We have validated the effectiveness of techniques below: Speech enhancement: Single-channel dereverberation method\(^1\) - reverberation time (RT) estimation Eight-channel beam-forming with direction of arrival estimation\(^2\) ASR using Kaldi toolkit\(^3\): Feature transformation and speaker adaptation - LDA, MLLT, basis FMLLR Discriminative training and discriminative feature transformation - boosted MMI and feature-space boosted MMI - Deep neural networks ASR System combination using ROVER\(^4\): Discriminative training for system combination - dual system approach\(^5\) - black-box optimization of ROVER parameters\(^6\)

Speech Enhancement Part

DS beamformer with direction of arrival estimation

\[
\hat{x}(t) = \sum_{j=1}^{J} w_j(t) y_j(t) = \sum_{j=1}^{J} \alpha_j(t) y_j(t)
\]

Estimation of direction of arrival

\[
\theta(t) = \arctan\left(\frac{y_2(t)}{y_1(t)}\right)
\]

\[
\theta(t) = \arctan\left(\frac{y_2(t)}{y_1(t)}\right)
\]

SS based derev. with RT estimation

Reverberation model

\[
x(t) = \sum_{\ell=0}^{\infty} h_\ell \cos(2\pi f_\ell t + \varphi_\ell)
\]

Instantaneous mixture model

\[
x(t) = \sum_{\ell=0}^{\infty} h_\ell \cos(2\pi f_\ell t + \varphi_\ell)
\]

Approx. dereverberation formula

\[
x(t) = \frac{\sum_{\ell=0}^{\infty} h_\ell \cos(2\pi f_\ell t + \varphi_\ell)}{\sum_{\ell=0}^{\infty} h_\ell^2}
\]

Polack model

\[
\phi(t) = \frac{1}{5} \sum_{k=0}^{4} \phi(k)
\]

To estimate RT, floored ratio of Eq.(1) is calculated for assumed RT

Two observations:

- \( r \) increases with \( T_a \) (assumed RT)
- \( r \) increases with \( T_a \) (actual RT)

Using these RT, can be estimated from the floored ratio

\[
\hat{T}_a = \frac{\sum_{k=0}^{4} \phi(k)}{5}
\]

ASR part

MMI discriminative training of acoustic models

MMI objective function to optimize \( \lambda \) and \( \nu_i \)

\[
J_{\text{MMI}}(\lambda, \nu_i) = \sum_{t=1}^{T} \sum_{c=1}^{C} P(c|\hat{x}(t)) \log \left( \frac{P(c|\hat{x}(t))}{\sum_{c'=1}^{C} P(c'|\hat{x}(t))} \right)
\]

Polack model

\[
\phi(t) = \frac{1}{5} \sum_{k=0}^{4} \phi(k)
\]

\[
\phi(t) = \frac{1}{5} \sum_{k=0}^{4} \phi(k)
\]

Discriminative training for system combination

Discriminative training principle: MT between ref., 1-best of base system, and hypotheses of comp. system

Proposed objective function\(^7\):

\[
J_{\text{MMI}}(\lambda, \nu_i) = \sum_{t=1}^{T} \sum_{c=1}^{C} P(c|\hat{x}(t)) \log \left( \frac{P(c|\hat{x}(t))}{\sum_{c'=1}^{C} P(c'|\hat{x}(t))} \right) + \lambda \left( \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{C} \sum_{c=1}^{C} P(c|\hat{x}(t)) \right) \right)
\]

\[
J_{\text{MMI}}(\lambda, \nu_i) = \sum_{t=1}^{T} \sum_{c=1}^{C} P(c|\hat{x}(t)) \log \left( \frac{P(c|\hat{x}(t))}{\sum_{c'=1}^{C} P(c'|\hat{x}(t))} \right) + \lambda \left( \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{C} \sum_{c=1}^{C} P(c|\hat{x}(t)) \right) \right)
\]

System overview

Experiments

Task description and setup

A middle-size vocabulary continuous speech recognition task

8 different reverberant environments: -3 rooms with near/far mic settings for SPMulated data -1 room with near/far mic settings for REAL data with noise

Speech enhancement

derev improves the performance

BF improves it further

Feature transformation and discriminative methods

LDA improves the performance

due to the use of long context basis FMLLR is effective

DMM is effective

SIT is unstable

Subspace GMM and DNN

bMMI is effective

Best performer is different for each environment

Complementary systems

Performance is moderate

Output tendencities are different

System combination

System combination improves the accuracies for all the cases

Proposed method is effective for all the environment

Combination of different types of systems is effective

<table>
<thead>
<tr>
<th>Method</th>
<th>WER</th>
<th>WER</th>
<th>WER</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIM</td>
<td>10.8</td>
<td>11.6</td>
<td>20</td>
<td>18.7</td>
</tr>
<tr>
<td>REAL</td>
<td>15.5</td>
<td>21.6</td>
<td>30.0</td>
<td>23.3</td>
</tr>
<tr>
<td>SIM REAL</td>
<td>10.8</td>
<td>11.6</td>
<td>20</td>
<td>18.7</td>
</tr>
</tbody>
</table>

Black box optimization on ROVER parameters

Selection of combined system

ROVER parameter

WER improves monotonically

100 iterations are enough

Evaluation set

<table>
<thead>
<tr>
<th>Method</th>
<th>WER</th>
<th>WER</th>
<th>WER</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIM</td>
<td>10.8</td>
<td>11.6</td>
<td>20</td>
<td>18.7</td>
</tr>
<tr>
<td>REAL</td>
<td>15.5</td>
<td>21.6</td>
<td>30.0</td>
<td>23.3</td>
</tr>
<tr>
<td>SIM REAL</td>
<td>10.8</td>
<td>11.6</td>
<td>20</td>
<td>18.7</td>
</tr>
</tbody>
</table>

---


1) Dereverberation technique: Single-channel dereverberation method (Tachioka et al., 2013), Eight-channel beamformer with direction of arrival estimation (Tachioka et al., 2015) and Multi-Channel (M. S. Shumway and D. T. Williams, 2004).