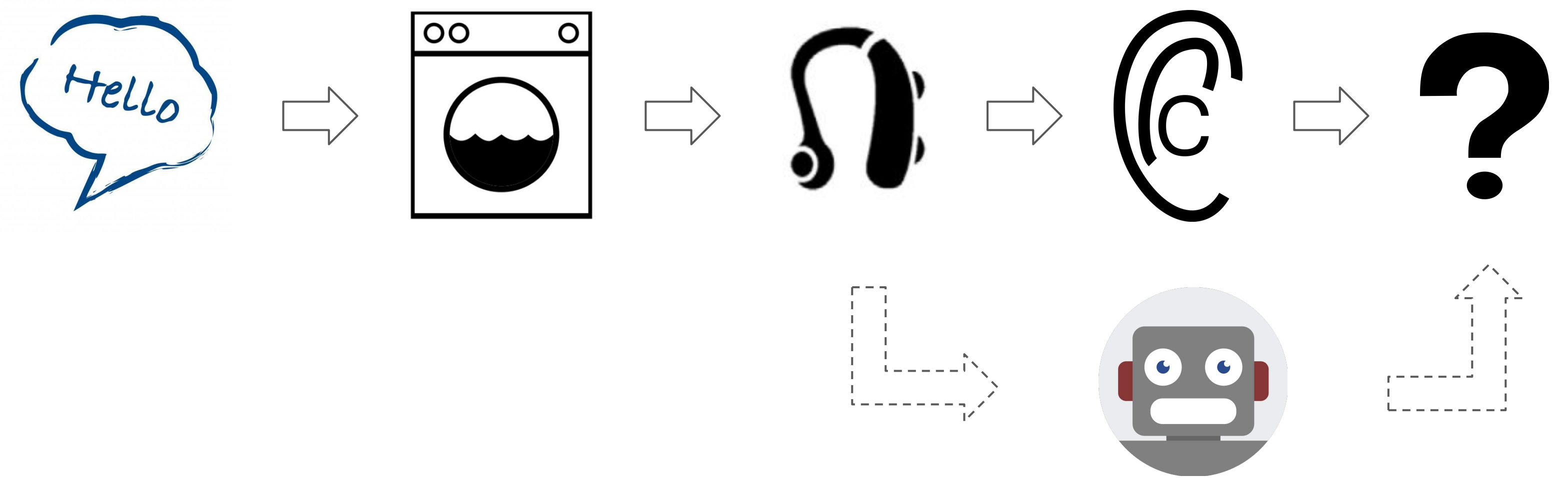
Sheffield Systems E029 & E032 for the First Round Clarity Prediction Challenge

Zehai Tu, Ning Ma, Jon Barker
Department of Computer Science, University of Sheffield, UK

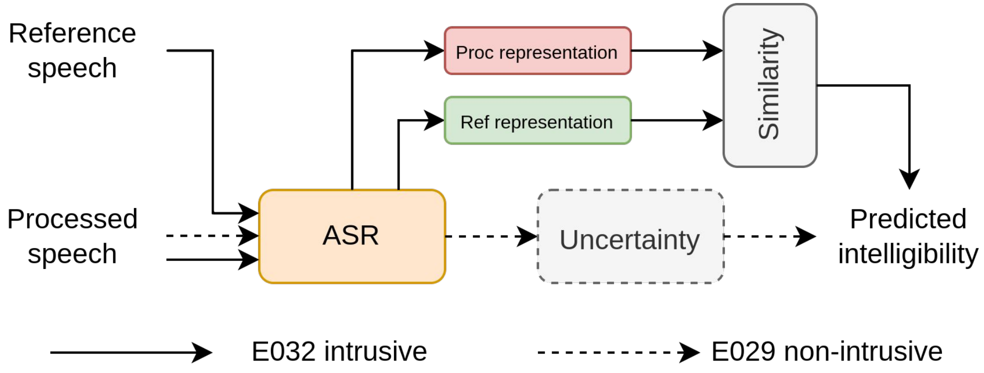
Background



How an automatic speech recogniser (ASR) can be used to predict the intelligibility?

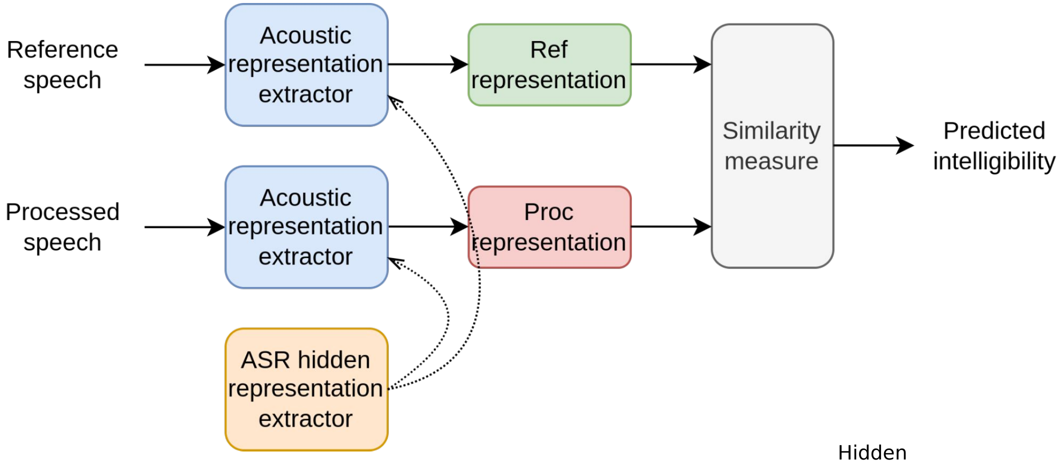
Overview

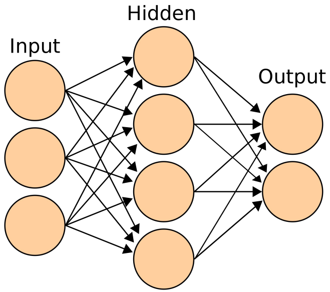
- E032 (intrusive): *Exploiting Hidden Representations from a DNN-based Speech Recogniser for Speech Intelligibility Prediction in Hearing-impaired Listeners*
- E029 (non-intrusive): *Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction*



E032 overview

- Method
  - DNN-based ASR model
  - Hidden representations
  - Similarity computation
- Experimental setup
- Results
- Conclusions





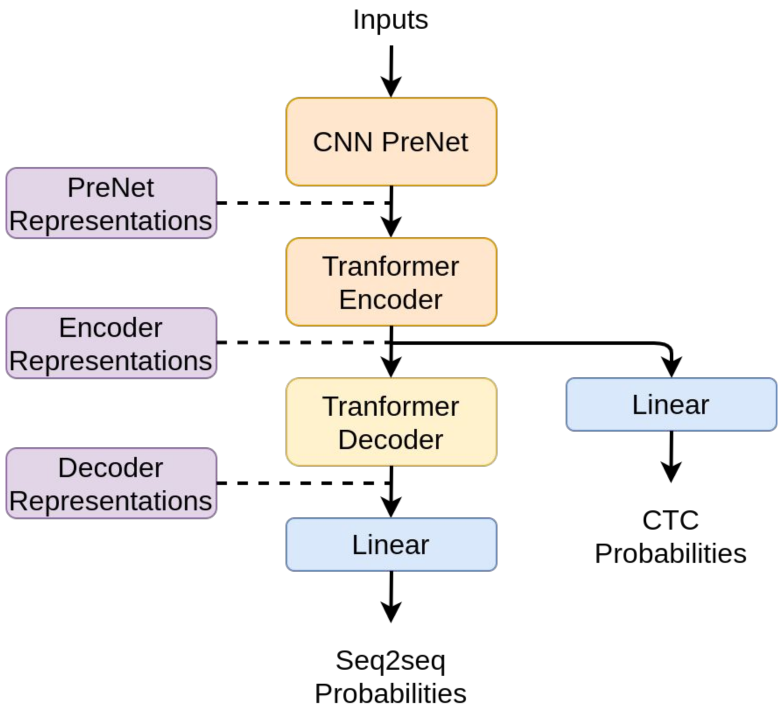
ASR model

- Transformer-based end-to-end ASR model
- Hybrid training: Connectionist Temporal Classification + Attention-based Sequence-to-sequence
- SpeechBrain LibriSpeech ASR transformer recipe[1]
- Fine-tuned from pretrained LibriSpeech (960h) model, i.e., strong knowledge on clean speech recognition

[1] https://github.com/speechbrain/speechbrain/tree/develop/recipes/LibriSpeech/ASR/transformer

Hidden representations

- PreNet representations: low-level acoustic features
- Encoder representations: high-level acoustic features
- Decoder representations: features with language model knowledge



Similarity computation

Given two hidden vectors of the reference and processed speech:

$$h, \hat{h} \in \mathcal{R}^d$$

The similarities of the PreNet and Encoder representations:

$$sim(H^{bi}, \hat{H}^{bi}) = \frac{1}{T}\sum_{t=1}^{T} \max\left\{\rho_t^{ll}, \rho_t^{lr}, \rho_t^{rl}, \rho_t^{rr}\right\}$$

The similarity of the Decoder representations, fast dynamic time warping is used:

$$sim(H_w, \hat{H}_w) = \frac{1}{T_w}\sum_{t=1}^{T_w} \cos(H_w(t), \hat{H}_w(t))$$

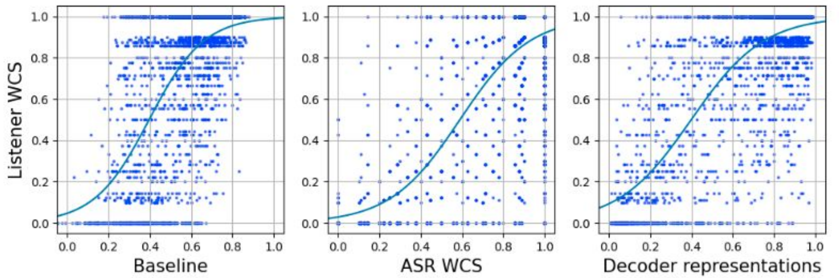
$$sim(H^{bi}, \hat{H}^{bi}) = \max\left\{sim(H_w^l, \hat{H}_w^l), sim(H_w^l, \hat{H}_w^r), sim(H_w^r, \hat{H}_w^l), sim(H_w^r, \hat{H}_w^r)\right\}$$

Experimental Setup

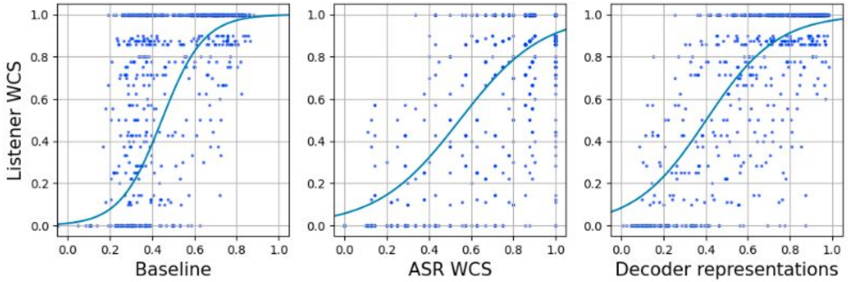
- Data split: 70% of CPC1_train_data is used as training set, 30% is used as dev set
- ASR training:
  - Cambridge MSBG hearing loss model as front-end to simulate hearing losses
  - Librispeech train-clean-100 + CEC1 training noise (CLS) for 10 epochs
  - CPC1 training set for 10 epochs
- Evaluation:
  - Root mean square error (RMS), normalised cross-correlation (NCC), Kendall’s Tau coefficient (KT)
  - A logistic function fitting on the dev set
- Baselines:
  - MSBG + MBSTOI (CPC1 baseline)
  - ASR word correctness score (WCS)

E032 Overall results

	RMSE ↓	NCC ↑	KT ↑
Closed-set			
Baseline	0.285	0.621	0.398
ASR WCS	0.250	0.729	**0.523**
PreNet representations	0.347	0.299	0.182
Encoder representations	0.237	0.758	0.487
Decoder representations	**0.231**	**0.773**	0.498
Open-set			
Baseline	0.365	0.529	0.391
ASR WCS	0.250	0.723	**0.534**
PreNet representations	0.356	0.254	0.136
Encoder representations	0.241	0.751	**0.534**
Decoder representations	**0.235**	**0.763**	0.530



(a) *Closed-set*



(b) *Open-set*

E032 results

Further analysis on the CPC1 closed-set:

- Different training data
- With or without using the MSBG hearing loss model

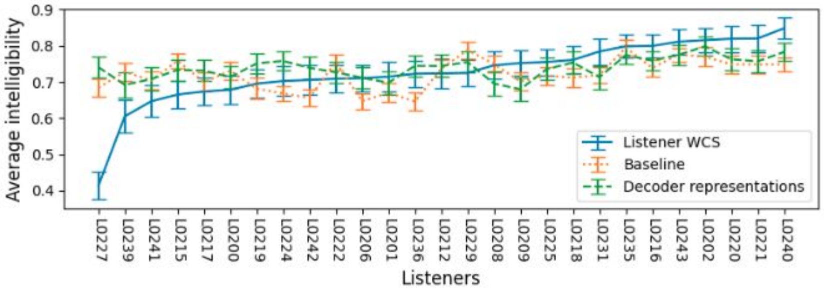
MSBG	Training data	RMSE ↓	NCC ↑	KT ↑
with	LS	0.264	0.692	0.449
	LS+CLS	0.243	0.746	0.464
	LS+CPC1	0.233	0.768	**0.503**
	LS+CLS+CPC1	**0.231**	**0.773**	0.498
w/o	LS+CLS+CPC1	0.234	0.767	0.476

E032 results

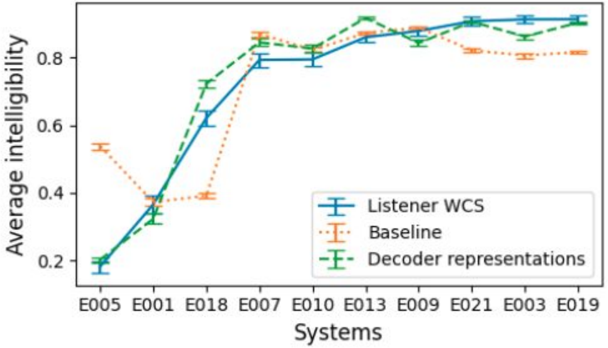
Further analysis on the CPC1 closed-set:

- Listener- and system-wise results

	RMSE ↓	NCC ↑	KT ↑
Listener-wise			
Baseline	0.078	0.414	0.311
Decoder representations	0.078	0.419	0.407
System-wise			
Baseline	0.147	0.798	0.244
Decoder representations	0.048	0.982	0.644



(a) *Listener-wise*



(b) *System-wise*

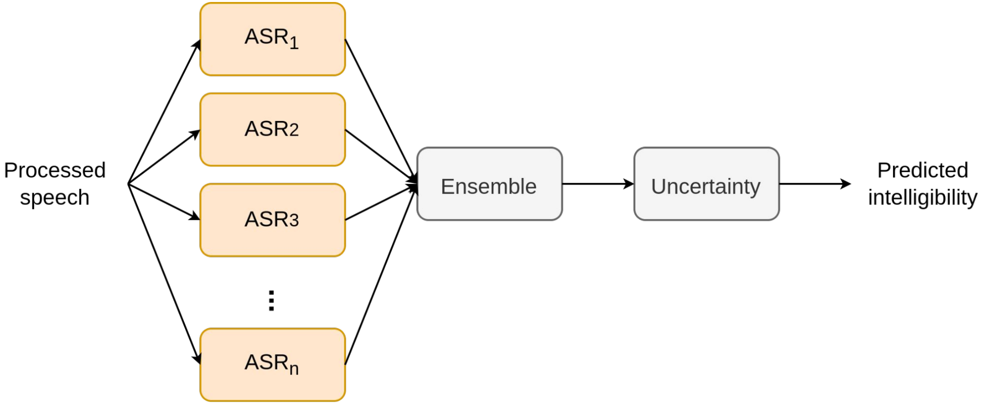
E032 conclusions

- Representations of a state-of-the-art ASR could be better at intelligibility prediction than acoustic representations of MBSTOI.
- Language knowledge matters, as Decoder > Encoder representations.
- ASR recognition results (WCS) might not be the best intelligibility predictor.
- ASR training data matters, meanwhile, ASR trained only by LS can still make good prediction.
- MSBG hearing loss simulation can improve performance.
- If considering only monotonicity, no listener intelligibility label is needed.

Questions on E032?

E029 overview

- Method
  - ASR model (same as E032)
  - **Unsupervised sequence-level uncertainty** estimation
- Experimental setup (same as E032)
- Results
- Conclusions



E029 method

- Why uncertainty?
  - Intelligibility can be characterised as the probability of correct word recognition by human, meanwhile, uncertainty of ASR is also associated with the probability of ASR making correct predictions.
  - Uncertainty avoids cases like correct guess.
- Why unsupervised?
  - No listener intelligibility labels are needed.
- Why sequence-level?
  - No alignment is needed.
  - Contextual information could matter.

E029 method

Given a sequence of input acoustic features and the corresponding transcript targets (BPE tokens):

$$\{x_1, \ldots, x_N\} = \boldsymbol{x}, \ \{y_1, \ldots, y_L\} = \boldsymbol{y}$$

The ASR posterior can be expressed as:

$$\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{x}; \boldsymbol{\theta}^{(m)}) = \lambda \mathrm{P}_{CTC}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{x}; \boldsymbol{\theta}^{(m)}) + (1-\lambda)\mathrm{P}_{seq2seq}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{x}; \boldsymbol{\theta}^{(m)})$$

The sequence-level confidence $\mathcal{C}_S$ is computed as:

$$\mathcal{C}_S = \exp\left[\frac{1}{L}\ln\sum_{l=1}^{L}\max\frac{1}{M}\sum_{m=1}^{M}\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{x}; \boldsymbol{\theta}^{(m)})\right]$$

E029 method

The sequence-level entropy $\mathcal{H}_S$ can be approximated with top samples in a beam-search candidates:

$$\mathcal{H}_S = -\sum_{b=1}^{B} \frac{\pi_b}{L^{(b)}} \ln \mathrm{P}(\boldsymbol{y}^{(b)}|\boldsymbol{x}, \boldsymbol{\theta})$$

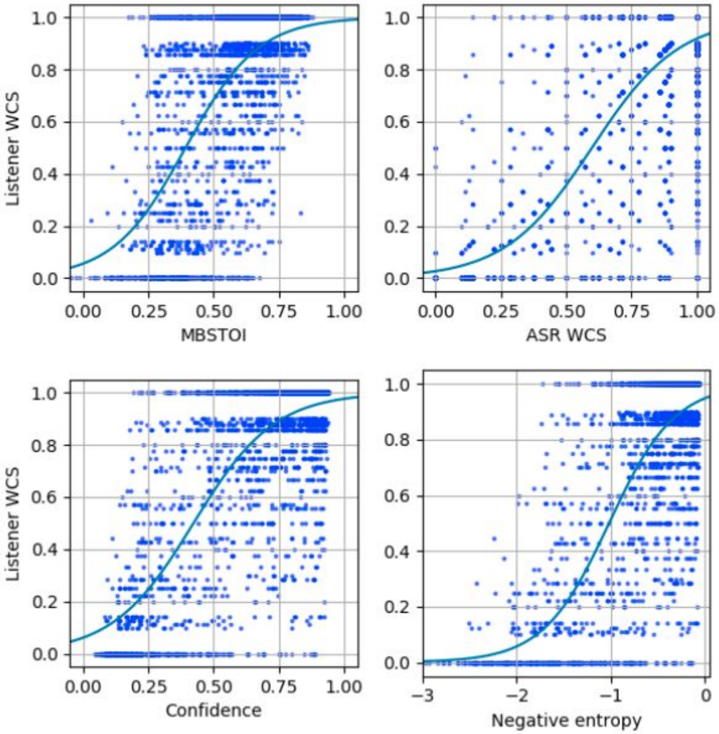
Where:

$$\pi_b = \frac{\exp \frac{1}{T} \ln \mathrm{P}(\boldsymbol{y}^{(b)}|\boldsymbol{x},\boldsymbol{\theta})}{\sum_{k}^{B} \exp \frac{1}{T} \ln \mathrm{P}(\boldsymbol{y}^{(k)}|\boldsymbol{x},\boldsymbol{\theta})}$$

$$\ln \mathrm{P}(\boldsymbol{y}^{(b)}|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{l^{(b)}=1}^{L^{(b)}} \ln \frac{1}{M} \sum_{m=1}^{M} \mathrm{P}(y_l^{(b)}|\boldsymbol{y}_{<l}^{(b)}, \boldsymbol{x}; \boldsymbol{\theta}^{(m)})$$

E029 Overall results

	WER	Measure	RMSE ↓	NCC ↑	KT ↑
Closed-set					
CPC1 Baseline	-	-	0.285	0.621	0.398
Proposed without MSBG	25.17	$\mathcal{C}_S$	0.241	0.751	0.472
		$-\mathcal{H}_S$	0.239	0.754	0.477
		ASR WCS	0.249	0.730	0.525
Proposed with MSBG	30.33	$\mathcal{C}_S$	0.234	0.767	0.497
		$-\mathcal{H}_S$	**0.233**	**0.768**	0.499
		ASR WCS	0.249	0.731	**0.526**
Open-set					
CPC1 Baseline	-	-	0.365	0.529	0.391
Proposed with MSBG	30.93	$\mathcal{C}_S$	0.248	0.729	0.512
		$-\mathcal{H}_S$	**0.246**	**0.734**	0.512
		ASR WCS	0.253	0.717	**0.530**



E029 conclusions

- The uncertainty estimated by an ensemble of powerful ASR models is naturally well correlated to speech intelligibility.
- MSBG hearing loss simulation can improve performance.
- The closed-set performance of the non-intrusive E029 is quite close to the intrusive E032.
- The intrusive E032 generalise better as its performance is better in the open-set.

System	RMSE ↓	NCC ↑	KT ↑
Closed-set			
E032(intrusive)	0.231	0.773	0.498
E029(non-intrusive)	0.233	0.768	0.499
Open-set			
E032(intrusive)	0.235	0.763	0.530
E029(non-intrusive)	0.246	0.734	0.512

Thank you for your attention!

The intrusive system E032 has been open-sourced in the Clarity Challenge github repository **recipes/cpc1/e032_sheffield**, please check ^_^: