Speech Processing 11-492/18-492

Speaker ID
Who is speaking?

- **Speaker ID, Speaker Recognition**
- **When do you use it**
  - Security, Access
  - Speaker specific modeling
    - Recognize the speaker and use their options
  - Diarization
    - In multi-speaker environments
    - Assign speech to different people
    - Allow questions like did Fred agree or not.
Voice Identity

What makes a voice identity

- Lexical Choice:
  - Woo-hoo,
  - I pity the fool ...

- Phonetic choice

- Intonation and duration

- Spectral qualities (vocal tract shape)

- Excitation
Voice Identity

What makes a voice identity

- Lexical Choice:
  - Woo-hoo,
  - I pity the fool …

- Phonetic choice

- Intonation and duration

- Spectral qualities (vocal tract shape)

- Excitation

But which is most discriminative?
GMM Speaker ID

- Just looking at spectral part
  - Which is sort of vocal tract shape
- Build a single Gaussian of MFCCs
  - Means and Standard Deviation of all speech
  - Actually build N-mixture Gaussian (32 or 64)
- Build a model for each speaker
- Use test data and see which model its closest to
GMM Speaker ID

- **How close does it need to be?**
  - One or two standard deviations?

- **The set of speakers needs to be different**
  - If they are closer than one or two stddev
  - You get confusion.

- **Should you have a “general” model**
  - Not one of the set of training speakers
GMM Speaker ID

- **Works well on constrained tasks**
  - In similar acoustic conditions
  - (not phone vs wide-band)
  - Same spoken style as training data
  - Cooperative users

- **Doesn’t work well when**
  - Different speaking style (conversation/lecture)
  - Shouting whispering
  - Speaker has a cold
  - Different language
Speaker ID Systems

- **Training**
  - Example speech from each speaker
  - Build models for each speaker
  - (maybe an exception model too)

- **ID phase**
  - Compare test speech to each model
  - Choose “closest” model (or none)
Basic Speaker ID system
**Accuracy**

- **Works well on smaller sets**
  - 20-50 speakers

- **As number of speakers increase**
  - Models begin to overlap – confuse speakers

- **What can we do to get better distinctions**
What about transitions

- Not just modeling isolated frames
- Look at phone sequences
- But ASR
  - Lots of variation
  - Limited amount of phonetic space
- What about lots of ASR engines
Phone-based Speaker ID

- Use *lots* of ASR engines
  - But they need to be different ASR engines
- Use ASR engines from lots of different languages
  - It doesn’t matter what language the speech is
  - Use many different ASR engines
  - Gives lots of variation
- Build models of what phones are recognized
  - Actually we use HMM states not phones
Phone-based SID (Jin)
Phone-based Speaker ID

- Much better distinctions for larger datasets
- Can work with 100 plus voices
- Slightly more robust across styles/channels
But we need more ...

- **Combined models**
  - GMM models
  - Ph-based models
  - Combine them
  - Slightly better results

- **What else ...**
  - Prosody (duration and F0)
Can VC beat Speaker-ID

- Can we fake voices?
- Can we fool Speaker ID systems?
- Can we make lots of money out of it?

- Yes to the first two
  - Jin, Toth, Black and Schultz ICASSP2008
Training/Testing Corpus

- **LDC CSR-I (WSJ0)**
  - US English studio read speech
  - 24 Male speakers
  - 50 sentences training, 5 test
  - Plus 40 additional training sentences
  - Sentence average length is 7s.

- **VT Source speakers**
  - Kal_diphone (synthetic speech)
  - US English male natural speaker (not all sentences)
Experiment I

- **VT GMM**
  - Kal_diphone source speaker
  - GMM train 50 sentences
  - GMM transform 5 test sentences

- **SID GMM**
  - Train 50 sentences
  - (Test natural 5 sentences, 100% correct)
GMM-VT vs GMM-SID

VT fools GMM-SID 100% of the time
Not surprising (others show this)
- Both optimizing spectral properties

These used the same training set
- (different training sets doesn’t change result)

VT output voices sounds “bad”
- Poor excitation and voicing decision

Human can distinguish VT vs Natural
- Actually GMM-SID can distinguish these too
- If VT included in training set
◆ **VT is always S17, S24 or S20**
◆ **Kal_diphone is recognized as S17 and S24**
◆ **Phone-SID seems to recognized source speaker**
What about Synthetic Speech?

- **Clustergen: CG**
  - Statistical Parametric Synthesizer
  - MLSA filter for resynthesis
- **Clunits: CL**
  - Unit Selection Synthesizer
  - Waveform concatenation
Synth vs GMM-SID

- Smaller is better
**Smaller is better**

**Opposite order from GMM-SID**
Conclusions

◆ **GMM-VT fools GMM-SID**

◆ **Ph-SID can distinguish source speaker**
  - **Ph-SID cares about dynamics**

◆ **Synthesis (pretty much) fools Ph-SID**
  - **We’ve not tried to distinguish Synth vs Real**
Much larger dataset
- 250 speakers (male and female)
- Open set (include background model)
- WSJ (0+1)

Use VT with long term dynamics
- HTS adaptation
- Articulatory position data
- Prosodics (F0 and duration)

Use ph-SID to tune VT model
VT that fools Ph-SID

- Develop X-SID (prosody?)
  - Develop X-VT that fools X-SID
    - Develop X2-SID
      - Develop X2-VT that fools …

…..
De-identification

- Using Speaker ID to score de-identification
  - Reverse of voice transformation
    - Masking source, rather than being like target

- Simplest view
  - Full ASR and TTS in new engine (two hard)

- Voice conversion to synthetic voice
  - Natural speech to TTS (kal_diphone)
De-identification

- Tested against 24 speakers
- GMM transformation
  - 50% de-identification
- GMM+duration normalization
  - 60% de-identification
- GMM+duration+transinterpolation
  - 80-100% de-identification
Speaker-ID and Language

- Identify which language someone is talking
- Identify their dialect
- In Cross-lingual voice conversion
  - Identify the accent (or lack of)
  - Identify the speaker
  - Want close to source speaker and close to target language
Annual international competitions

- Given this data set (1000s speakers)
- How well can you identify the test speakers
- Vary the issues:
  - Channel conditions (phone, non-phone)
  - Language/Speaker style
  - Realtime vs fully offline