To Separate Speech!
A System for Recognizing Simultaneous Speech

John McDonough\textsuperscript{1,2}, Kenichi Kumatani\textsuperscript{2,3}, Tobias Gehrig\textsuperscript{4},
Emilian Stoimenov\textsuperscript{4}, Uwe Mayer\textsuperscript{4}, Stefan Schacht\textsuperscript{1},
Matthias Wölfe\textsuperscript{4} and Dietrich Klakow\textsuperscript{1}

\textsuperscript{1}Spoken Language Systems, Saarland University
\textsuperscript{2}Institute for Intelligent Sensor-Actuator Systems, University of Karlsruhe
\textsuperscript{3}IDIAP Research Institute
\textsuperscript{4}Institute for Theoretical Computer Science, University of Karlsruhe

June 28, 2007
Introduction

• In a typical meeting or telephone conversation, 50% of the speech segments contain overlapping speech.

• The automatic recognition of such overlapping speech segments remains an unsolved problem, which makes the *Speech Separation Challenge (SSC)* a very interesting task.

• Our system consists of three principal components:
  – Person tracking and utterance segmenter;
  – Beamformer based on a minimum mutual information criterion;
  – An automatic speech recognition engine based on weighted finite-state transducers.

• This talk will briefly introduce each component.

• Our results on the SSC development and evaluation data will also be presented.
Time Delay of Arrival

- Let \( x \) and \( m \) respectively denote the three-dimensional position of a speaker and a microphone.

- The time delay of arrival between \( x \) and \( m \) is given by

\[
\delta(x, m) = \frac{\sqrt{(x - m_x)^2 + (y - m_y)^2 + (z - m_z)^2}}{c} = \frac{\|x - m\|}{c}
\]

where \( c \) is the speed of sound.
Consider the $i$-th pair of microphones $(m_{i1}, m_{i2})$.

The difference $T_i(x)$ in TDOAs for a speaker at position $x$ can be expressed as:

$$T_i(x) = \frac{\|x - m_{i1}\| - \|x - m_{i2}\|}{c}$$
Time Delay of Arrival Estimation

- The difference $\tau_i$ in TDOAs for the $i$-th pair of microphones can be estimated with the phase transform (PHAT) function:

$$R_{12}(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{X_1(e^{j\omega \tau})X_2^*(e^{j\omega \tau})}{|X_1(e^{j\omega \tau})X_2^*(e^{j\omega \tau})|} e^{j\omega \tau} d\omega$$  \hspace{1cm} (1)

- For efficiency, $R_{12}(\tau)$ is typically calculated with an inverse FFT.

- After the FFT, interpolation is performed to overcome the granularity in the estimate introduced by the sampling interval.

- The estimated TDOA difference $\hat{\tau}_i$ is given by

$$\hat{\tau}_i = \arg \max_{\tau} R_{12}(\tau)$$
For a given set $\{\hat{\tau}_i\}_i$ of estimated time delay differences, we can formulate the error function

$$
\epsilon(x; t) = \sum_{i=0}^{N-1} \frac{[\hat{\tau}_i - T_i(x)]^2}{\sigma_i^2}
$$

(2)

where $\sigma_i^2$ is the observation variance for the $i$-th microphone pair.

The position $x$ of the speaker can then be simply estimated as

$$
\hat{x}(t) = \arg \min_x \epsilon(x; t)
$$

So formulated, $\hat{x}(t)$ is clearly obtained as a nonlinear least squares estimate.
Kalman Filter

- Consider a dynamical system described by the process and observation equations

\[ x(t + 1) = F(t + 1, t)x(t) + \nu_1(t) \]  
\[ y(t) = C(t)x(t) + \nu_2(t) \]

where \( x(t) \) is the state and \( y(t) \) is the observation.

- The process noise \( Q_1(t) \) and observation noise \( Q_2(t) \) are zero-mean stochastic processes with covariance matrices

\[ Q_1(t) = E\{\nu_1(t)\nu_1^H(t)\} \]  
\[ Q_2(t) = E\{\nu_2(t)\nu_2^H(t)\} \]

respectively.

- The innovations process associated with \( y(t) \) is by definition

\[ \alpha(t) = y(t) - \hat{y}(t|Y_{t-1}) \]

where \( \hat{y}(t|Y_{t-1}) = C(t)\hat{x}(t|Y_{t-1}) \) is the predicted observation.
The Kalman filter has a *predictor-corrector* structure.
• The Kalman filter maintains *predicted* and *filtered state error covariance matrices*, $K(t + 1, t)$ and $K(t)$ respectively.

• $K(t + 1, t)$ is used to calculate the *innovations covariance matrix* $R(t)$. 

---

**Extended Kalman Filter Schematic**
Joint Probabilistic Data Association Filter

- The joint probabilistic data association filter is a generalization of the Kalman filter.
- The JPDAF uses multiple observations at each time step and is capable of tracking multiple targets.
For the PASCAL *Speech Separation Challenge* (SSC), far-field speech data was collected with eight-element circular arrays.

Two speakers at different locations in a room simultaneously read sentences from the *Wall Street Journal* 5K ASR tasks.

Segmentation and language models are provided.

Speech is collected with two eight-element circular arrays with 20 cm diameters.
SSC Beam Patterns

Figure 1: Circular array beam pattern for $f_s = 686$ Hz.

Figure 2: Circular array beam pattern for $f_s = 3,430$ Hz.
Figure 3: Circular array beam pattern for $f_s = 5,145$ Hz.

Figure 4: Circular array beam pattern for $f_s = 6,860$ Hz.
The weight vector of a generalized sidelobe canceller (GSC) is partitioned into a quiescent weight vector $w_q$ and an active weight vector $w_a$.

For a pure look direction constrain, the constraint space is defined by $w_q$.

Define the blocking matrix $B$ as an $N \times 1$ matrix with linearly independent columns such that

$$w_q^H B = 0$$

The orthogonal space is defined by the columns of $B$.  

---

**Figure 5:** Generalized sidelobe canceller.
Beamforming and BSS Problems

• Biggest problem in beamforming is *source cancellation* due to:
  – Steering errors,
  – Reverberation.

• Biggest problems in BSS are:
  – Permutation ambiguities.
  – Scaling ambiguities.

• **Idea:** Attempt to overcome deficiencies of beamforming and BSS by combining best parts of both.
Approach

• **Idea:**
  – Form two GSCs, one for each speaker.
  – Use JPDAF source localizer to find polar coordinates of each speaker.
  – Jointly optimize the active weights of each GSC to achieve *minimum mutual information* (MMI) of the GSC outputs.

• **Mutual information** (Gallager, 1968) is defined as

\[
I(Y_1, Y_2) = \mathcal{E} \left\{ \log \frac{p(Y_1, Y_2)}{p(Y_1)p(Y_2)} \right\}
\]
**Gaussian Assumption**

- Under a Gaussian assumption, the mutual information criterion reduces to
  \[
  I(Y_1, Y_2) = -\log \left(1 - |\rho_{12}|^2\right)
  \]
  for
  \[
  \rho_{12} = \frac{\epsilon_{12}}{\sigma_1 \sigma_2}
  \]
  where
  \[
  \sigma_i^2 = \mathcal{E}\{Y_i Y_i^*\} = (w_{q,i} - B_i w_{a,i})^H \Sigma_X (w_{q,i} - B_i w_{a,i})
  \]
  \[
  \epsilon_{12} = \mathcal{E}\{Y_1 Y_2^*\} = (w_{q,1} - B_1 w_{a,1})^H \Sigma_X (w_{q,2} - B_2 w_{a,2})
  \]

- Note that the variances \( \sigma_1 \) and \( \sigma_2 \) appear in the denominator of \( \rho_{12} \).
- Hence, the MMI optimization criterion offers the possibility of performing both *null and sidelobe* steering.
Non-Gaussianity

- The subband samples have super-Gaussian pdfs.
- Possible super-Gaussian pdfs are

\[ p(x) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|x|} \]  
  \text{(Laplace)}

\[ p(x) = \frac{1}{\pi} K_0(|x|) \]  
  \text{\(K_0\) or Bessel)}

\[ p(x) = \frac{\sqrt{3}}{4\sqrt{\pi}} \left( \frac{\sqrt{3}|x|}{2} \right)^{-1/2} e^{-\sqrt{3}|x|/2} \]  
  \text{(Gamma)}
Super-Gaussian pdfs

Figure 6: Plot of the log-likelihood of the super-Gaussian and Gaussian pdfs.
Empirical evidence clearly demonstrates that subband samples of speech are *not* Gaussian.

This confirms the assumptions made in ICA.

Table 1: Average log-likelihoods of subband speech samples for various pdfs.

<table>
<thead>
<tr>
<th>pdf</th>
<th>( \frac{1}{TK} \sum_{t=0}^{T-1} \sum_{m=0}^{M-1} \log p(X_{t,m}; \text{pdf}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>migammed ( \Gamma )</td>
<td>-0.779</td>
</tr>
<tr>
<td>K0</td>
<td>-1.11</td>
</tr>
<tr>
<td>Laplace</td>
<td>-2.48</td>
</tr>
<tr>
<td>Gaussian</td>
<td>-9.93</td>
</tr>
</tbody>
</table>
Speech Recognition Experiments

- We have adapted the JPDAF person tracker to return azimuth and elevation estimates for the circular array.
- For each utterance, we calculated “global” covariance matrices for the snapshots of each subband, then global active weight vectors.
- Our initial ASR system was trained on four RT meeting corpora plus WSJCAM0.
We used four decoding passes.

For all passes save the first unadapted pass, speaker adaptation parameters were estimated using the word lattices generated during the prior pass.

A description of the four decoding passes follows:

1. Decode with the unadapted, conventional ML acoustic model and bigram language model (LM).
2. Estimate vocal tract length normalization (VTLN) parameters and constrained maximum likelihood linear regression parameters (CMLLR) for each speaker, then redecode with the conventional ML acoustic model and bigram LM.
3. Estimate VTLN, CMLLR, and maximum likelihood linear regression (MLLR) parameters for each speaker, then redecode with the conventional model and bigram LM.
4. Estimate VTLN, CMLLR, MLLR parameters for each speaker, then redecode with the ML-SA T model and bigram LM.
Table 2: Word error rates for every beamforming algorithm after every decoding passes, as well as the close-talking microphone (CTM).

<table>
<thead>
<tr>
<th>Beamforming Algorithm</th>
<th>Pass (%WER)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Delay &amp; Sum</td>
<td>85.1</td>
</tr>
<tr>
<td>MMI: Gaussian</td>
<td>79.7</td>
</tr>
<tr>
<td>MMI: Laplace</td>
<td>81.1</td>
</tr>
<tr>
<td>MMI: $K_0$</td>
<td>78.0</td>
</tr>
<tr>
<td>MMI: $\Gamma$</td>
<td>80.3</td>
</tr>
<tr>
<td>CTM</td>
<td>37.1</td>
</tr>
</tbody>
</table>
Table 3: Network sizes for the shrunken and full trigram.

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Bigrams</th>
<th>Trigrams</th>
<th>Nodes</th>
<th>Arcs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shrunken Bigram</td>
<td>323,703</td>
<td>0</td>
<td>4,974,987</td>
<td>16,672,798</td>
</tr>
<tr>
<td>Full Bigram</td>
<td>835,688</td>
<td>0</td>
<td>4,366,485</td>
<td>10,639,728</td>
</tr>
<tr>
<td>Shrunken Trigram</td>
<td>431,131</td>
<td>435,420</td>
<td>14,187,005</td>
<td>32,533,593</td>
</tr>
<tr>
<td>Full Trigram</td>
<td>1,639,687</td>
<td>2,684,151</td>
<td>49,082,515</td>
<td>114,304,406</td>
</tr>
</tbody>
</table>

Table 4: ASR results on the SSC development data.

<table>
<thead>
<tr>
<th>Language Model/Pass</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>shrunken bigram</td>
<td>85.7</td>
<td>64.8</td>
<td>53.8</td>
<td>52.5</td>
</tr>
<tr>
<td>full bigram</td>
<td>85.7</td>
<td>65.5</td>
<td>53.8</td>
<td>52.4</td>
</tr>
<tr>
<td>shrunken trigram</td>
<td>86.1</td>
<td>61.7</td>
<td>49.8</td>
<td>47.7</td>
</tr>
<tr>
<td>full trigram (JHLM)</td>
<td>88.3</td>
<td>61.0</td>
<td>48.9</td>
<td>47.0</td>
</tr>
</tbody>
</table>
Filter Bank Comparisons

- As beamforming is performed in the frequency or subband domain, a *filter bank* is required.
- Our initial experiments we conducted with a *perfect reconstruction* (PR) filter bank (Vaidyanathan, 1993, §8).
- As the PR filter bank is *not* optimal for adaptive filtering applications, we also considered the de Haan (2003) filter bank.
- In addition we tried a simple FFT as well as a new filter bank based on a *Nyquist*($M$) *constraint*. 
Figure 7: Frequency response of proposed composite analysis-synthesis filter bank prototypes.
Effect of Filter Bank

- We tested each different type of filter bank under the Gaussian assumption.
- The results were very *counterintuitive*.

Table 5: ASR results on the SSC evaluation data.

<table>
<thead>
<tr>
<th>Filter Bank</th>
<th>Word Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Nyquist</td>
<td>89.3</td>
</tr>
<tr>
<td>Normal FFT</td>
<td>88.5</td>
</tr>
<tr>
<td>de Haan</td>
<td>88.9</td>
</tr>
<tr>
<td>PR</td>
<td>87.7</td>
</tr>
</tbody>
</table>
Conclusions

• Speech Separation Challenge: Great Task! We learned many new things.
• What worked:
  – Speaker tracking and segmentation with the JPDAF;
  – Minimum mutual information beamforming;
  – Super-gaussian pdfs;
  – Speaker adaptation on the far-field data based on word lattices;
  – Improved techniques for constructing the recognition network.
• What did not work (this time):
  – Non-PR filter banks (very strange);
  – Super-duper pdf (but it’s out there);
  – Discriminative HMM training.
• There are still many unanswered questions.
• Let’s do it again next year.
References


