Multimodal speech recognition and enhancement

Dorothea Kolossa & Steffen Zeiler
Cognitive Signal Processing Group /
Kavli Institute for Theoretical Physics

Ruhr-Universität Bochum
Fakultät für Elektrotechnik und Informationstechnik
Cognitive Signal Processing Group

Research Associate:
Steffen Zeiler

PhD students:
Benedikt Bönninghoff
Hendrik Meutzner
Christopher Schymura
Mahdie Karbasi
Dennis Orth
Lea Schönherr

Former Scientists
Ahmed Hussen Abdelaziz
Sebastian Gergen

Student Research Assistants:
Juan Rios Grajales
Jan Hünnemeyer
Tobias Isenberg
Diana Castano Marin

with many thanks to Robert Nickel, Ning Ma, Guy Brown, Ramon Fernandez Astudillo
Audiovisual speech perception

Human speech perception utilizes video information

One piece of evidence:

is the “McGurk Effect” [McGurk1976]
Audiovisual speech perception

Early, extensive intelligibility tests:

Introduction & Overview

Idea:
Integrate video information in machine listening

Useful for two purposes:

• Multimodal speech recognition
• Audiovisual Speech Enhancement (to improve intelligibility)
Introduction & Overview

Outline:

- Audiovisual speech recognition
  - Methods and models for audiovisual integration
  - Stream weighting
- Audiovisual Speech Enhancement
- Conclusions and perspectives
Audiovisual Speech Recognition

Levels of integration
Levels of integration

**Graphical models** [Whittaker1990, Jordan1999]

Describe statistical dependencies of multiple variables

“Visible”/”Measureable” variables are often denoted by shaded circles

\[ O_t \]
Levels of integration

**Graphical models** [Whittaker1990, Jordan1999]

Describe statistical dependencies of multiple variables
“Hidden” variables are often denoted by empty circles

\[ q_t \]
Levels of integration

**Graphical models** [Whittaker1990, Jordan1999]

Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:

\[ q_t \rightarrow o_t \]
Levels of integration

*Graphical models* [Whittaker1990, Jordan1999]

Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:
Levels of integration

*Graphical models* [Whittaker1990, Jordan1999]

Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:
Levels of integration

*Graphical models* [Whittaker1990, Jordan1999]

Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:

\[
\begin{align*}
q_{t-1} & \quad \rightarrow \quad q_t \\
O_{t-1} & \quad \rightarrow \quad O_t \\
q_t & \quad \rightarrow \quad q_{t+1}
\end{align*}
\]
Levels of integration

*Graphical models* [Whittaker1990, Jordan1999]

Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:
Levels of integration

**Graphical models** [Whittaker1990, Jordan1999]

Specifically in “Bayesian Networks”, direct statistical dependencies are denoted by arrows:
Levels of integration

Graphical models  [Whittaker1990, Jordan1999]

Indirect statistical dependencies are not:

This model encodes the dependency assumptions of (1st order) Hidden Markov Models in speech recognition.
Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

\[ \ldots \quad q_{t-1} \quad q_t \quad q_{t+1} \quad \ldots \]

\[ \ldots \quad o_{t-1} \quad o_t \quad o_{t+1} \quad \ldots \]
Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

**Graphical Model of Audiovisual Speech Recognition with Feature Fusion**

![Diagram showing the graphical model of audiovisual speech recognition with feature fusion. The diagram includes nodes and arrows representing speech features (Audio features) and video features, leading to recognized text.]

**System

“Standard” speech recognition setup**
Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]

*Graphical model for audiovisual speech recognition with late integration*
Levels of integration

Multimodal speech recognition can take place at three levels
a) Early integration = Feature fusion
b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]

System for late integration
Two “standard” ASR systems, whose outputs are later combined
Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion

b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]

c) Intermediate integration = within the classifier/DNN

*Graphical model, intermediate integration*
Levels of integration

Multimodal speech recognition can take place at three levels

a) Early integration = Feature fusion
b) Late integration = combine multiple recognition results (ROVER) [Fiscus1997]
c) Intermediate integration = within the classifier/DNN

Graphical Model

Most successful model in wide range of experiments [Nefian2002a, Zeiler 2016, Receveur2016]
Coupled Hidden Markov Models

*An example of intermediate integration*
Coupled HMMs for asynchronous Audio- & Video Streams

Cartesian product of audio and video. HMM can cope with time-varying delay of audio and video.

\[ q_a = 1 \quad 2 \quad \ldots \quad m \]

\[ q_v = 1 \quad \ldots \quad n \]

HMM for Stream 1

Coupled HMM

Stream 2

[Luettin2001]
Audiovisual speech recognition

Audio Only: Lay blue in o 8 now.
Video Only: Set blue by h 1 soon.
AVSR Result: Bin blue at s 1 soon.

bin blue at s one soon

Coupled HMMs for asynchronous Audio- & Video Streams

Audiovisual Speech Recognition

- Audiovisual recognition using coupled HMMs always outperforms audio-only and video-only ASR when stream weights (more later!) are appropriately set.

<table>
<thead>
<tr>
<th>SNR</th>
<th>-6dB</th>
<th>-3dB</th>
<th>0dB</th>
<th>3dB</th>
<th>6dB</th>
<th>9dB</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video</td>
<td>27.8</td>
<td>27.8</td>
<td>27.8</td>
<td>27.8</td>
<td>27.8</td>
<td>27.8</td>
<td>27.8</td>
</tr>
<tr>
<td>Audio</td>
<td>27.9</td>
<td>23.0</td>
<td>18.1</td>
<td>15.4</td>
<td>12.8</td>
<td>10.4</td>
<td>17.9</td>
</tr>
<tr>
<td>Audiovisual CHMM</td>
<td>17.2</td>
<td>14.1</td>
<td>12.0</td>
<td>10.1</td>
<td>9.0</td>
<td>7.7</td>
<td>11.7</td>
</tr>
</tbody>
</table>

Coupled HMMs for asynchronous Audio- & Video Streams

Audiovisual Speech Recognition

- Audiovisual recognition using coupled HMMs always outperforms audio-only and video-only ASR when stream weights (more later!) are appropriately set.
- Best results are achieved with noise-adaptive LDA + ground truth uncertainties

<table>
<thead>
<tr>
<th>Keyword Error Rates (%) on CHiME 2 Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR</td>
</tr>
<tr>
<td>Video</td>
</tr>
<tr>
<td>Audio</td>
</tr>
<tr>
<td>Audio + NALDA</td>
</tr>
<tr>
<td>Audiovisual CHMM</td>
</tr>
<tr>
<td>Audiovisual CHMM + NALDA</td>
</tr>
</tbody>
</table>

Coupled HMMs for asynchronous Audio- & Video Streams

Audiovisual speech recognition always outperforms audio-only and video-only ASR when stream weights are appropriately set, even under complete mismatch, here, training on clean & testing on noisy data.
Stream weighting

*Can’t live with it, can’t seem to live without it...*
Stream Weighting for Audiovisual Speech Recognition

Emission probabilities of coupled HMM:

\[
p(o \mid q) = b_a (o_a \mid q_a)\lambda \cdot b_v (o_v \mid q_v)^{1-\lambda}
\]

The \( b_{a/v} (o_{a/v} \mid q_{a/v})^\lambda \) are observation likelihoods, \( \lambda \) the stream weight.

Stream weighting not only applicable in coupled model but in all early and intermediate integration schemes including deep neural network-based ones.

Question: Is this really necessary?
   Most recently, e.g. [Ninomiya2015, Ngiam2011, Tamura2015, Noda2015, Meutzner2017]

Question 2: If yes, how?
Stream Weighting for Audiovisual Speech Recognition

Idea of dynamic stream weighting system

Train neural network or logistic regression function to map some reliability features onto stream weights, using optimal dynamic stream weights as training targets.

During test time, this trained regression model or DNN will then map reliability measures (frame by frame) onto frame-wise stream weights

Reliability measure features

- Estimated observation uncertainties
- Estimated SNR
- Soft and hard VAD cues based on IMCRA noise estimation
- Dispersion and entropy of audio and video HMM

Coupled HMMs for asynchronous Audio- & Videostreams

Results of dynamic stream weighting, comparing three strategies

- Equal weights, $\lambda = 0.5$ ("Bayes Fusion")
- Exponential Function [Estellers 2012]
- MLP: Dynamic stream weight estimation using multiple reliability features

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>SNR [dB]</th>
<th>Audio only</th>
<th>Video only</th>
<th>Audio-visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Babble</td>
<td>15</td>
<td>0.8516</td>
<td></td>
<td>0.9401</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.6853</td>
<td></td>
<td>0.8840</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.4675</td>
<td></td>
<td>0.7523</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.3065</td>
<td></td>
<td>0.6040</td>
</tr>
<tr>
<td>White</td>
<td>15</td>
<td>0.8399</td>
<td>0.8476</td>
<td>0.9385</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.6819</td>
<td></td>
<td>0.8854</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.5133</td>
<td></td>
<td>0.8130</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.3701</td>
<td></td>
<td>0.7296</td>
</tr>
<tr>
<td>Clean</td>
<td>-</td>
<td>0.9886</td>
<td></td>
<td>0.9856</td>
</tr>
<tr>
<td>Avg.</td>
<td>-</td>
<td>0.6339</td>
<td></td>
<td>0.8369</td>
</tr>
</tbody>
</table>

Stream Weighting in Deep Neural Networks
Stream Weighting in Deep Neural Networks

**Fundamental question:**
Shouldn’t we just train one large neural network?
Two considered alternatives

1) **Concatenation of uncertainties**
Train one large network with uncertainties as an additional input.

2) **Explicit stream weighting**
Train two networks and fuse their posterior probabilities according to

$$
\log p(o^{AV} | q) = \gamma \log(b^A (o^A | q)) + (1 - \gamma) \log(b^V (o^V | q))
$$
Stream Weighting in Deep Neural Networks

Evaluation:
Again, on the CHiME 2 data, as above.

Kaldi recipe based on Wall-Street-Journal training scripts*, using

1) Concatenation of uncertainties
2) Explicit stream weighting

https://github.com/hmeutzner/kaldi-avsr

*Hybrid system, so the DNN estimates state posteriors. Trained starting by GMM/HMM training, including LDA, fMLLR & speaker-adaptive training and continuing onto DNN/HMM. For this purpose, we use a topology with 11 frames of context, for 440d input, 6 hidden layers with 2048 neurons each, 1453 neurons in softmax output layer. RBM layer-wise pre-training is followed by minimum-cross-entropy training, followed by minimum Bayes risk fine-tuning.
Stream Weighting in Deep Neural Networks

Concatenation of uncertainties

<table>
<thead>
<tr>
<th>Features</th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter-bank</td>
<td>DCT</td>
<td>13.01</td>
<td>11.90</td>
<td>11.14</td>
<td>9.95</td>
<td>8.33</td>
<td>9.10</td>
</tr>
<tr>
<td>Filter-bank</td>
<td>DCT</td>
<td>Uncertainty</td>
<td>13.86</td>
<td>12.59</td>
<td>10.80</td>
<td>8.93</td>
<td><strong>7.40</strong></td>
</tr>
<tr>
<td>MFCC</td>
<td>DCT</td>
<td>17.94</td>
<td>16.67</td>
<td>15.73</td>
<td>14.63</td>
<td>13.27</td>
<td>12.33</td>
</tr>
<tr>
<td>MFCC</td>
<td>DCT</td>
<td>Uncertainty</td>
<td>14.71</td>
<td>13.18</td>
<td>11.39</td>
<td>10.03</td>
<td>9.18</td>
</tr>
<tr>
<td>Rate-map</td>
<td>DCT</td>
<td><strong>11.99</strong></td>
<td><strong>11.48</strong></td>
<td><strong>10.29</strong></td>
<td>8.08</td>
<td>7.91</td>
<td>8.08</td>
</tr>
<tr>
<td>Rate-map</td>
<td>DCT</td>
<td>Uncertainty</td>
<td>14.29</td>
<td>12.16</td>
<td>10.63</td>
<td><strong>7.65</strong></td>
<td>7.65</td>
</tr>
</tbody>
</table>

does help

Explicit stream weighting

<table>
<thead>
<tr>
<th>Features</th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate-map</td>
<td>28.36</td>
<td>23.45</td>
<td>15.17</td>
<td>11.55</td>
<td>7.93</td>
<td>6.72</td>
<td>15.53</td>
</tr>
<tr>
<td>DCT</td>
<td>27.50</td>
<td>27.50</td>
<td>27.50</td>
<td>27.50</td>
<td>27.50</td>
<td>27.50</td>
<td>27.50</td>
</tr>
<tr>
<td>Rate-map</td>
<td>DCT</td>
<td>13.79</td>
<td>12.76</td>
<td>10.00</td>
<td><strong>8.88</strong></td>
<td>8.88</td>
<td>8.79</td>
</tr>
<tr>
<td>Rate-map</td>
<td>DCT</td>
<td>Uncertainty</td>
<td>14.91</td>
<td>13.88</td>
<td>11.21</td>
<td><strong>8.88</strong></td>
<td><strong>8.79</strong></td>
</tr>
<tr>
<td>(\lambda) set per sentence, oracle-SNR-based</td>
<td>27.67</td>
<td>18.10</td>
<td>11.12</td>
<td>6.64</td>
<td>5.86</td>
<td>6.72</td>
<td>12.69</td>
</tr>
<tr>
<td>(\lambda) set per frame, uncertainty-based</td>
<td><strong>13.28</strong></td>
<td><strong>11.12</strong></td>
<td><strong>8.10</strong></td>
<td><strong>6.81</strong></td>
<td><strong>5.60</strong></td>
<td><strong>4.22</strong></td>
<td><strong>8.19</strong></td>
</tr>
</tbody>
</table>

may or may not help
Stream Weighting in Deep Neural Networks

Intermediate Conclusion

With appropriate stream weighting, audiovisual recognition can reliably give accuracies that are equal to or better than the single best modality.

Stream weighting can be guided by reliability measures composed of recognition confidence measures and observation uncertainties. The composition is better than the single best measure.

Such stream weighting also appears to be helpful in the fusion of audiovisual multi-stream DNNs.

Next question

How can such audiovisual recognition systems benefit speech enhancement (e.g. for extremely noisy environments)?
...and moving on to the second part:

*Audiovisual speech enhancement*
...and many thanks for your attention!
References


References


Coupled HMMs for asynchronous Audio- & Video Streams

Block diagram of dynamic stream weighting system

Train neural network or logistic regression function using oracle dynamic stream weights (ODSWs) as training targets.

**Reliability measure features**
- Estimated observation uncertainties
- Estimated SNR
- Soft and hard VAD cues based on IMCRA noise estimation
- Dispersion and entropy of audio and video HMM

Introducing the Turbo-Twin-HMM for Audio-Visual Speech Enhancement

Steffen Zeiler, Hendrik Meutzner, Ahmed Hussen Abdelaziz, Dorothea Kolossa
June 28th 2017
How can we recover speech from noise?
How can we recover speech from noise?

clean: “bin white with L3 again”
How can we recover speech from noise?

clean: "bin white with l3 again"

noisy: "bin white with l3 again"
How can we recover speech from noise?
How can we recover speech from noise?

**Intro**
TurboTwin-HMM
Twin-HMM
Turbo Decoding
Results
Conclusion

Steffen Zeiler
System Overview

\[
(1 - g_t) \hat{x}_t^m + g_t \hat{x}_t^h
\]

\[
\hat{x}_t^a e^{i\phi_t}
\]

% Intro TurboTwin-HMM Twin-HMM Turbo Decoding Results Conclusion

Steffen Zeiler

2 / 9
The Twin-HMM

features → recognition model → synthesis output density functions → output

features:
- AP: MMSE estimate of the clean speech amplitude spectrum
  \[
  \hat{x}(t) = E(x_t|o_t) = \sum_{i=1}^N p(q_t=i|o_t) E(x_t|q_t=i)
  \]
- BP: use the most probable state \(i^*_t\) in each frame
  \[
  \hat{x}(t) = E(x_t|q_t=i^*_t)
  \]
The Twin-HMM

**AP:** MMSE estimate of the clean speech amplitude spectrum

\[ \hat{x}(t) = \mathbb{E}(x_t|o) = \sum_{i=1}^{N} p(q_t = i|o) \mathbb{E}(x_t|q_t = i) \]
The Twin-HMM

**AP**: MMSE estimate of the clean speech amplitude spectrum

\[ \hat{x}(t) = E(x_t|o) = \sum_{i=1}^{N} p(q_t = i|o) E(x_t|q_t = i) \]

**BP**: use the most probable state \( i^*_t \) in each frame \( t \)

\[ \hat{x}(t) = E(x_t|q_t = i^*_t) \]
Turbo Decoding

\[ \tilde{b}_a (o_a | q_a) = b_a (o_a | q_a) \cdot g_a (q_a) \lambda_T \lambda_P, \]

\[ \tilde{b}_v (o_v | q_v) = b_v (o_v | q_v) \cdot g_v (q_v) (1 - \lambda_T) \lambda_P. \]

---

Shivappa, Rao, Trivedi: *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition*, ICASSP 2008
Turbo Decoding

\[ \tilde{b}_a (o_a | q_a) = b_a (o_a | q_a) \cdot g_a (q_a) \lambda_T \lambda_P, \]
\[ \tilde{b}_v (o_v | q_v) = b_v (o_v | q_v) \cdot g_v (q_v) (1 - \lambda_T) \lambda_P. \]

Shivappa, Rao, Trivedi: *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition*, ICASSP 2008
Turbo Decoding

1Shivappa, Rao, Trivedi: Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition, ICASSP 2008
Turbo Decoding

![Diagram of Turbo Decoding]

**likelihood modification:**

\[
\tilde{b}_a(o_a | q_a) = b_a(o_a | q_a) \cdot g_a(q_a)^{\lambda_T \lambda_P},
\]
\[
\tilde{b}_v(o_v | q_v) = b_v(o_v | q_v) \cdot g_v(q_v)^{(1-\lambda_T) \lambda_P}
\]

---

1Shivappa, Rao, Trivedi: *Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition*, ICASSP 2008
# Instrumental Measures

<table>
<thead>
<tr>
<th>SNR</th>
<th>PESQ&lt;sup&gt;2&lt;/sup&gt;</th>
<th>STOI&lt;sup&gt;3&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 dB</td>
<td>-3 dB</td>
</tr>
<tr>
<td></td>
<td>noisy</td>
<td>1.95</td>
</tr>
<tr>
<td></td>
<td>log-MMSE</td>
<td>1.90</td>
</tr>
<tr>
<td>E1AP</td>
<td>2.11</td>
<td>2.02</td>
</tr>
<tr>
<td>E1BP</td>
<td>2.02</td>
<td>1.92</td>
</tr>
<tr>
<td>E2AP</td>
<td>2.08</td>
<td>2.01</td>
</tr>
<tr>
<td>E2BP</td>
<td>1.99</td>
<td>1.91</td>
</tr>
</tbody>
</table>

588 files per SNR

**recognizer features**

- **E1** minimize synthesis distortions
- **E2** optimize recognition results

**clean speech estimation**

- **AP** all path synthesis
- **BP** best path synthesis

<sup>2</sup>PESQ: Perceptual Evaluation of Speech Quality,  
<sup>3</sup>STOI: Short Time Objective Intelligibility
Listening Tests

- large-scale listening experiment (CrowdFlower)
- 690 individual participants, 27,118 transcribed utterances
- quality control to identify cheaters or language deficits

<table>
<thead>
<tr>
<th>SNR [dB]</th>
<th>Noisy</th>
<th>log-MMSE</th>
<th>E1AP</th>
<th>E2AP</th>
<th>E2BP</th>
</tr>
</thead>
<tbody>
<tr>
<td>-9</td>
<td>48.65</td>
<td>69.56</td>
<td>73.43</td>
<td>74.82</td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td>60.23</td>
<td>53.88</td>
<td>72.72</td>
<td>77.28</td>
<td>79.35</td>
</tr>
<tr>
<td>-3</td>
<td>70.47</td>
<td>61.41</td>
<td>75.82</td>
<td>81.58</td>
<td>81.43</td>
</tr>
<tr>
<td>0</td>
<td>77.24</td>
<td>71.08</td>
<td>78.94</td>
<td>82.65</td>
<td>84.64</td>
</tr>
</tbody>
</table>

260 utterances per column
Conclusion

- video assisted single channel speech enhancement works

- our best system [E2BP] improves word accuracy of human listeners for the GRID task
  
  - from 48.6% to 74.8% at -9dB
  
  - from 77.2% to 84.6% at 0dB SNR

- predictions of reference-based objective speech intelligibility measures are *unreliable* for non-linearly processed speech
Perspectives

- better intelligibility estimators are needed - we are considering speech-recognition-based measures

- AV speech enhancement needs to be extended to open vocabularies and arbitrary recording conditions (taking video reliability information into account)

- for this purpose, and others, we are working on large-vocabulary AV speech recognition, combining our more general topologies with TensorFlow training of convolutive/recurrent nets
Thank you
## Recognition Accuracy for variants E1 and E2

<table>
<thead>
<tr>
<th>Method</th>
<th>-9 dB</th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>$\infty$ dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E1$</td>
<td>87.95%</td>
<td>90.27%</td>
<td>91.13%</td>
<td>93.38%</td>
<td>97.15%</td>
</tr>
<tr>
<td>$E2$</td>
<td>89.67%</td>
<td>91.81%</td>
<td>93.98%</td>
<td>95.34%</td>
<td>98.18%</td>
</tr>
</tbody>
</table>

$E1$ : optimized for minimal distortion during synthesis

$E2$ : optimized for best recognition results
## Listening Test Recognition Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Noisy</th>
<th>log-MMSE</th>
<th>E1AP</th>
<th>E2AP</th>
<th>E2BP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>64.27%</td>
<td>57.09%</td>
<td>74.30%</td>
<td>78.78%</td>
<td>80.10%</td>
</tr>
</tbody>
</table>

- average word accuracy over all SNRs
- Each score is based on 1037 utterances