Norway grants

NORDTRANS

T A
C R

Technology for automatic speech transcription in selected Nordic languages

Project code: TO01000027
2021/01 – 2024/04

newton technologies

TUL

NTNU
Norwegian University of Science and Technology

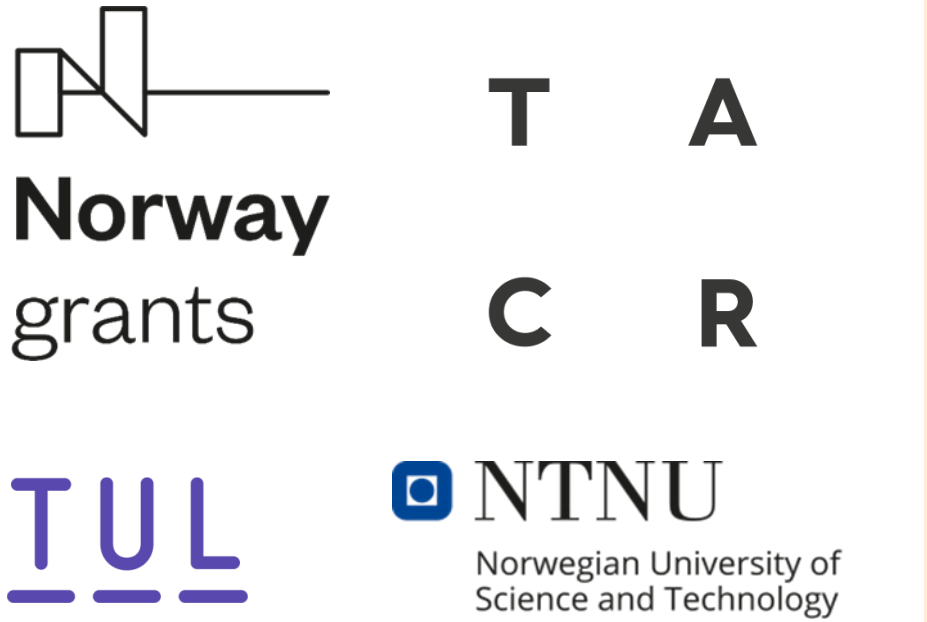
Norway grants

Project goal

T A C R

Developing a **highly accurate automatic speech recognition (ASR) algorithm** for **Swedish, Norwegian** and **Danish**

- Project consortium:
  - Newton Technologies (Czechia)
  - Norwegian University of Science and Technology (Norway)
  - Technical University of Liberec (Czechia)



Norway grants

Unleashing the power of language

T A C R

Speech-to-Text as a **pivot**

- Language technologies (incl. LLMs) work mostly on **written text**
- But digital life is mostly **audio** (and video – roughly 2/3 of all internet traffic is **audiovisual)**
- **S2T systems are the way to unleash the power of the latest AI systems on pristine content**

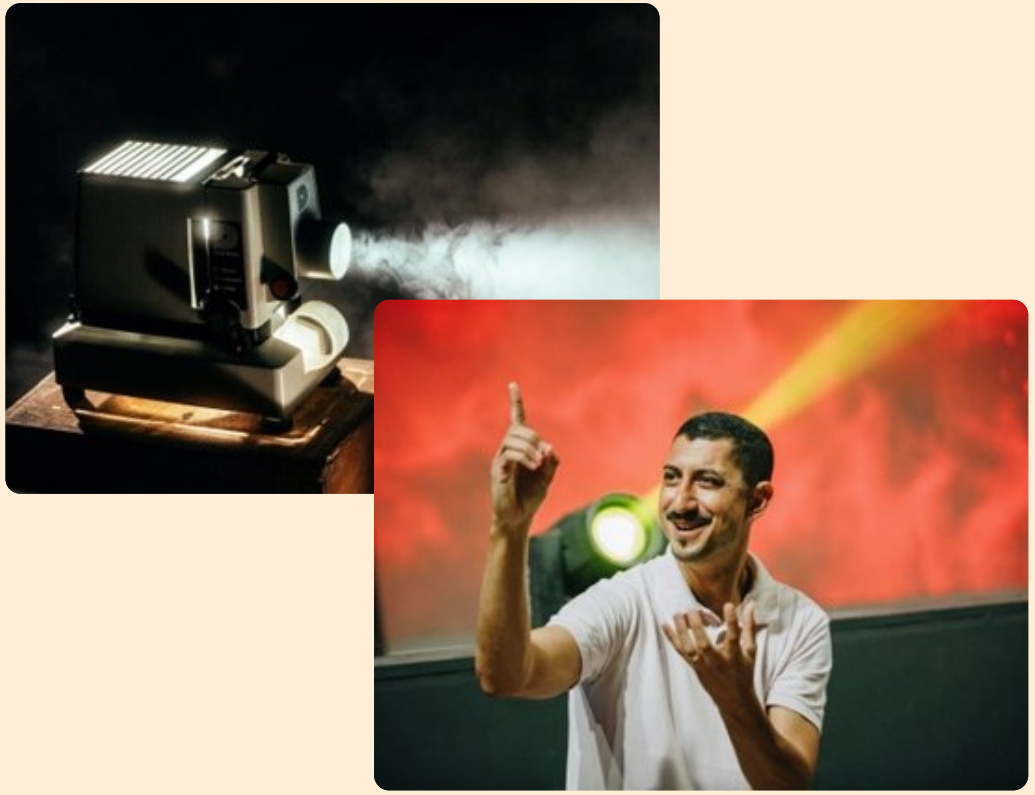
Photo by Brett Jordan on Unsplash

Norway grants

Why is it relevant

T A C R

- Enabling the disabled
- Overcoming language and cultural barriers
- Bridging the gap through data and information embedded in different media
- Use cases are varied and growing: **media monitoring, education, movies and television, government and institutions, consulting, security**



Photos by Jeremy Yap, Thiago Barletta on Unsplash

Norway grants

Why Newton Technologies

T A C R

Newton Tech as the industrial partner

- Committed to innovation
- **Strong IT and linguistic skills and experience**
- Large (and growing) international footprint
- **Gradual deployment from captive environment to partners' network**

Norway grants

End-to-end models

T A C R

- **Key innovation value** of this project in the AI field
- One single **massive neural network**
- No dictionary, language or acoustic models

+ Much **higher accuracy**
+ More effective for **noisy speech**
+ Strongly reduced **manual work**

- **Vast amounts** of training data needed
- Neural network cannot be **corrected**

Norway grants

Project results

T A C R

Practical

Modules for automatic transcription and processing of spoken **Norwegian, Swedish and Danish**

- Available in the **Beey platform** by Newton Technologies
- Transcription of TV, radio or any audiovisual content
- Subtitling, captioning, media monitoring

Scientific

1 journal article, 10 conference papers

Norway grants

Developed ASR modules

T A C R

- Based on recent **Deep Neural Networks (DNN) architectures**
  - Each module has 140M of parameters
  - Trained on a cluster of 4 GPUs (for several days)
- Utilized **end-to-end approach**
  - Only one large DNN is employed
  - Trained using **speech recordings** and **text transcripts**
    - No lexical on pronunciation model is needed
  - **Large amounts** of data are required for training
    - Own approach for data harvesting from public sources
    - Around **1000 hours** for each language

Image generated by Microsoft Copilot

Norway grants

Main advantages

T A C R

- **real-time processing** just using a CPU
  - no GPU is needed in inference time
- **automatic punctuation** and **capitalization**
  - using large pre-trained language models (LLMs)
- **fully customizable** post-processing (formatting) for individual users
  - numbers, abbreviations, etc.
- possibility of adding **user words**

Image generated by Microsoft Copilot

Norway grants

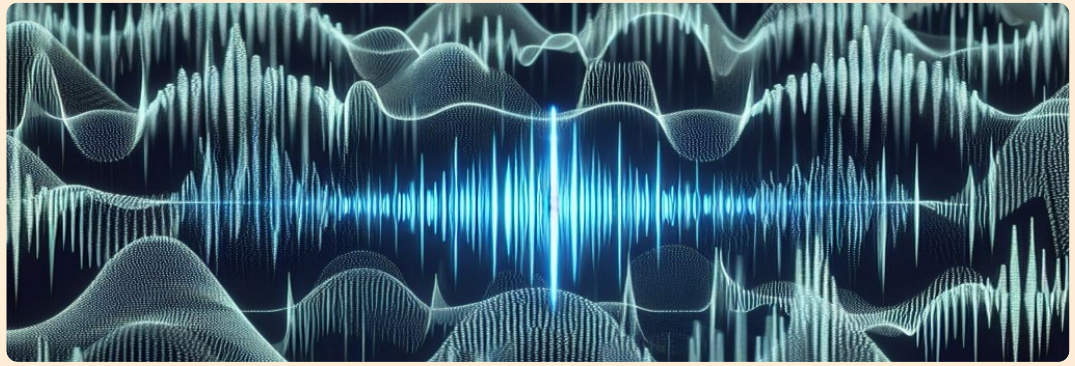
Achieved performance

T A C R

SWEDISH		READ NST5h	PARLIAMENT	SVT TV	YOUTUBE	READ FLEURS	READ CV 9	AUDIOBOOKS		TOTAL
WHISPER LARGE (using GPU)		88.9	88.8	78.9	86.6	90.0	88.4	93.5		89.3
MICROSOFT AZURE		94.5	88.5	**89.2**	**89.9**	87.1	89.5	88.6		88.8
NEWTON		**97.0**	**92.7**	87.9	88.5	**89.1**	**94.1**	**96.3**		**92.3**
DANISH		READ NST3h	PARLIAMENT	DR-TV2023	YOUTUBE	READ FLEURS	READ CV 13	AUDIOBOOKS		ALL
WHISPER LARGE (using GPU)		84.1	85.8	79.2	86.3	76.1	85.6	79.1		82.5
MICROSOFT AZURE		95.5	90.8	90.1	89.5	85.3	90.2	90.6		90.3
NEWTON		**97.1**	**95.7**	**90.7**	**92.7**	**91.1**	**94.0**	**93.4**		**93.6**
NORWEGIAN		READ NST5h	READ NPSC5h	SPONTANEOUS TALE	RUNDKAST TV	NRK TV	READ CV 12	AUDIOBOOKS		TOTAL
WHISPER LARGE (using GPU)		84.0	81.3	82.9	86.5	86.1	56.8	88.3		84.2
MICROSOFT AZURE		**98.8**	92.5	**89.7**	**90.0**	**89.7**	57.6	88.8		91.9
NEWTON		96.6	**94.9**	86.2	88.7	87.1	**81.6**	**89.1**		**92.3**

Metric used: Accuracy (percentage of correct words)

Extraction of Target Speaker from Mixtures of Speech Signals and Environmental Noise



J. Malek, J. Jansky, Z. Koldovsky, T. Kounovsky, J. Cmejla and J. Zdansky, "Blind Extraction of Target Speech Source Guided by Supervised Speaker Identification via X-vectors," *Accepted to IEEE Transaction on Audio, Speech and Language Processing*

Image generated by Microsoft Copilot

Norway grants

Motivation

T A C R

- Speech recordings from real environments contain **noise** and sometimes **unwanted speakers**
- These phenomena reduce
  - **intelligibility** for human listeners
  - **accuracy** of ASR systems
- Can be compensated by **extracting** the voice of the target speaker from mixture of speech signals
  - Other signals are suppressed

Image generated by Microsoft Copilot

Norway grants

Proposed extraction method

T A C R

- The method consists of two modules:
- **Identification module**
  - requires training data from the target speaker (30 s of speech)
  - provides the extraction module with information about the desired speaker
- **Extraction module**
  - No training is required
  - Selects individual signals from the mixture just based on their **statistical independence**
- **Advantages of the proposed modular approach**
  - Requires almost no adaptation to different environments
  - Almost language independent
    - Just the identification module is slightly language dependent

Norway grants

Demo of the proposed method

T A C R

- Extraction of target English speaker from a mixture of two speeches and coffee shop noise
  - The speakers are moving

Original mixture 

Target speaker 

Norway grants

Benefits for ASR

T A C R

- Extraction of the target speaker significantly **improves transcription accuracy**:

	Noisy signal (WER [%])	Extracted signal (WER[%])
CHIME4 English data set	19,9	9,3

Metric used: WER – word error rate (ratio of errors to all words, the lower the better)

Norway grants

NTNU
Norwegian University of Science and Technology

T A
C R

Semantically Meaningful Metrics for Norwegian ASR Systems

Janine Rugayan, Torbjørn Svendsen, Giampiero Salvi

Trondheim, March 14, 2022

Norway grants

Measuring ASR performance

T A C R

- Word error rate (WER)
  - Widely used metric
  - WER = total number of errors / total number of words
  - All errors are weighed equally

Reference: This is a cat.

ASR1: This is a **bat**.

ASR2: **It** is a cat.

Norway grants

What is the problem?

T A C R

- Not all errors are equally important
- We want a more robust and semantically meaningful measure compared to WER
- Norwegian language's special characteristics
  - two written standards: Bokmål and Nynorsk

    "to come"

    Bokmål: *å komme*

    Nynorsk: *å kome, å koma, å komme, å komma*
  - orthography is not strict
  - no standard way of speaking
  - high number of compound words

    småbarnsfamiliehovedadministrator

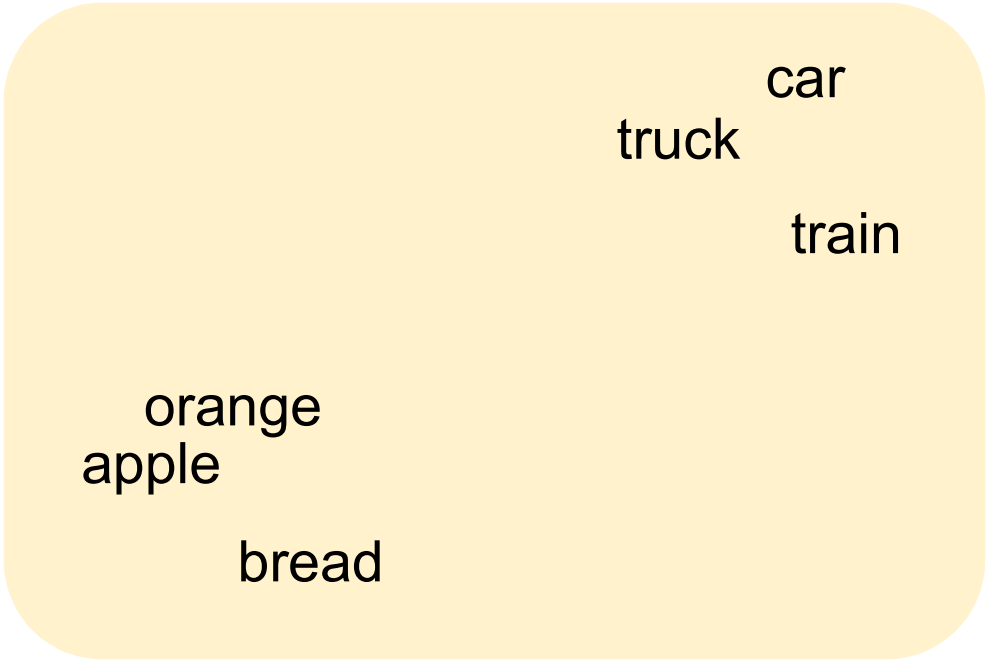
    "the chief administrator of a family with small children"

Norway grants

Solution: Use semantic information

T A C R

- Recently developed language models **capture semantic information**
  - Utilized to extract embeddings which are numerical representations of words in a vector space
  - Proximity in the vector space indicates semantic similarity

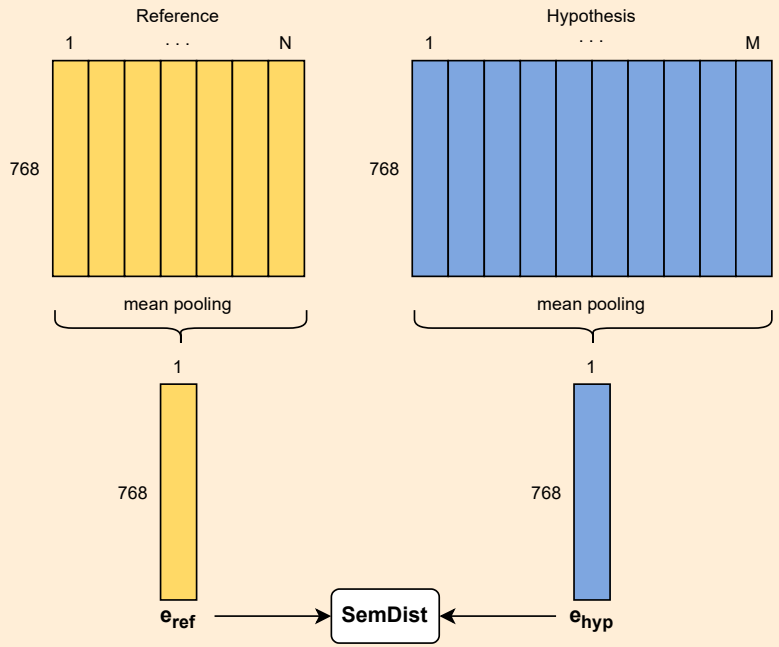


Norway grants

Semantic-based metric

T A C R

Semantic Distance[1]

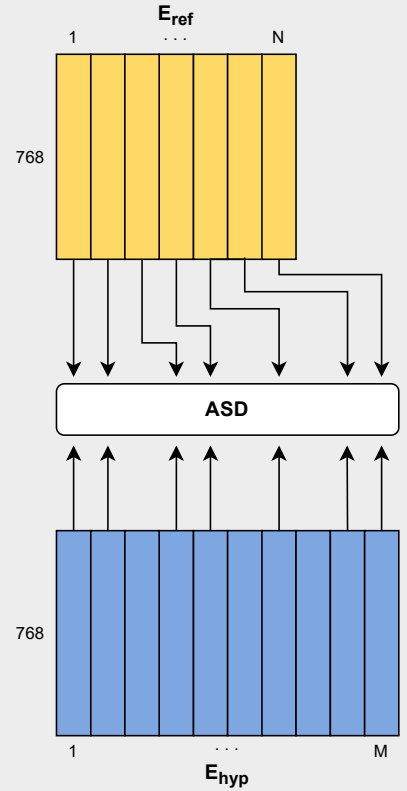


[1] Kim, S., Arora, A., Le, D., Yeh, C.-F., Fuegen, C., Kalinli, O., Seltzer, M.L. (2021) Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. Proc. Interspeech 2021, 1977-1981, doi: 10.21437/Interspeech.2021-1929

[2] Rugayan, J., Svendsen, T., Salvi, G. (2022) Semantically Meaningful Metrics for Norwegian ASR Systems. Proc. Interspeech 2022, 2283-2287, doi: 10.21437/Interspeech.2022-817

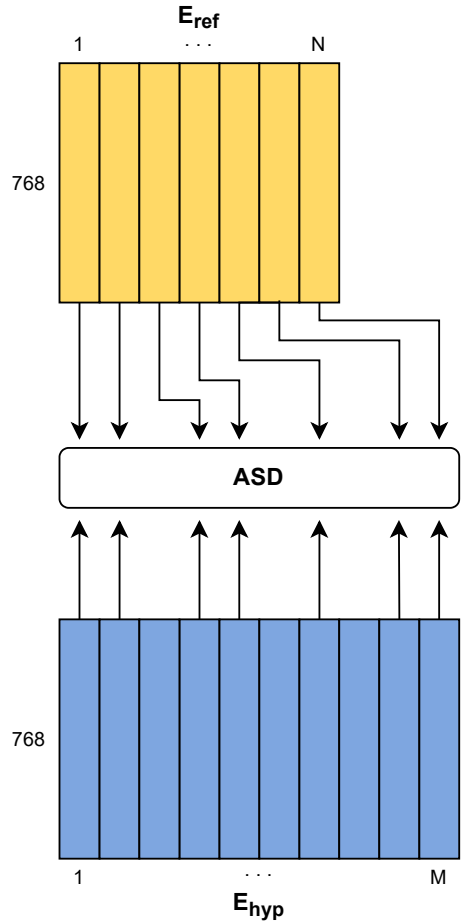
Aligned Semantic Distance[2]

- our proposed method



Aligned semantic distance (ASD)

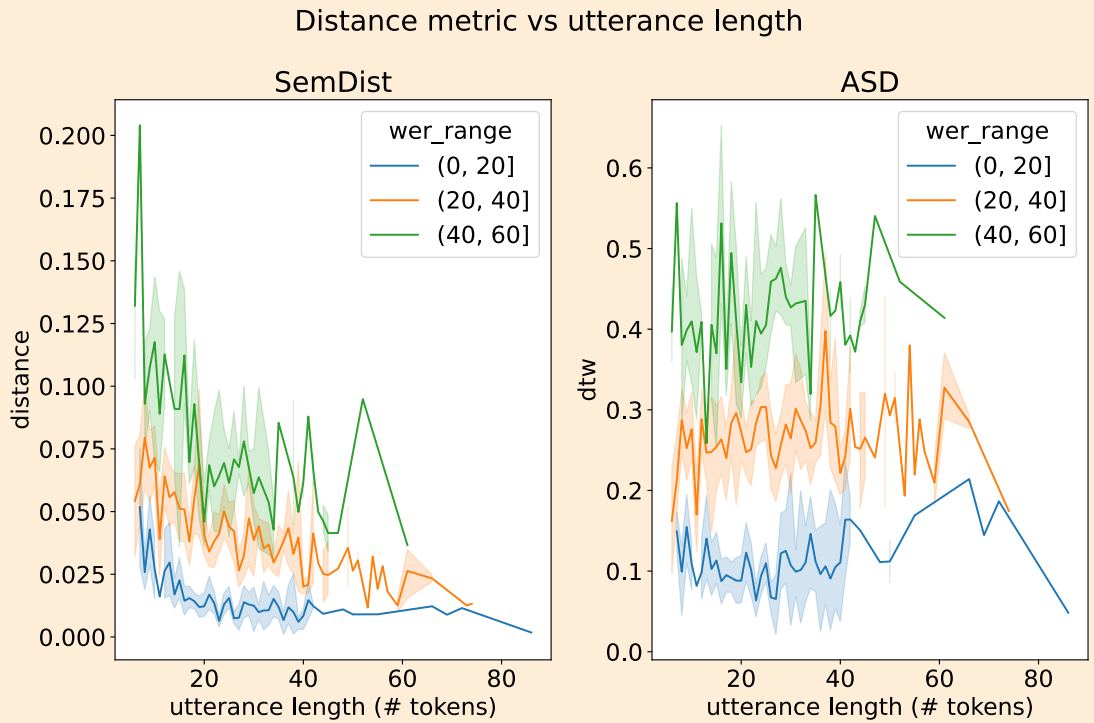
- Word-to-word comparison of embeddings
- Find the optimal alignment between the reference and ASR hypothesis
- Experiments:
  - used existing Norwegian language model for extracting embeddings
  - applied ASD to transcriptions of various speech data sources (NB Tale, Rundkast, Stortinget)



Norway grants

Our results

T A C R

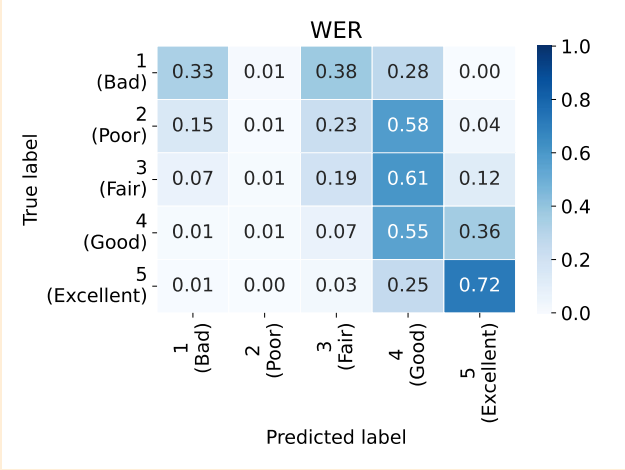


- Our proposed method ASD is stable with respect to utterance length
- ASD provides a more meaningful metric compared to word error rate
- ASD is useful for Norwegian
  - low penalty for equivalent Bokmål and Nynorsk words

ASD vs. Human Perception[1]

- ASD has better correlation to human perception compared to WER
- Prediction model for human perception is feasible
  - with ASD being a better input vs. WER

Normalized confusion matrices of Gaussian Naive Bayes classifier with 10-fold cross-validation.



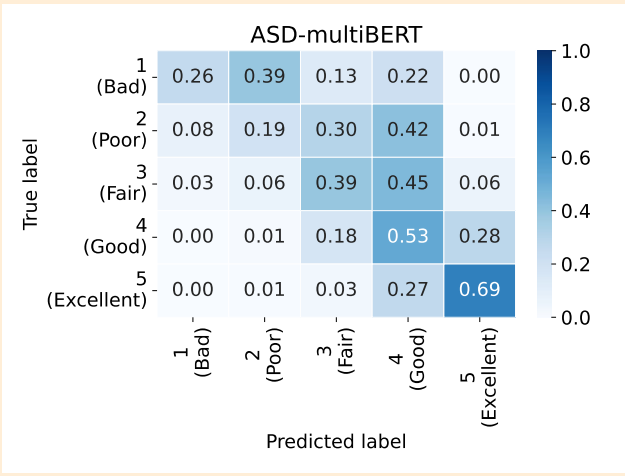


Table 1. Correlation of evaluation metrics with human evaluation scores of ASR quality

Metric	Layers	Correlation Coeff.
WER	1 - 12	-0.604
ASD-NorBERT	1 - 12	-0.646
ASD-multiBERT	1 - 12	**-0.683**
ASD-multiBERT	1 - 4	-0.659
ASD-multiBERT	5 - 8	**-0.698**
ASD-multiBERT	9 - 12	-0.684

Table 2. Comparison of linear regression models

Metric	MSE	MAE	R^2
WER	0.98	0.79	0.34
ASD-NorBERT	0.94	0.78	0.37
ASD-multiBERT	**0.82**	**0.72**	**0.45**

[1] Rugayan, J., Salvi, G., Svendsen, T. (2023) Perceptual and Task-Oriented Assessment of a Semantic Metric for ASR Evaluation. Proc. INTERSPEECH 2023, 2158-2162, doi: 10.21437/Interspeech.2023-1778

Norway grants

ASD vs. NLP tasks[1]

T A C R

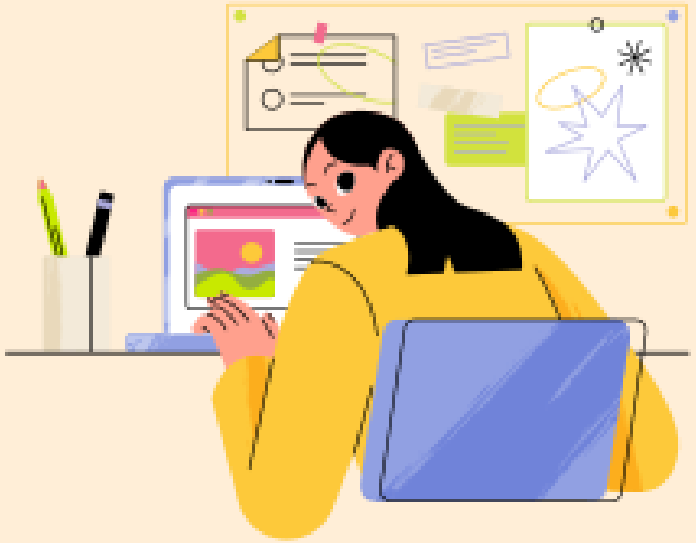
WER, ASD, and F1-scores on the NER & sentiment classification task.

Hypothesis Set	WER	ASD	F1 Score NER	F1 Score Sentiment
Baseline	0.074	0.066	0.878	0.938
Better ASD	0.074	0.051	0.944	0.957
Worse ASD (random)	0.074	0.080	0.893	0.898
Worse ASD (entity priority)	0.074	0.093	0.704	0.926

Compared to WER, ASD is a better indicator of downstream NLP tasks - named entity recognition (NER) and sentiment classification

[1] Rugayan, J., Salvi, G., Svendsen, T. (2023) Perceptual and Task-Oriented Assessment of a Semantic Metric for ASR Evaluation. Proc. INTERSPEECH 2023, 2158-2162, doi: 10.21437/Interspeech.2023-1778

What‘s next?

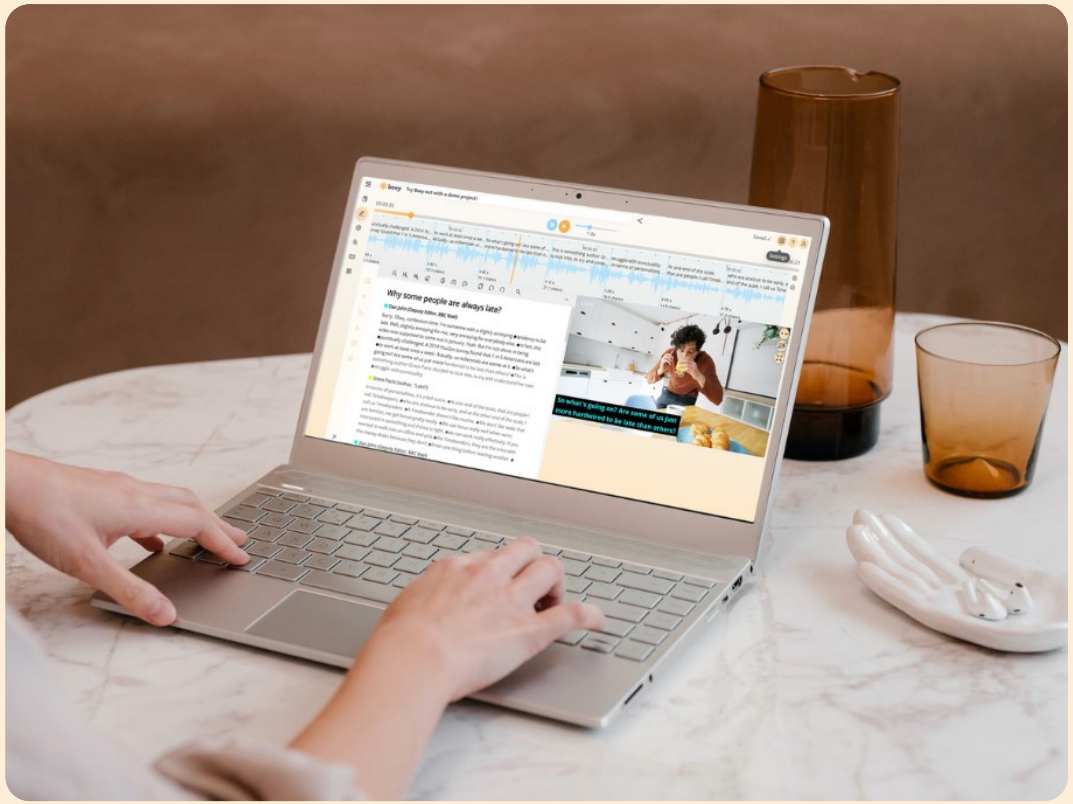


- Utilize ASD as optimization criteria for improving ASR models
- Demonstrate the applicability of the metric on other languages or other tasks

Norway grants

From technology to the user: Beey

T A C R



- Introducing **Beey**: Platform for fast transcribing, editing & more
- 23 languages - now including **Norwegian**, **Swedish** and **Danish**

Sign here to try it yourself
2 hours of free transcriptions
beey.io/trondheim2024