Speech Separation in Humans and Machines

Dan Ellis

Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

dpwe@ee.columbia.edu             http://labrosa.ee.columbia.edu/

1. The Speech Separation Problem
2. Human Performance
3. Source Separation
4. Source Inference
5. Concluding Remarks
1. Speech Separation

- Speech rarely occurs in isolation
  - but recognizing mixed speech is a problem
  - for humans and machines
Speech Separation Scenarios

- **Interactive voice systems**
  - human-level understanding is expected

- **Speech prostheses**
  - crowds: #1 complaint of hearing aid users

- **Archive analysis**
  - identifying and isolating speech

- **Surveillance...**
How Can We Separate?

- **By between-sensor differences** (spatial cues)
  - ‘steer a null’ onto a compact interfering source

- **By finding a ‘separable representation’**
  - spectral? but speech is broadband
  - periodicity? maybe – for voiced speech
  - something more signal-specific...

- **By inference** (based on knowledge/models)
  - speech is redundant
    → use part to guess the remainder
Outline

1. The Speech Separation problem
2. Human Performance
   - scene analysis
   - speech separation by location
   - speech separation by voice characteristics
3. Source Separation
4. Source Inference
5. Concluding Remarks
Auditory Scene Analysis

• Listeners **organize** sound mixtures into discrete perceived **sources** based on within-signal **cues** (audio + ...)

  ○ common
  ○ onset
  ○ + continuity
  ○ harmonicity

  ○ spatial, modulation, ...
  ○ learned “schema”

---

Speech Separation - Dan Ellis
Speech Mixtures: Spatial Separation

- **Task:** Coordinate Response Measure
  - “Ready Baron go to green eight now”
  - 256 variants, 16 speakers
  - correct = color and number for “Baron”

- **Accuracy as a function of spatial separation:**

  ![Graph showing accuracy as a function of spatial separation]

  - A, B same speaker

  Brungart et al.’02

A, B same speaker
Separation by Vocal Differences

• CRM varying the level and voice character
  ○ (same spatial location)

energetic vs. informational masking
Varying the Number of Voices

- Two voices OK;
- More than two voices harder
  - (same spatial origin)

- mix of $N$ voices tends to speech-shaped noise...
Outline

1. The Speech Separation problem
2. Human Performance
3. Source Separation
   - Independent Component Analysis
   - Computational Auditory Scene Analysis
4. Source Inference
5. Concluding Remarks
Machine Separation

- Problem: Features of combinations are not combinations of features
  - voice is easy to characterize when in isolation
  - redundancy needed for real-world communication
Separation Approaches

**ICA**
- Multi-channel
- Fixed filtering
- Perfect separation – maybe!

\[
\text{target } x - n
\]

**CASA / Model-based**
- Single-channel
- Time-varying filtering
- Approximate Separation

\[
\text{mix } x + n - \overset{\text{stft}}{\text{spectrogram}} - \overset{\text{mask}}{\text{proc}} - \overset{\text{istft}}{\hat{x}}
\]

• Very different approaches!
Independent Component Analysis

- **Central idea:**
  Search **unmixing space** to maximize independence of outputs

  \[
  \begin{bmatrix}
  a_{11} & a_{12} \\
  a_{21} & a_{22}
  \end{bmatrix} \times \begin{bmatrix}
  s_1 \\
  s_2
  \end{bmatrix} \rightarrow \begin{bmatrix}
  x_1 \\
  x_2
  \end{bmatrix}
  \]

  \[
  \begin{bmatrix}
  w_{11} & w_{12} \\
  w_{21} & w_{22}
  \end{bmatrix} \times \begin{bmatrix}
  x_1 \\
  x_2
  \end{bmatrix} \rightarrow \begin{bmatrix}
  \hat{s}_1 \\
  \hat{s}_2
  \end{bmatrix}
  \]

  \[
  \frac{-\delta \text{MutInfo}}{\delta w_{ij}}
  \]

- simple mixing
  \[\rightarrow \text{a good solution (usually) exists}\]
ICA Limitations

- **Cancellation** is very finicky
  - hard to get more than ~ 10 dB rejection

- The world is not instantaneous, fixed, linear
  - subband models for reverberation
  - continuous adaptation

- **Needs** spatially-compact interfering sources
Central idea: Segment time-frequency into sources based on perceptual grouping cues.

... principal cue is harmonicity.
CASA Preprocessing

- **Correlogram**: a 3rd “periodicity” axis
  - envelope of wideband channels follows pitch

- c/w Modulation Filtering \[\text{Schimmel & Atlas '05}\]

![Diagram of Correlogram process](image)
Time-Frequency (T-F) Masking

• “Local Dominance” assumption

![Time-frequency spectrograms for male and female voices](image)

- Oracle masks are remarkably effective!
- $|\text{mix} - \max(\text{male}, \text{female})| < 3\text{dB}$ for $\sim80\%$ of cells
CASA limitations

- **Driven by local features**
  - problems with aperiodic sources...
- **Limitations of T-F masking**
  - need to identify single-source regions
  - cannot undo overlaps – leaves gaps

\[ \text{from Hu \\ & Wang '04} \]
Combining Spatial + T-F Masking

- **T-F masks** based on inter-channel properties
  - [Roman et al. '02], [Yilmaz & Rickard '04]
  - multiple channels make CASA-like masks better

- **T-F masking after ICA**
  - [Blin et al. '04]
  - cancellation can remove energy within T-F cells
Outline

1. The Speech Separation problem
2. Human Performance
3. Source Separation
4. Source Inference
   - Separation vs. inference
   - Model-based separation
   - Speech Fragment Decoding
5. Concluding Remarks
Separation vs. Inference

- **Ideal** separation is rarely possible
  - i.e. no projection can completely remove overlaps

- **Overlaps** ⇒ **Ambiguity**
  - scene analysis = find “most reasonable” explanation

- **Ambiguity** can be expressed probabilistically
  - i.e. posteriors of sources \( \{S_i\} \) given observations \( X \):

\[
P(\{S_i\} | X) \propto P(X | \{S_i\}) P(\{S_i\})
\]

- **Better source models** → **better inference**
  - .. learn from examples?
Model-Based Separation

- **Central idea:**
  Employ strong **learned constraints** to **disambiguate** possible sources
  \[ \{S_i\} = \arg\max_{S_i} P(X | \{S_i\}) \]

- e.g. fit speech-trained **Vector-Quantizer** to mixed spectrum:

  - separate via T-F mask (again)

Varga & Moore'90
Roweis’03...
Can Models Do CASA?

- **Source models** can learn *harmonicity*, *onset*
  - ... to *subsume* rules/representations of CASA
  
  - can capture *spatial* info too [Pearlmutter & Zador’04]

- **Can also capture** *sequential structure*
  - e.g. consonants follow vowels
  - ... like people do?

- **But: need source-specific models**
  - ... for *every possible source*

- Use model *adaptation*? [Ozerov et al. 2005]
Separation with ASR Models

• Drive separation engine to **match** outputs to existing speech models

*ASR includes a very detailed source model*

---

From Manuel Reyes's WASPAA 2003 presentation
Separation or Description?

- Are isolated **waveforms** required?
  - clearly sufficient, but may not be necessary
  - not part of perceptual source separation!
- **Integrate** separation with application?
  - e.g. speech recognition

words output = abstract description of signal
Speech Separation - Dan Ellis

2005-11-28 - 26/35

Missing Data Recognition

• Speech models $p(x|M)$ are multidimensional...
  ○ need values for all dimensions to evaluate $p(\bullet)$

• But: can make inferences given just a subset of dimensions $x_k$
  ○ $p(x_k|M) = \int p(x_k,x_u|M)dx_u$

• Hence, missing data recognition:
  ○ hard part is finding the mask (segregation)
The Speech Fragment Decoder

• Match ‘uncorrupt’ spectrum to ASR models using missing data

• Joint search for model $M$ and segregation $S$ to maximize:

$$P(M, S|Y) = P(M) \int P(X|M) \cdot \frac{P(X|Y, S)}{P(X)} dX \cdot P(S|Y)$$

Isolated Source Model

Segregation Model

Source $X(f)$

Observation $Y(f)$
Using CASA cues

\[ P(M, S|Y) = P(M) \int P(X|M) \cdot \frac{P(X|Y, S)}{P(X)} dX \cdot P(S|Y) \]

- CASA can help search
  - consider only segregations made from CASA chunks
- CASA can rate segregation
  - construct \( P(S|Y) \) to reward CASA qualities:
Speech-Fragment Recognition

- CASA-based fragments give extra gain over missing-data recognition

![Graph showing digit recognition accuracy vs SNR (dB) for Factory Noise.](from Barker et al. ’05)
Outline

1. The Speech Separation problem
2. Human Performance
3. Source Separation
4. Source Inference
5. Concluding Remarks
   - Evaluation
   - Connecting to Perception
Evaluation

- How to measure separation performance?
  - depends what you are trying to do
- SNR?
  - energy (and distortions) are not created equal
  - different nonlinear components [Vincent et al. ’06]
- Intelligibility?
  - rare for nonlinear processing to improve intelligibility
  - listening tests expensive
- ASR performance?
  - separate-then-recognize too simplistic; ASR needs to accommodate separation
“Speech Separation Challenge”

• Mixed and Noisy Speech ASR task defined by Martin Cooke and Te-Won Lee
  ○ short, grammatically-constrained utterances:
    <command:4><color:4><preposition:4><letter:25><number:10><adverb:4>
e.g. "bin white at M 5 soon"

• Results to be presented at Interspeech’06
  ○ http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

• See also “Statistical And Perceptual Audition” workshop
  ○ http://www.sapa2006.org/
More Realistic Evaluation

- **Real-world speech tasks**
  - crowded environments
  - applications: communication, command/control, transcription

- **Metric**
  - human intelligibility?
  - ‘diarization’ annotation (not transcription)
Reconnecting to Perception

- People are (still) much better at speech recognition, including mixtures
- Can we model human separation with ASR?
  - “Glimpse model”: MD ASR using oracle local SNR
  - Listeners identify high SNR islands?

from Cooke’06
Summary & Conclusions

• **Listeners** do well separating speech
  - using spatial location
  - using source-property variations

• **Machines** do less well
  - difficult to apply enough constraints
  - need to exploit signal detail

• **Models** capture constraints
  - learn from the real world
  - adapt to sources

• **Inferring state** (≈ recognition)
  is a promising approach to separation
Sources / See Also

• NSF/AFOSR Montreal Workshops ’03, ’04
  o www.ebire.org/speechseparation/
  o labrosa.ee.columbia.edu/Montreal2004/
  o as well as the resulting book...

• Hanse meeting:
  o www.lifesci.sussex.ac.uk/home/Chris_Darwin/Hanse/

• DeLiang Wang’s ICASSP’04 tutorial
  o www.cse ohio-state.edu/~dwang/presentation.html

• Martin Cooke’s NIPS’02 tutorial
  o www.dcs.shef.ac.uk/~martin/nips.ppt
References 1/2


References 2/2


