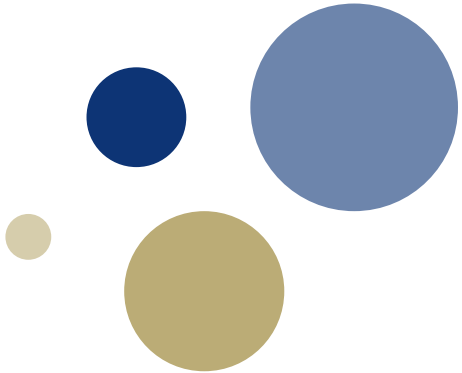
Norwegian University of
Science and Technology



Collaborative Projects in Speech Technology

Some highlights

Torbjørn Svendsen

Child speech recognition

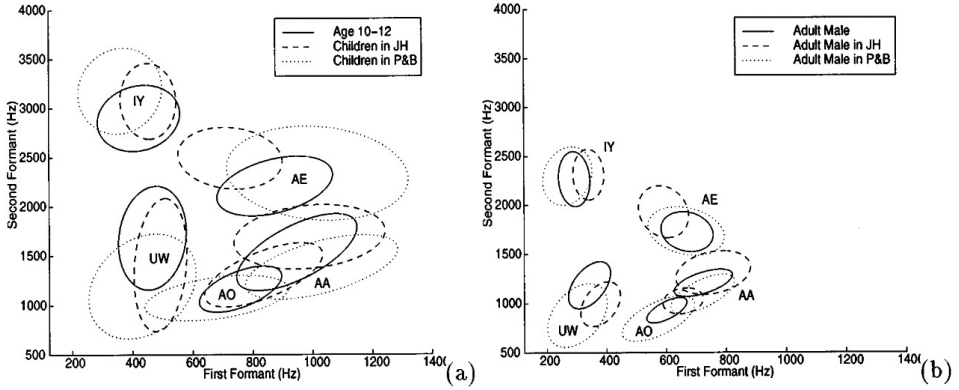
- Research (master projects++) since 2011
  - Fjær 2011; Walsøe 2016; Thorsrud 2017; Steinskog 2021
  - Interspeech 2013 (Doddipatla&Svendsen)
- Focus on voice conversion approaches to utilize adult speech databases for training systems also for children
- Currently two research projects with PhD addressing child speech recognition

e-LADDA: Early language development in the digital age

- Interdisciplinary Marie Sklodowska-Curie Innovative Training Network
- Goal: establish whether the new and intuitive interactions afforded by digital tools impact on young children's language development and language outcomes in a positive or adverse way.
- Focus on child/computer interaction and computer assisted early language acquisition
- NTNU: Child speech recognition and synthesis
- https://www.ntnu.edu/e-ladda/
- PhD Student: **Zijian Fan**

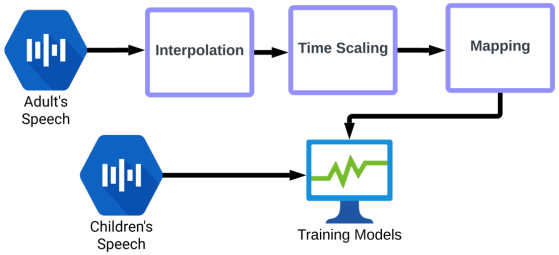
Children vs adults

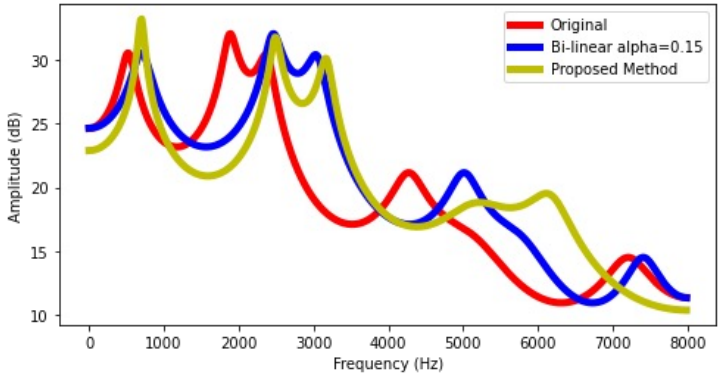
- More acoustic variability due to language development
- Non-linear spectral changes due to physiology
- Also: cognitive development and stage of language acquisition



From: Lee, Potamianos, & Narayanan : Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am* **105**, 1455–1468 (1999).

Non-linear spectral modification





- Traditional methods: VTLN, bilinear warp
- New proposal, including speed modification and linear warp
- Also: Transfer learning, end-to-end methods

Teflon

- NordForsk project
  - Aalto University, Tampere University (Finland)
  - Karolinska Insitutet (Sweden)
  - University of Oslo, NTNU (Norway)
- Recognizing speech from 2nd language child learners
  - Norwegian, Swedish, Finnish
  - pronunciation assessment in a gamified learning environment
  - child speech
  - speech pathologies
- PhD Student: **Cao Xinwei**

Goodness of pronunciation

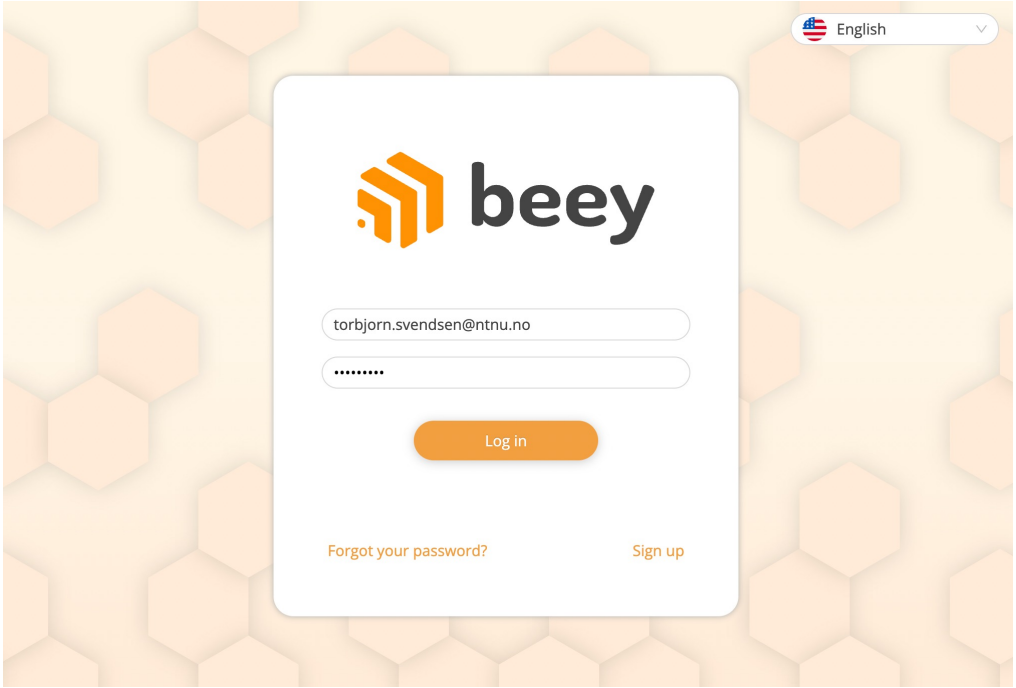
- Estimation of how well the pronunciation of a student matches the correct pronunciation

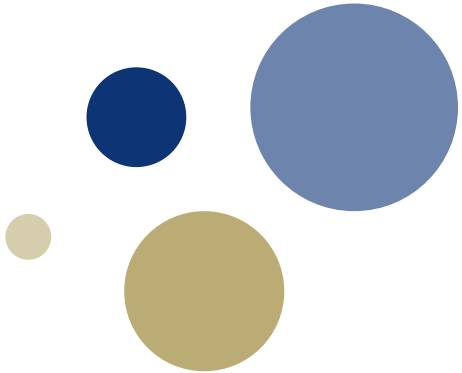
$$GoP(p) = \frac{1}{T}\left|\log \mathcal{P}(p|\mathbf{O})\right| = \frac{1}{T}\left|log\frac{\mathcal{P}(\mathbf{O}|p)\mathcal{P}(p)}{\sum_{q\in Q}\mathcal{P}(\mathbf{O}|q)\mathcal{P}(q)}\right|$$

- Corrective feedback to student

NordTrans - Technology for automatic speech transcription in selected Nordic languages

- EEA-grants project, NTNU, Newton Technologies (CZ), TU Liberec (CZ)
- Improve the state-of-the-art quality and usability of the automatic speech recognition (ASR) technology for Swedish and Norwegian.
- Focus on streaming audio, e.g radio, TV, internet podcasts, etc.
  - Other applications: transcription of speeches in parliaments and similar public institutions, as well as spoken archive mining
- Based on ASR engines developed by Newton and TUL
- 1 PhD student
  - Algorithm development
  - Semantically meaningful performance metrics


English
beey
torbjorn.svendsen@ntnu.no
Log in
Forgot your password?
Sign up



Semantically Meaningful Metrics for Norwegian ASR Systems

Janine Rugayan, Torbjørn Svendsen, Giampiero Salvi

Prague, November 24, 2022

How do we measure ASR performance?

- 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.

What is the problem?

- 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"

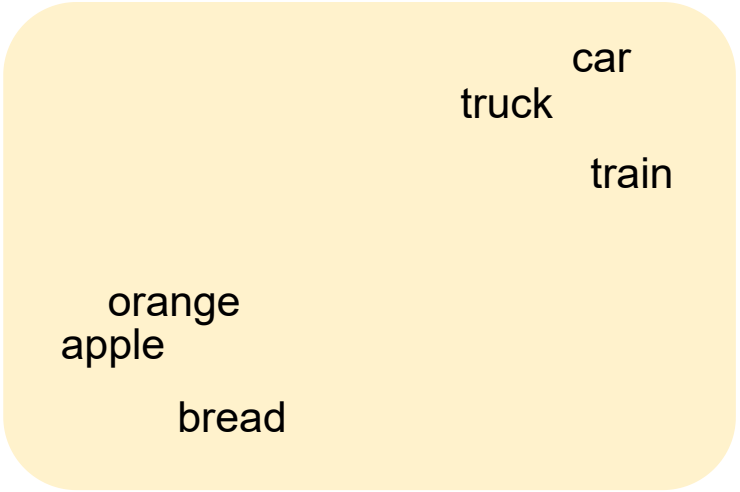
    Bokmål: *å komme*

    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"

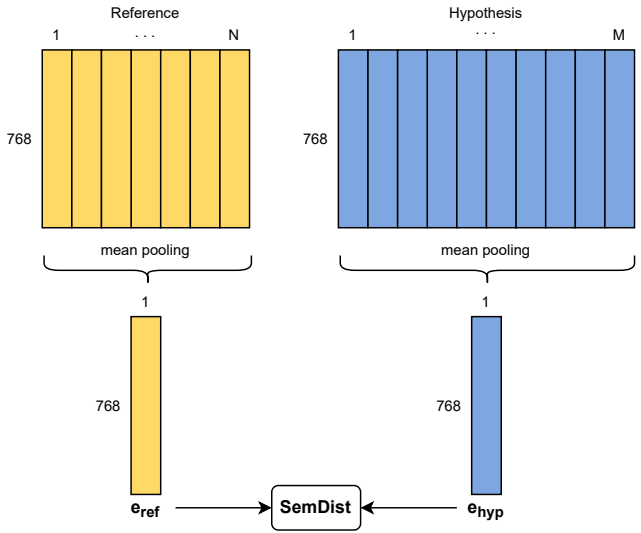
Solution: use semantic information



- 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

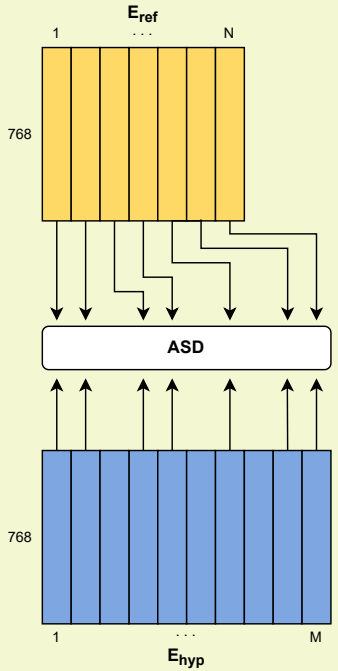
Semantic-based metrics

Semantic Distance[1]



Aligned Semantic Distance[2]

- our proposed method

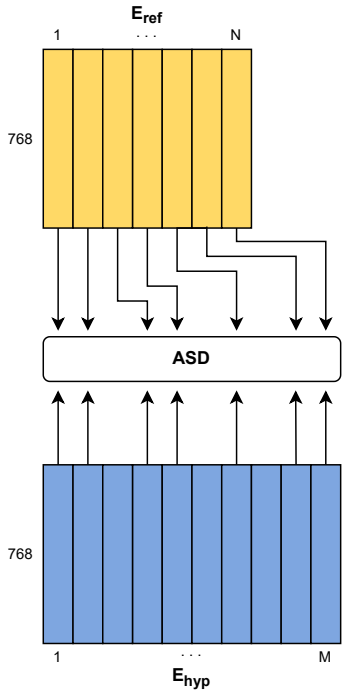


[1] Kim, S., Arora, A., Le, D., Yeh, C.-F., Fuegen, C., Kalinli, O., Seltzer, M.L. (2021) Semantic Distance: A New Metric for ASR Performance Analysis Towards Spoken Language Understanding. Proc. Interspeech 2021, 1977-1981, doi: 10.21437/Interspeech.2021-1929

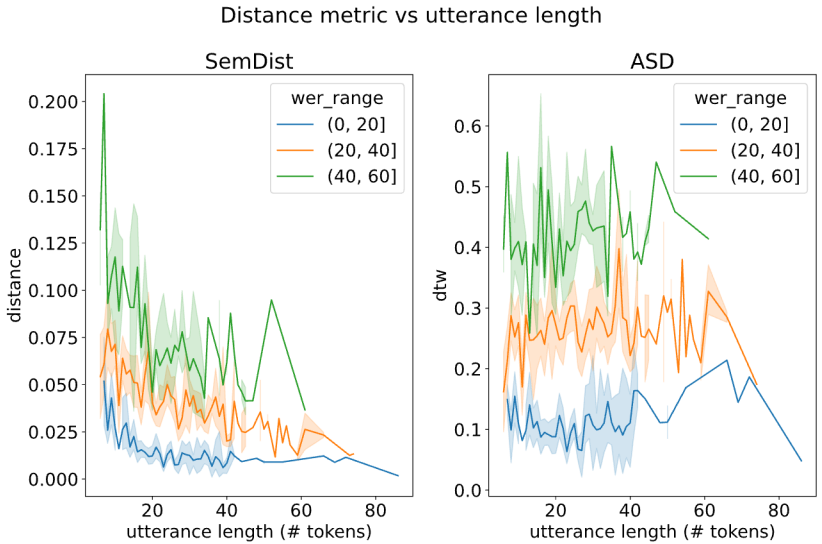
[2] Rugayan, J., Svendsen, T., Salvi, G. (2022) Semantically Meaningful Metrics for Norwegian ASR Systems. Proc. Interspeech 2022, 2283-2287, doi: 10.21437/Interspeech.2022-817

Aligned Semantic Distance (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)



Our Results



- 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

Future Work

- Perform a correlation study between user-rated transcriptions and ASD (ongoing)
- Evaluate the metric against a downstream task
- Explore application of ASD on other languages, for instance, English

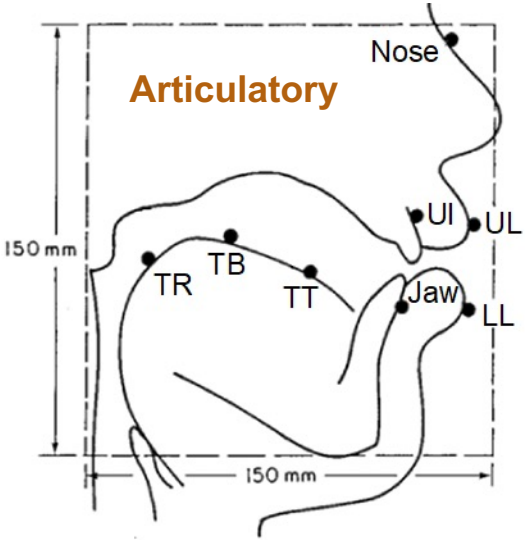
SCRIBE – Machine Transcription of Conversational Speech

Develop a Norwegian speech-to-text transcription system for multi-party conversations in realistic recording conditions

- Develop models that are robust to disfluencies that are typical in spontaneous conversational speech, that can cope with turn taking and take advantage of the context in the dialog.
- The models will also support the use of spoken dialects and different orthographies (Bokmål, Nynorsk, or dialect specific).
- Define evaluation metrics that predict the quality of the transcription based on semantics rather than merely word error rate.
- Contribute to the theoretical and methodological development of machine learning with sparse data.

Recent PhD Projects

Reza Sabzi: Articulatory inversion



$y = f(x)$

AA

I

$x = f^{-1}(y)$

AI

Acoustic

'The birch canoe slid on the smooth planks.'

Linguistic

Recent PhD Projects

Femke Gelderblom: Evaluating PerformanceMetrics for Deep Neural Network-based Speech Enhancement Systems

Important finding: Popular metrics (PESQ, STOI, HASPI…) for assessment of speech quality does not match human perception for quality evaluation of enhanced speech.