Speech Separation: Evaluation

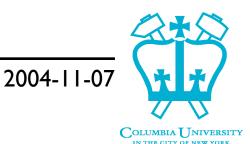
Dan Ellis

Laboratory for Recognition and Organization of Speech and Audio Dept. Electrical Eng., Columbia Univ., NY USA dpwe@ee.columbia.edu

- I. Task and goal
- 2. Metrics
- 3. Corpora



Speech Separation: Evaluation - Dan Ellis



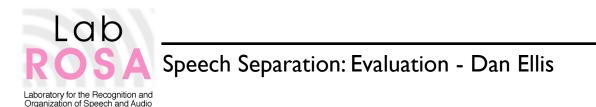
I. Why Evaluation?

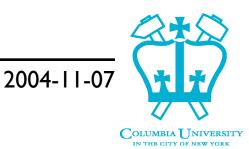
- Evaluation helps progress

 by revealing what works
 by making a field 'fundable'

 Speech separation is ready

 diverse range of approaches
 - .. some of which work quite well!

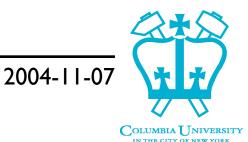




What Task?

- Evaluation task should be a real problem
 - .. because it may be all that is solved
- What are the motivating problem(s)?
 - spoken commands for machines
 - machine-directed speech \rightarrow words
 - indexing/surveillance
 - natural speech \rightarrow words
 - hearing aids
 - natural speech \rightarrow audio
- Compare to human performance?

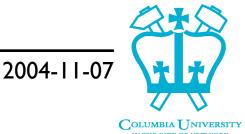




What Data?

- Need audio + ground truth
 - pre-mixing sources?
 - word-level transcripts
- Real vs. synthetic recordings
 - artificial mixtures are too easy (ICA case)
 - real field recording is inflexible (enumerate cases)
 - high-realism synthetic mixtures?
 - including time-varying reverb, many sources...
 - developed to exactly match properties of field recordings
 - .. but then much more flexible





What Metric?

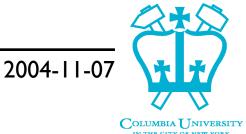
• SNR is sufficient but not necessary

- needs pre-mixture sources
- .. can factor out optimal spectral normalization
- PEAQ includes psychoacoustic model of distortion prominence

• Subjective content measures

- play resynthesis to human listeners for MOS?
 - confounds separation and resynthesis
- have listeners transcribe some aspect of content (e.g. words); score on ability to reproduce that
 - solving only part of the problem





Existing Datasets

- ICSI Meeting Recorder (2001)
 - 100 hr, transcribed, ~5% overlapping, 16 ch

• ShATR Crossword Task (1994)

- simultaneous discussions, 37 min, transcribed, 8 ch
- SPINE (2000)
 - o game in noisy room, 12 hr, transcribed, 1 ch
- LDC Fletcher corpus
- Aurora-3 (SpeechDat-Car)
- AV-TREC
- CHIL
- Personal Audio





COLUMBIA I INIVERSITY

Conclusions

- Evaluation is important
 .. and the field is ready
- Task: Something real and relevant
- Metric: Portable and relevant
- What now?
 - what percentage of effort should be devoted to evaluation?
 - and how can the load be spread?

