
3. Apply the mel filterbank to the power spectra, sum the energy in each filter.
4. Take the logarithm of all filterbank energies.
5. Take the DCT of the log filterbank energies.
6. Keep DCT coefficients 2-13, discard the rest.

MFCC Computation Pipeline



Step 1: Windowing

- Audio is continuously changing.
- Break into short, overlapping segments (20-40 milliseconds)
- Can assume audio does not change in short window

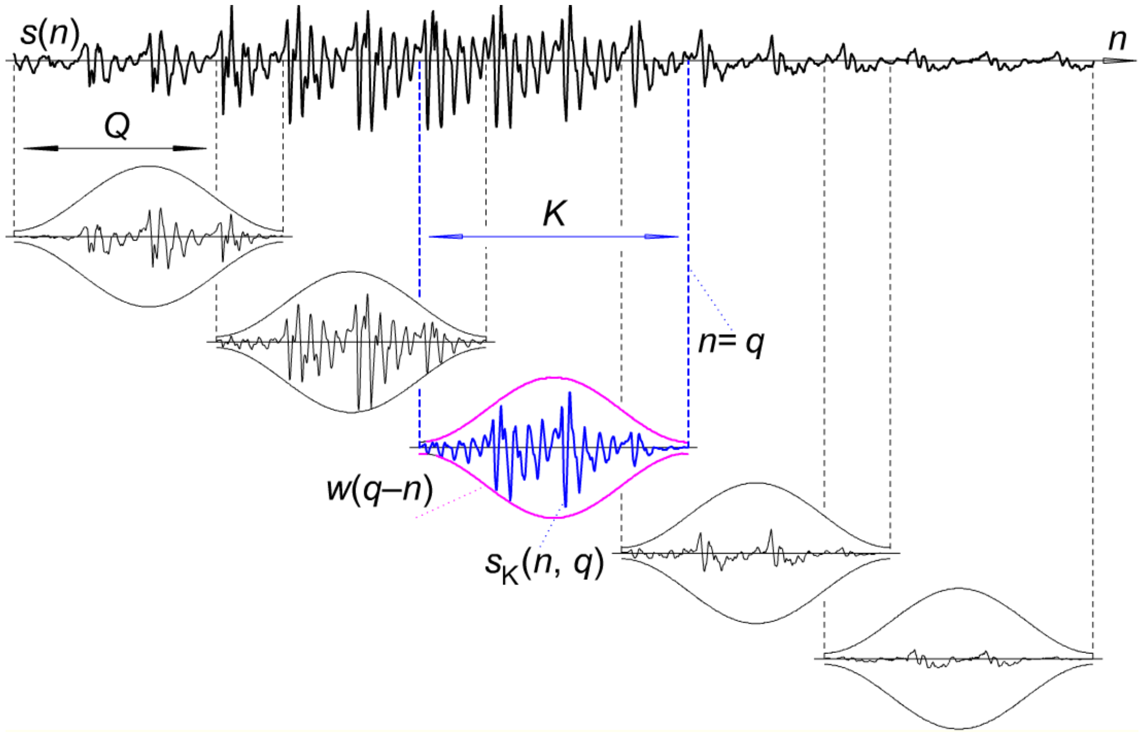
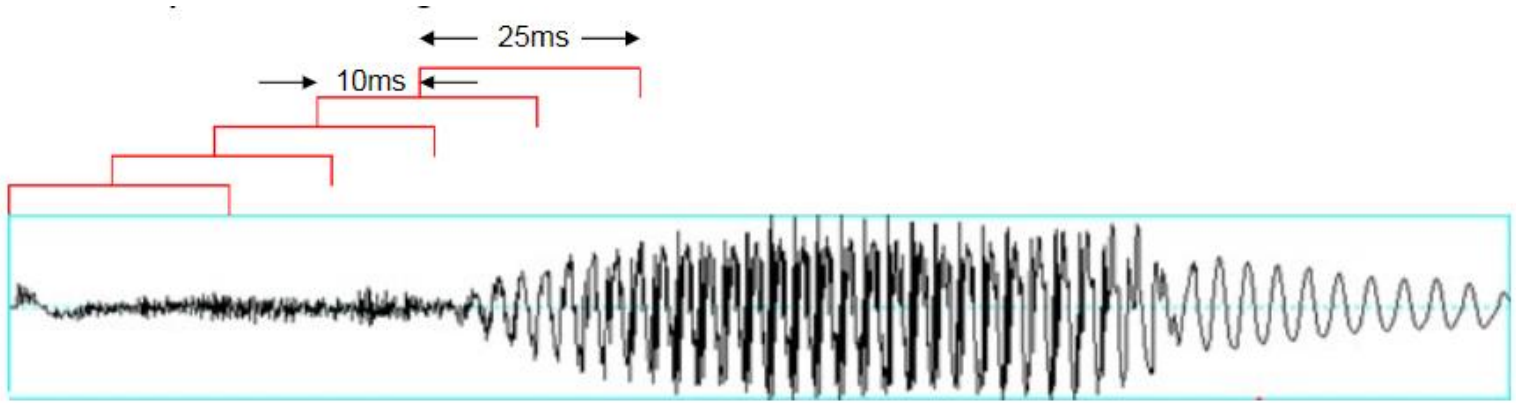


Image credits: http://recognize-speech.com/preprocessing/cepstral-mean-normalization/10-preprocessing

Step 1: Windowing

- Essentially, break into smaller overlapping frames
- Need to select frame length (e.g. 25 ms), shift (e.g. 10 ms)

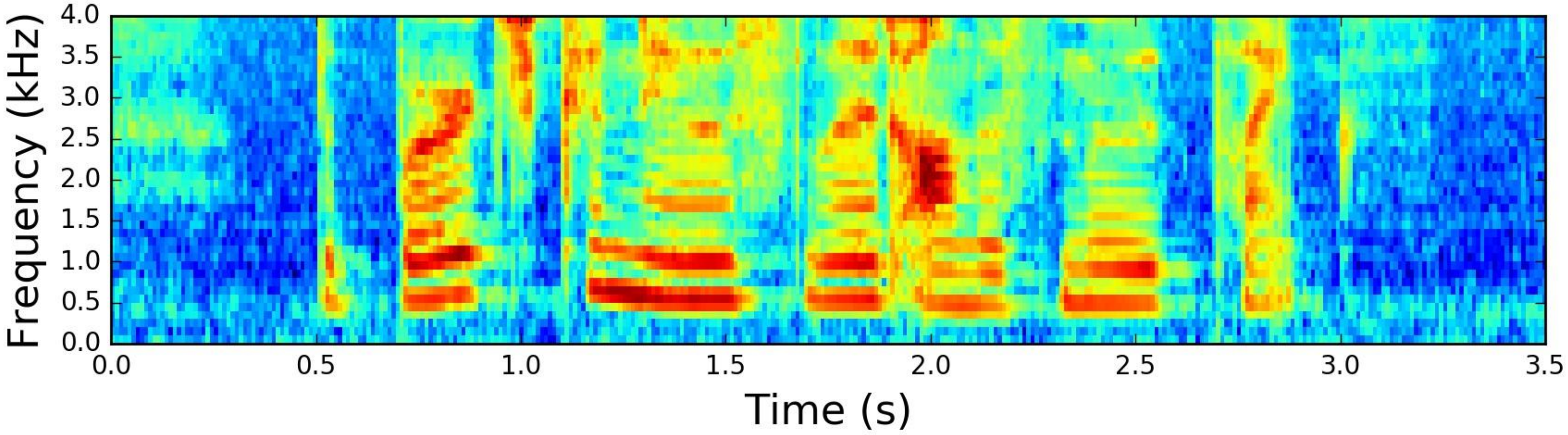


- So what? Can compare frames from reference vs test audio (i.e. calculate distances between them)

http://slideplayer.com/slide/7674116/

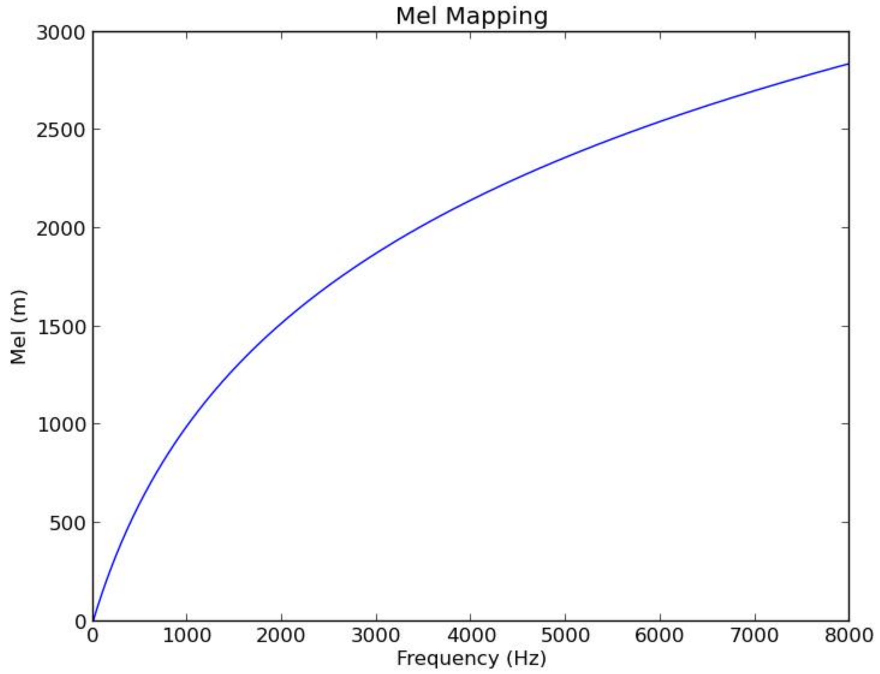
Step 2: Calculate Power Spectrum of each Frame

- Cochlea (Part of human ear) vibrates at different parts depending on sound frequency
- Power spectrum Periodogram similarly identifies frequencies present in each frame



Background: Mel Scale

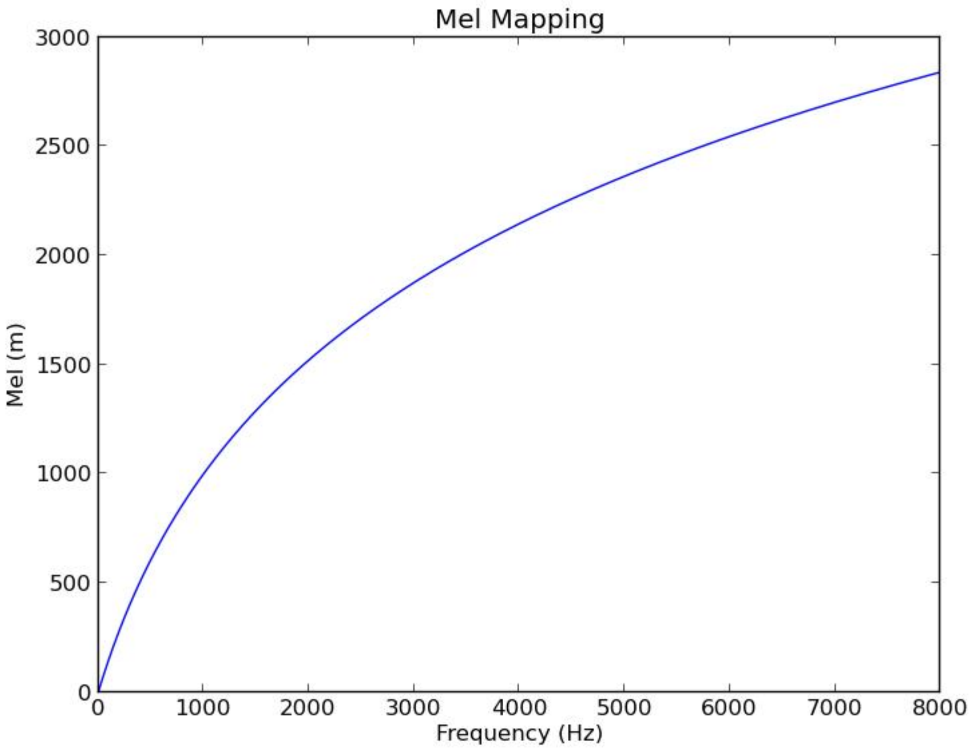
- Transforms speech attributes (frequency, tone, pitch) on non-linear scale based on human perception of voice
- Result: non-linear amplification, MFCC features that mirror human perception
    - E.g. humans good at perceiving small change at low frequency than at high frequency



Step 3: Apply Mel FilterBank

- Non-linear conversion from frequency to Mel Space

$$M(f) = 1125 \ln(1 + f/700) \quad (1)$$



Step 4: Apply Logarithm of Mel Filterbank

- Take log of filterbank energies at each frequency
- This step makes output mimic human hearing better
  - We don't hear loudness on a linear scale
  - Changes in loud noises may not sound different

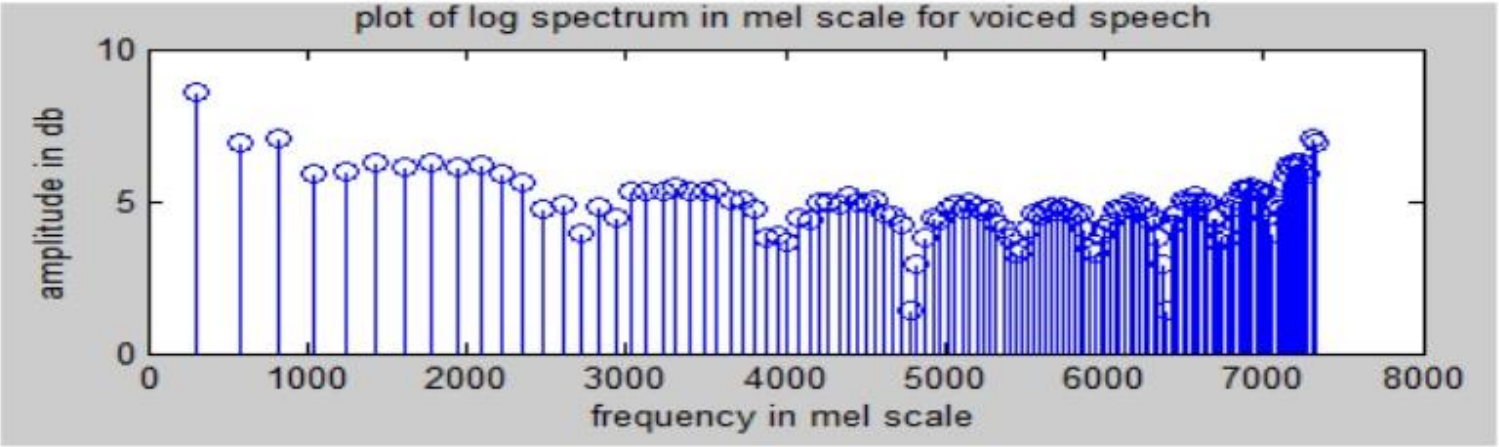


Fig.7. Spectrum of voiced speech

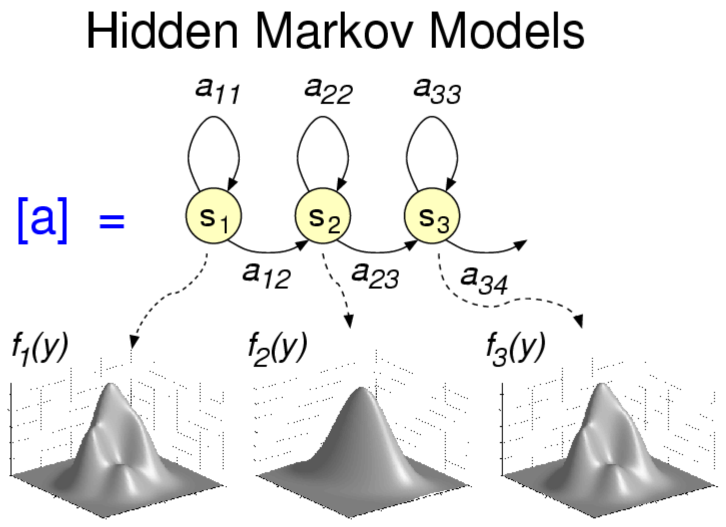
Step 4: Apply Logarithm of Mel Filterbank

- **Step 5: DCT of log filterbank:**
  - There are correlations between signals at different frequencies
  - Discrete Cosine Transform (DCT) extracts most useful and independent features

- Final result: 39-element acoustic vector used in speech processing algorithms

Speech Classification

- Human speech can be broken into phonemes
- Example of phoneme is /k/ in the words (**c**at, s**ch**ool, s**k**ill)
- Classic Speech recognition tries to recognize sequence of phonemes in a word
- Typically uses Hidden Markov Model (HMM)
  - Recognizes letters, then words, then sentences
  - Like a state machine that strings together sequence of sounds recognized



Speech/Language Analytics/NLP

Audio Project Ideas

- OpenAudio project, http://www.openaudio.eu/
- Many tools, dataset available
  - OpenSMILE: Tool for extracting > 1000 audio features
    - Windowing
    - MFCC
    - Pitch
    - Statistical features, etc
    - Supports popular file formats (e.g. Weka)
  - OpenEAR: Toolkit for automatic speech emotion recognition
  - iHeaRu-EAT Database: 30 subjects recorded speaking while eating

Affect Detection

Definitions

- Affect
  - Broad range of feelings
  - Can be either emotions or moods
- Emotion
  - Brief, intense feelings (anger, fear, sadness, etc)
  - Directed at someone or something
- Mood
  - Less intense, not directed at a specific stimulus
  - Lasts longer (hours (4?) or days)

Physiological Measurement of Emotion

- **Biological arousal:** heart rate, respiration, perspiration, temperature, muscle tension

- **Expressions:** facial expression, gesture, posture, voice intonation, breathing noise

Emotion	Physiological Response
Anger	Increased heart rate, blood vessels bulge, constriction
Fear	Pale, sweaty, clammy palms
Sad	Tears, crying
Disgust	Salivate, drool
Happiness	Tightness in chest, goosebumps

Affective State Detection from Facial + Head Movements

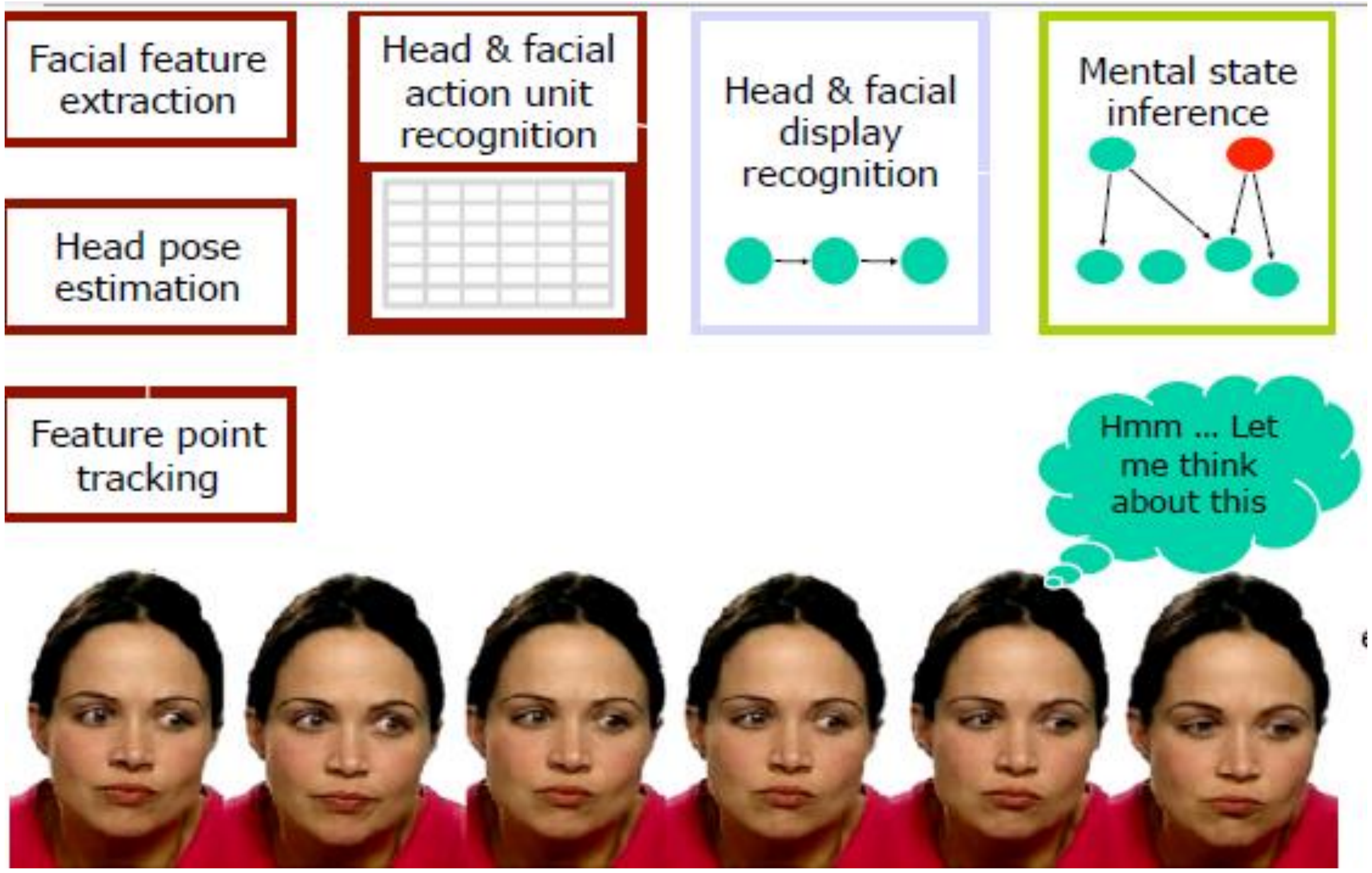


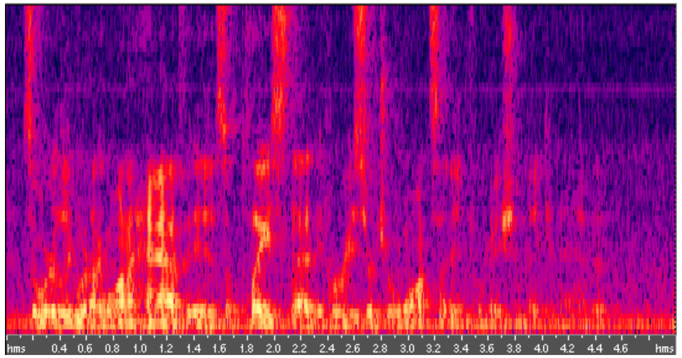
Image credit: Deepak Ganesan

Audio Features for Emotion Detection

- MFCC widely used for analysis of speech content, Automatic Speaker Recognition (ASR)
  - Who is speaking?
- Other audio features exist to capture sound characteristics/dynamics (prosody)
  - Useful in detecting emotion in speech
- **Pitch:** the frequency of a sound wave. E.g.
  - Sudden increase in pitch => Anger
  - Low variance of pitch => Sadness

Audio Features for Emotion Detection

- **Intensity:** Energy of speech, intensity. E.g.
  - Angry speech: sharp rise in energy
  - Sad speech: low intensity

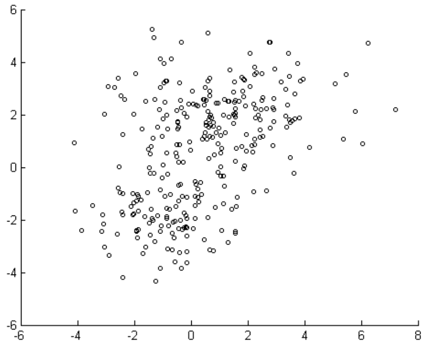


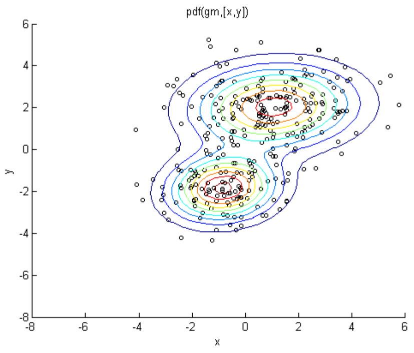
- **Temporal features:**
  - Speech rate, voice activity (e.g. pauses)
  - E.g. Sad speech: slower, more pauses

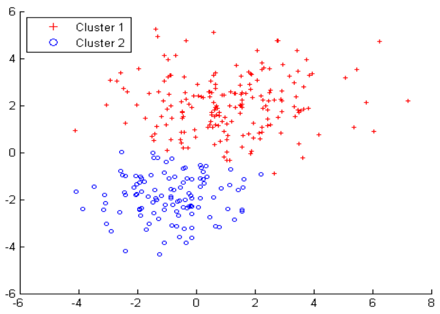
- **Other emotion features:** Voice quality, spectrogram, statistical measures

Gaussian Mixture Model (GMM)

- GMM used to classify audio features (e.g. depressed vs not depressed)
- **General idea:**
  1. Plot subjects in a multi-dimensional feature space
  2. Cluster points (e.g. depressed vs not depressed)
  3. Fit to gaussian (normal) distribution (assumed)
  4. Parameters of GMM are features for classification of health condition







Uses of Affect Detection
E.g. Using Voice on Smartphone

- Audio processing (especially to detect affect, mental health) can revolutionize healthcare
  - Detection of mental health issues automatically from patients voice
  - Population-level (e.g campus wide) mental health screening
  - Continuous, passive stress monitoring
    - Suggest interventions: breathing exercises, play relaxing music
  - Monitoring social interactions, recognize conversations (number and duration per day/week, etc)

Voice Analytics Example: SpeakerSense (Lu et al)

Lu, H., Brush, A.B., Priyantha, B., Karlson, A.K. and Liu, J., 2011, June. Speakersense: Energy efficient unobtrusive speaker identification on mobile phones. In *International conference on pervasive computing* (pp. 188-205). Springer, Berlin, Heidelberg.

- Identifies speaker, who conversation is with
- Used GMM to classify pitch and MFCC features

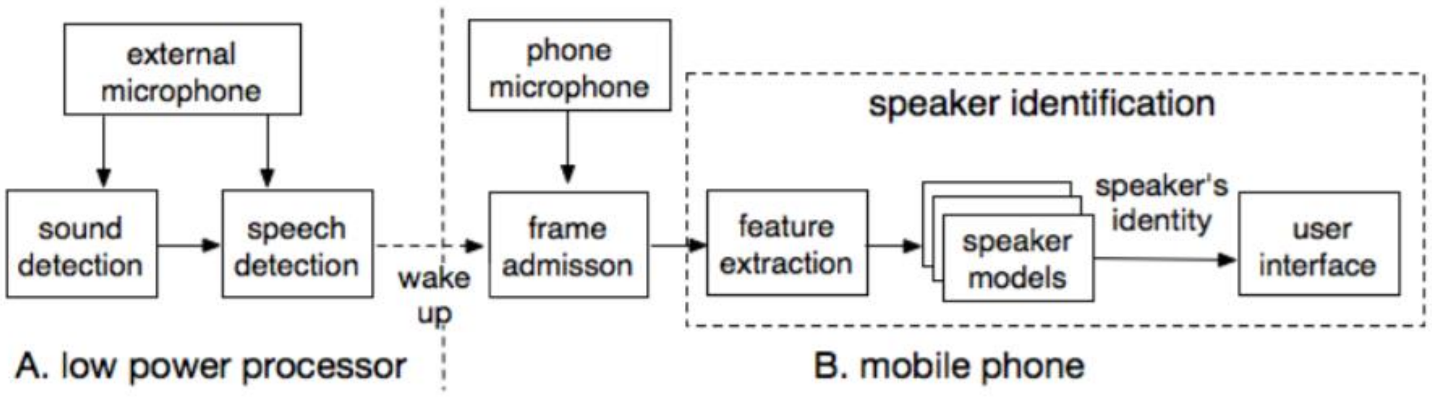


Fig. 1. The SpeakerSense architecture.

Voice Analytics Example: StressSense (Lu et al)

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M.S., Chittaranjan, G.T., Campbell, A.T., Gatica-Perez, D. and Choudhury, T., 2012, September. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 351-360).

- Detected stress in speaker’s voice
- Features: MFCC, pitch, speaking rate
- Classification using GMM
- Accuracy: indoors (81%), outdoors (76%)

Voice Analytics Example: Mental Illness Diagnosis

- What if depressed patient lies to psychiatrist, says “I’m doing great”
- Mental health (e.g. depression) detectable from voice, can be used to detect lying patient
- Doctors pay attention to speech aspects when examining patients

Category	**Patterns**
Rate of speech	slow, rapid
Flow of speech	hesitant, long pauses, stuttering
Intensity of speech	loud, soft
Clarity	clear, slurred
Liveliness	pressured, monotonous, explosive
Quality	verbose, scant

- E.g. depressed people have slower responses, more pauses, monotonic responses and poor articulation

Detection of COVID from Respiratory sounds

Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P. and Mascolo, C., 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. *arXiv preprint arXiv:2006.05919*.

- large-scale crowdsourced dataset of respiratory sounds collected to aid diagnosis of COVID-19.
- Coughs and breathing to understand how discernible COVID-19 sounds are from those in asthma or healthy controls.
- Simple binary machine learning classifier is able to classify correctly healthy and COVID-19 sounds.
- Were able to distinguish
  - User who had COVID-19 + cough vs healthy user with a cough
  - Users who had COVID-19 + cough vs. Users with asthma and a cough.
- Models achieved an Area Under the Curve (AUC) of above 80% across all tasks.

Detection of COVID from Respiratory sounds

Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P. and Mascolo, C., 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. *arXiv preprint arXiv:2006.05919*.

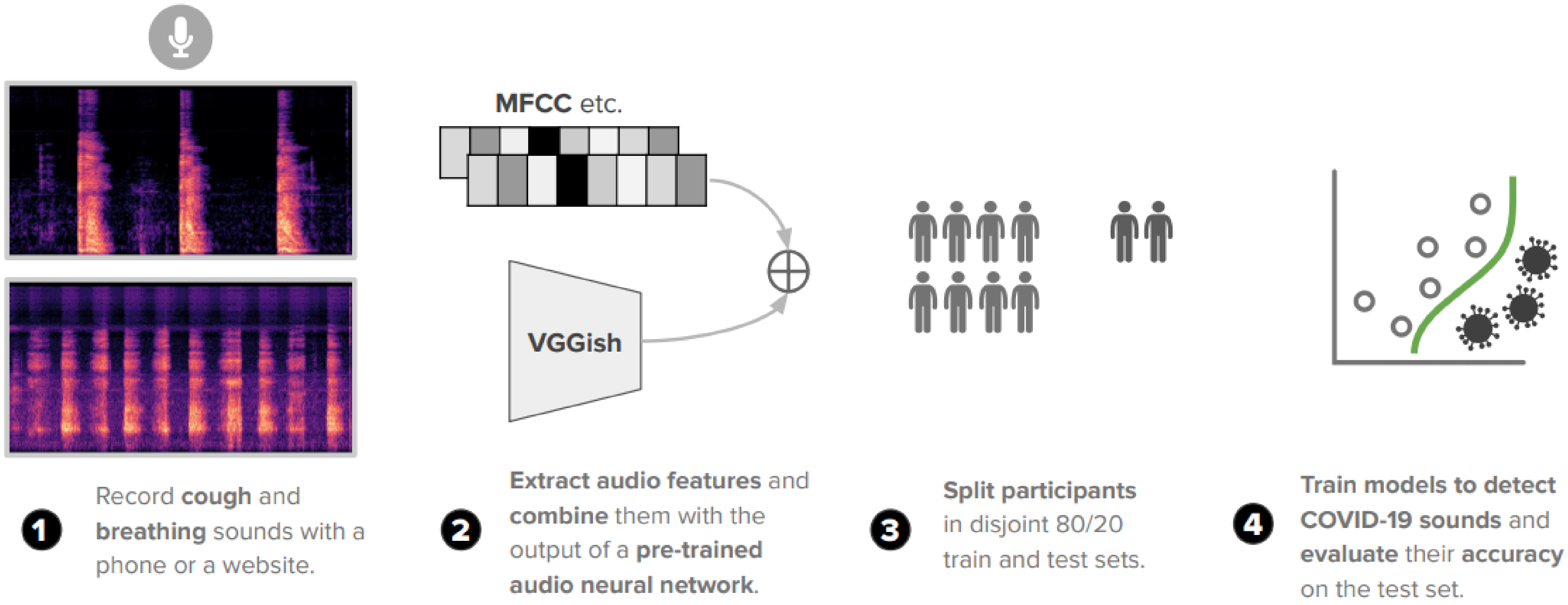


Figure 4: Description of our machine learning pipeline, describing sounds input (coughs and breathing), the extracted feature vector, and our training and testing split of the users that are used to train classification models.