Speech Processing 15-492/18-492

Voice Conversion 2
De-identification

- **Remove speaker identity**
  - But keep it still human like

- **Health Records**
  - HIPAA laws require this
  - Not just removing names and SSNs

- **Use Voice conversion to get “new” voices**
De-identification

- **Best would be ASR to text and TTS**
  - *But it would all sound the same*
  - *And it would loose spontaneity*
- **VC to some example speaker**
  - *But then you’d need lots of example speakers*
  - *And you can still detect some properties of source speaker*
De-identification by GMM VC

- **Using standard GMM VC**
  - Can still identify 50% of the voices (out of 24)
  - (Human’s cant but machine can)

- **Need something more extreme**
GMM VC plus duration

- Find the average duration of source and target speakers
- Modify length of speech by factor
  - Has to be overall factor as no phoneme information is available
- For de-identification
  - < 30% still identified
- Need to be more extreme
What about transterpolation > 1.0?

- Certainly gives another voice
- Moved further away from source
- Original
- GMM to kal
- GMM to kal trans 1.2
- GMM to kal trans 2.0
How can tell if VC works?

Ask people
  - Does it sound like the target
  - (Does it not sound like the source)

Use objective metrics
  - Some automatic score
Human listening tests

- **ABX tests**
  - Source, Target and transformed speech
  - Does X sound more like A or B

- **AX tests**
  - Source/Target and transformed speech
  - Were A and X produced by same or different speakers

- Over multiple listeners you can get consensus

- Note different results for A->B than B->A
What if you know the speakers?

- **Test with CMU voices**
  - CMU listeners
  - Non-CMU listeners

- **Results are still basically the same**
  - Voices that convert better are the same
  - Know the speakers doesn’t make a difference
Objective Measures

• Need to have a automatic measure too

• Mel-Cepstral Distortion
  – Euclidean distance between MFCC
    \[ \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{24} \left( mc_d^{(t)} - mc_d^{(e)} \right)^2} \]
  – Lower order MFCCs have bigger magnitudes
  – Thus this is scaled to favor lower order MFCCs
  – Scaling factor is to make it a nice number

• Results are between 3.5 and 6.5 (smaller is better)
Cross Lingual Voice Conversion

- **Have your voice in another language**
  - Speech to Speech translations systems
- **It’s OK to be non-native accented**
- **But we need parallel data**
  - But I can’t speak X
  - Fake it but it should *very* accented
  - Use source phones, but it sound *very* *very* accented
Cross-Lingual VC

- **Find a bilingual speaker**
  - A(german) and A(english)

- **For your non-bi-lingual speaker B(english)**
  - Build A(english)->B(english) VC model

- **For cross-lingual B(german)**
  - A(german) plus A(english)->B(english)

- **Sort of works, but not very well**
Cross-lingual VC

- A voice has both
  - Speaker specific components
  - Language specific components
- For CLVC you want to separate these
- How do you evaluate it?
  - With bilingual speaker
  - With human listeners
    - Does it have an accent
    - Where is this person from
    - Is this the same person speaking
Backwards speech

- **Playing speech backwards**
  - Still has speaker properties
  - Can be language independent
  - Speaker 1
  - Speaker 2
  - Speaker 2 or 1
New Language with VC

- **In order to build support in new languages**
- **Use existing language databases**
  - Find “similar” phones in different languages
  - Synthesis with these “similar” phones”
- **Collect parallel data**
  - Xenophone synthesis (X)
  - Native speaker (N)
- **Build VC between X->N**
  - Use as filter on xenophone synthesis
- **Sort of works**
  - Some people don’t do the VC stage!
VC factors

- *Can be done in real time*
  - *Delay needn’t be more than a few frames*

- *Don’t need phonetic information*
  - *Though some people now do this*
  - *Synthesis plus VC can do this*

- *Can be done in any language*
  - *Only need F0, MFCC and power for features*
VC vc Synthesis

- **In general**
  - Voice conversion
    - Clear and understandable but
    - Not have full target voice properties
  - Synthesis
    - Bad joins and disfluencies
    - Clear properties of target voice
- **Really bad unit selection**
  - E.g. with too small databases (10 utts)
  - Still has good target voice properties
VC and SPS

- **Becoming closely related**
  - Small amount of target speaker
  - Use larger background models