Generating More Audios for End-to-End Spoken Language Understanding

Patrick Schrottenbacher

Agenda

1. Einleitung
2. Transcription, ASR, SLU
3. Spoken Language Understanding
4. GMA-SLU
5. Ergebnisse
6. Fazit

Einleitung

- **G**enerating **M**ore **A**udios for End-to-End Spoken Language Understanding (**GMA**-SLU)
- Xuxing Cheng und Yuexian Zou präsentieren in dieser Arbeit einen Prozess, um Datensätze zu erweitern indem weitere synthetische Daten erzeugt werden
- Dieser Prozess führt zu statistisch signifikanten Unterschieden bezüglich der Genauigkeit von SLU-Modellen auf gängigen Datensätzen

Agenda

1. Einleitung
2. Transcription, ASR, SLU
3. Spoken Language Understanding
4. GMA-SLU
5. Ergebnisse
6. Fazit

Transcription, ASR, SLU

- Natural Language Processing (NLP) bzw. Computerlinguistik beschäftigt sich mit der Fragestellung, inwiefern natürliche Sprachen mit (computer) Algorithmen verarbeitet werden können
- Ein Teilaspekt beschäftigt sich hierbei mit gesprochenen Aufnahmen

- Automatic Speaker Recogniton (ASR)
  $\Rightarrow$ Gesprochenes zu Text verarbeiten (Audio zu Text)
- Diarization
  $\Rightarrow$ **Wer** hat **wann** gesprochen?

Transcription (Transkribierung) als Oberbegriff

Transcription, ASR, SLU

- Wichtige Modelle und Innovation:
  - WaveNet (2016)
  - BERT (2019)
  - Whisper(2021)
  - Large Language Models
    - Phi-4 (2024 Dez)
    - Granite-speech (June 2025)

Agenda

1. Einleitung
2. Transcription, ASR, SLU
3. Spoken Language Understanding
4. GMA-SLU
5. Ergebnisse
6. Fazit

Spoken Language Understanding

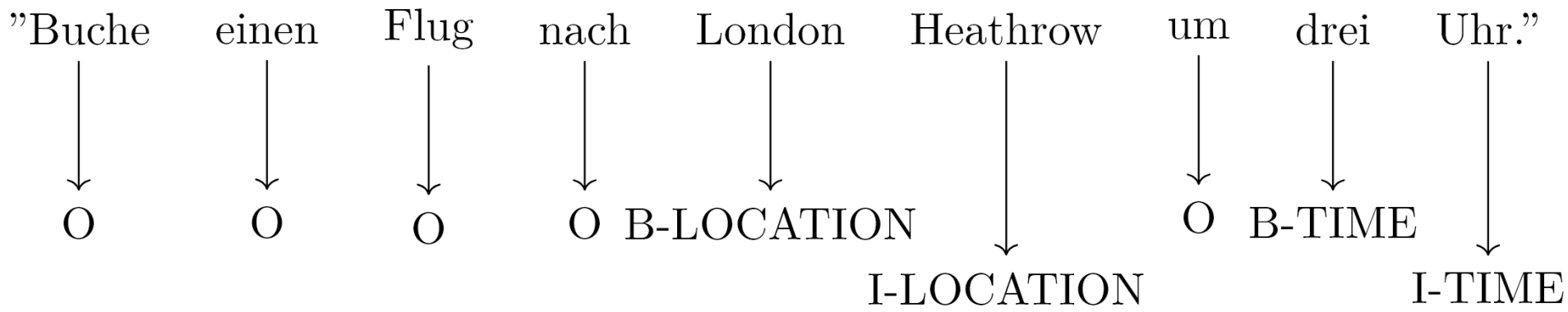
- Spoken Language Understanding versucht gesprochenen Audioaufnahmen so zu verarbeiten das darauf relevante Aktionen getätigt werden können

  ➔ Sprachassistenten, “spoken dialogue systems”
- Unterteilung in zwei Unteraufgaben Slot Filling (SF) und Intention Detection (ID)

Spoken Language Understanding

Slot Filling

Slot Filling nach BIO-Format:



Spoken Language Understanding

Intent Detection

- Intent Detection versucht hierbei ein (Intent)-Label für den kompletten Satz zu finden
- Für unser Beispiel "*Buche einen Flug nach London Heathrow um drei Uhr*" könnte dies z.B eine interne Funktion namens *"book_flight_with_time"* sein.
- Zumeist erfolgt dies durch traditionelle Machine Learning verfahren, wie Random Forests
- Seit ein paar Jahren kommen jedoch auch Transformer Modelle zum Einsatz wie z.B BERT, welche auch für **End-to-End** Lösungen angewendet werden.

Spoken Language Understanding

Problematiken

- Traditionellerweise wird, um SLU durchzuführen, ein ASR-Model herbeigezogen, um die SF und ID mit gängigen Modellen zu ermöglichen
- ASR-Modelle selbst jedoch sind nicht zu 100% genau, selbst die besten Modelle haben eine *Word Error Rate* (WER) von ca. 6%
- Ebenfalls kommt es durch die Modalitäts-umwandlung zu potentiellen Informationsverlusten
- Warum den ASR-Prozess also nicht umgehen?

  $\Rightarrow$**End-To-End** Systeme

Agenda

1. Einleitung
2. Transcription, ASR, SLU
3. Spoken Language Understanding
4. GMA-SLU
5. Ergebnisse
6. Fazit

Generating More Audios for End to End SLU

- **End-to-End** Systeme umgehen die ASR-Komponente, indem sie Modelle verwenden, welche trainiert worden sind um direkt von den Audiodaten die SF und ID durchzuführen
- Modelle, welche dies jedoch tun, sind i.d.R. jedoch schlechter als jene welche ASR verwenden
- Dies liegt unter anderem daran, dass es deutlich schwieriger ist end-to-end Modelle zu trainieren(?)
- **End-to-End** Modelle werden deswegen öfters mithilfe von Transkripten trainiert

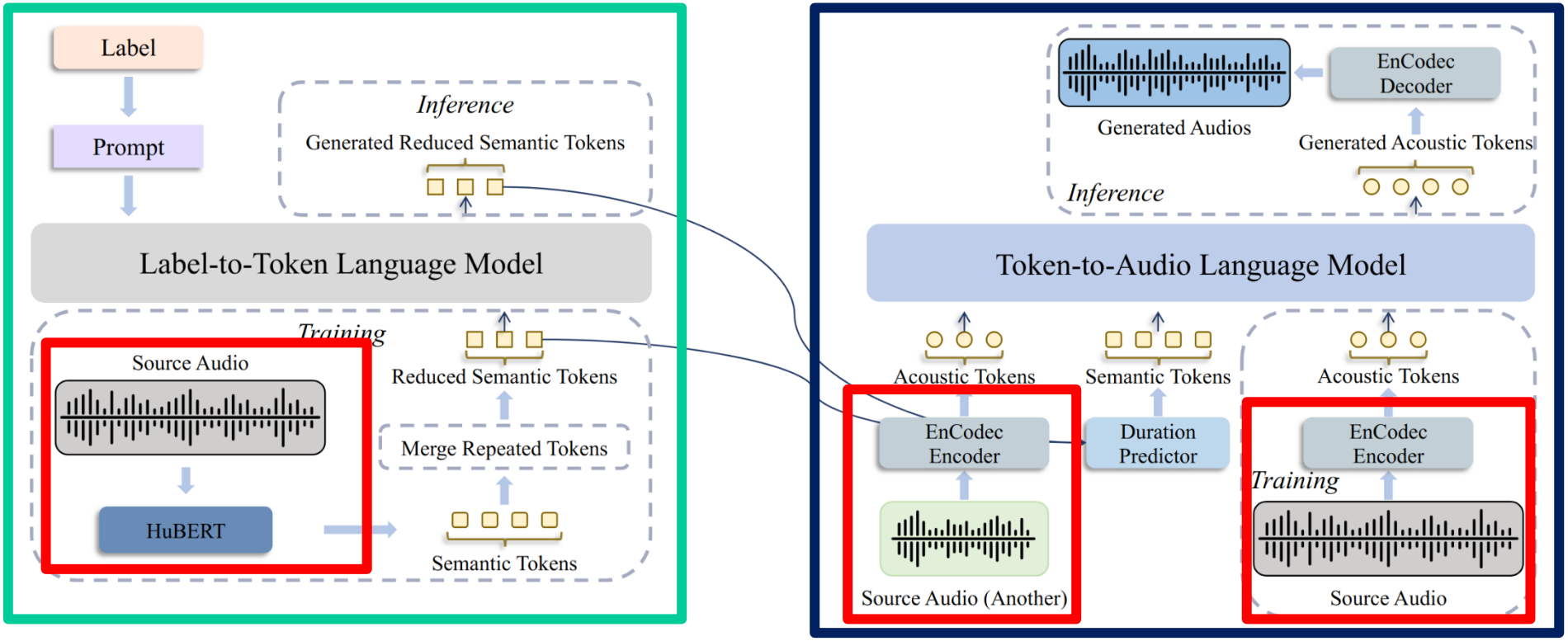
Generating More Audios for End to End SLU

- Die Anzahl an Datensätzen mit Transkripten für SLU-Aufgaben ist jedoch gering
- Vor allem existieren für viele Sprachen nur sehr kleine Datensätze

  $\Rightarrow$Wie können wir, am besten ohne großen Aufwand, an mehr Daten gelangen um bessere Modelle zu trainieren?

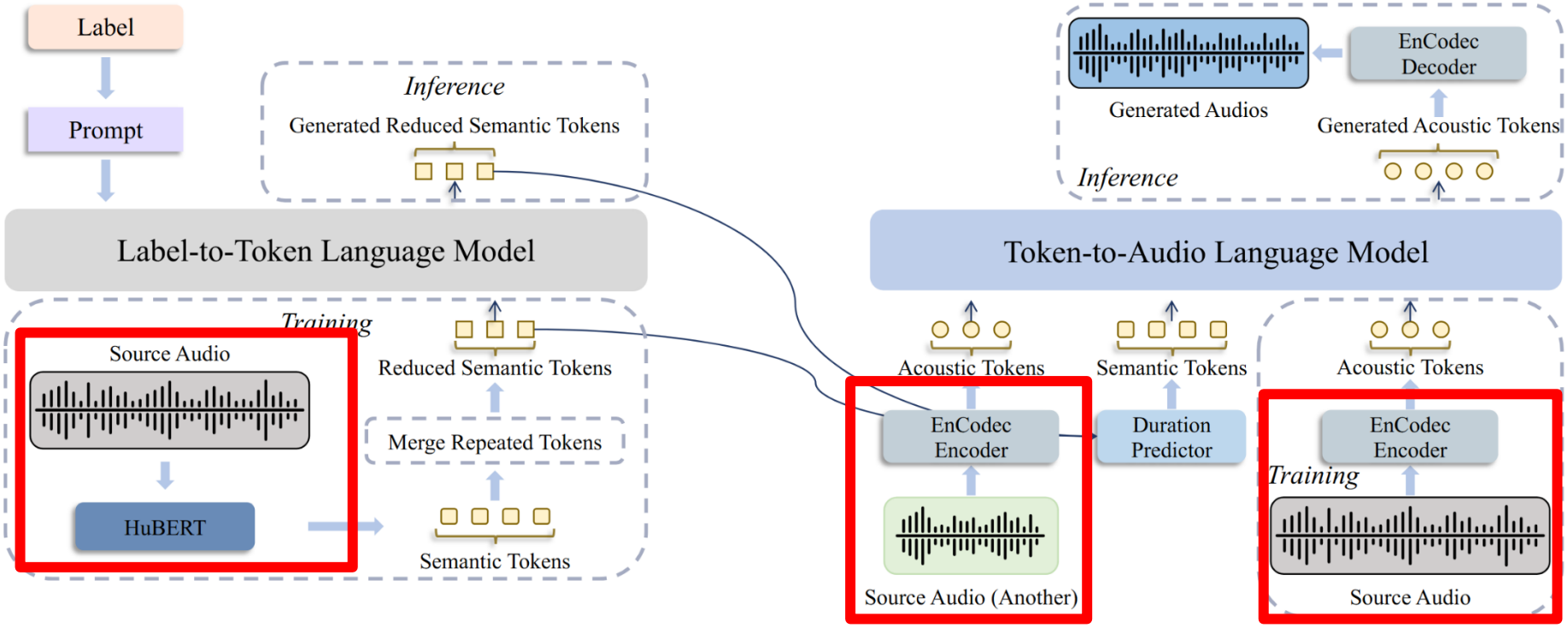
- **Idee**: Wir generieren mehr synthetische Audio Trainingsdaten basierend auf den existierenden

Generating More Audios



Generating More Audios

Discrete Tokens Generation



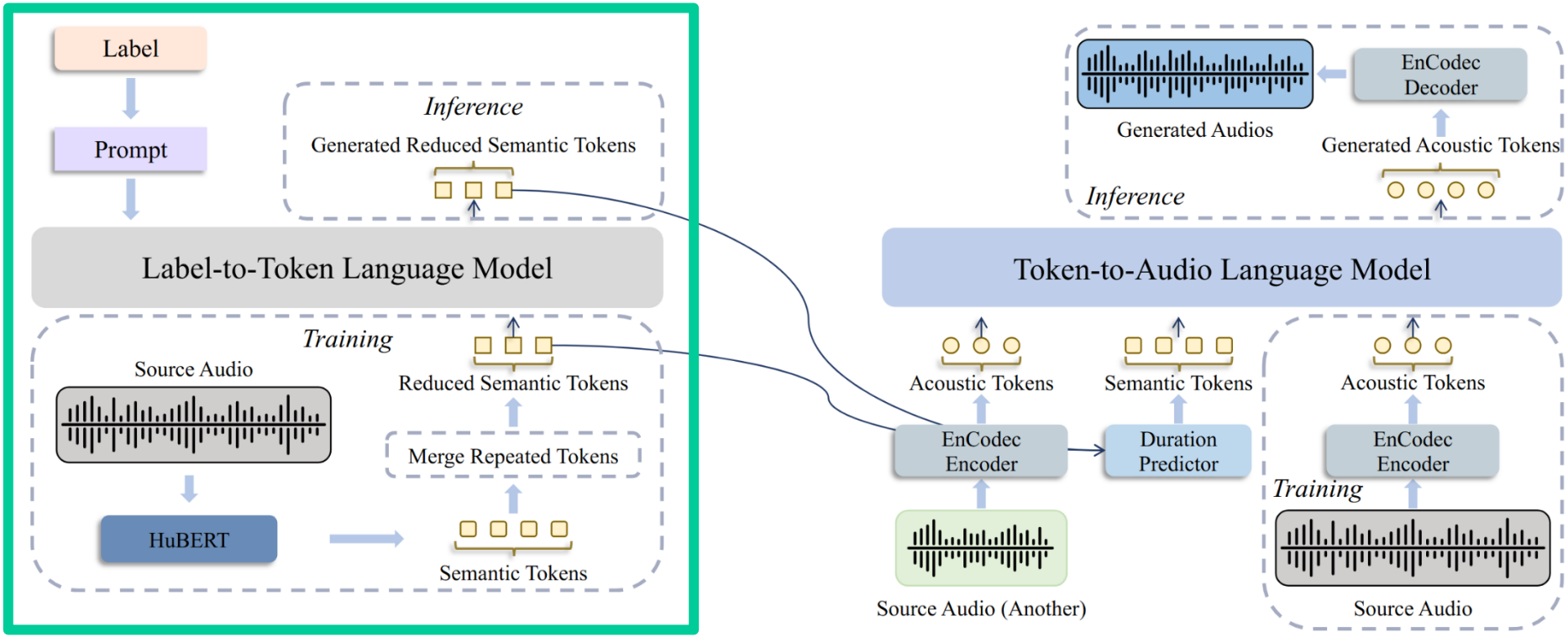
Generating More Audios

Discrete Tokens Generation

- Akustische und Semantische Tokens
- Die akustischen Tokens werden mithilfe von EnCodec erzeugt
- Die semantischen Tokens werden mithilfe von HuBERT erzeugt
  - Indiziert durch k-means clustering
  - Die Indizes repräsentieren dann ein Token

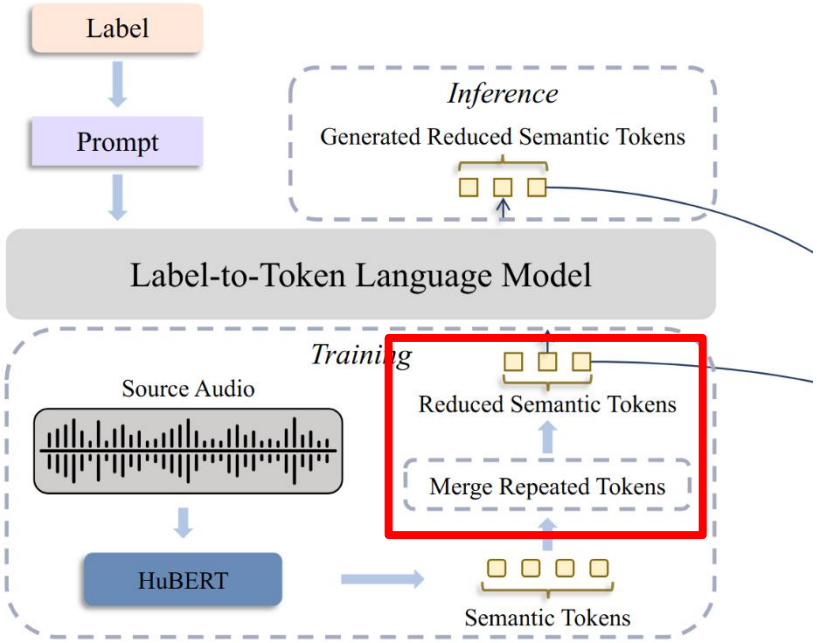
Generating More Audios

Lable-To-Token Language Model



Generating More Audios

Lable-To-Token Language Model



Generating More Audios

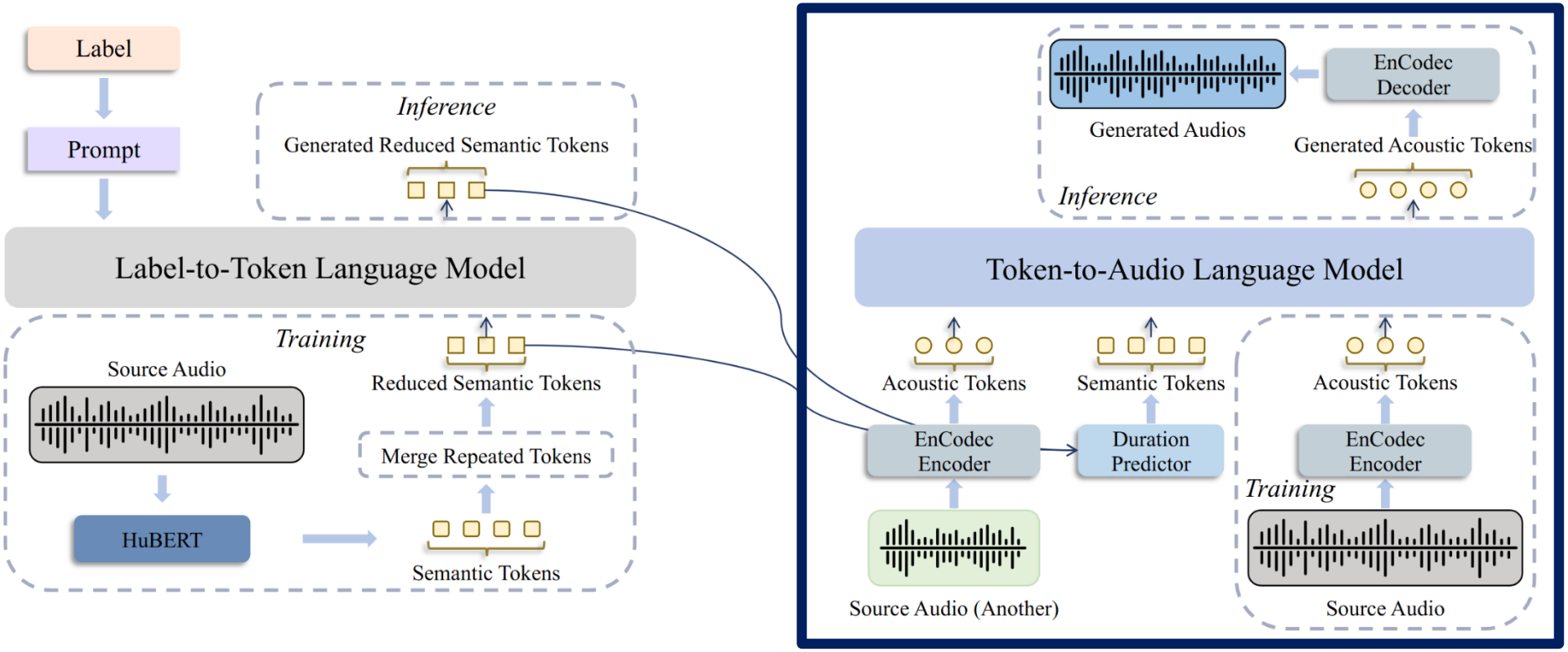
Lable-To-Token Language Model

- Zum trainieren des LLMs werden verschiedene Prompts für die Tokens Zufällig ausgewählt
- Sie entsprechen Grundsätzlich folgendem Schema:

  $\Rightarrow$“Generate the utterance whose intent is [I_label] and slot is [S_label]: [semantic token]”

Generating More Audios

Token-To-Audio Language Model



Generating More Audios for End to End SLU

- **Vorgang:**
  - Zunächst werden die Teilkomponenten auf den Datensätzen trainiert
  - Hierzu wird ein 30-20-50% train-dev-test split verwendet
  - Die besten Modelle, welche auf dem dev set trainiert wurden, werden dann verwendet
  - Das Model wird dann final zunächst auf den generierten Audiodaten trainiert und dann auf den echten Audioaufnahmen fine-tuned

Agenda

1. Einleitung
2. Transcription, ASR, SLU
3. Spoken Language Understanding
4. GMA-SLU
5. Ergebnisse
6. Fazit

Ergebnisse

- Um die Leistung der Architektur zu überprüfen wird sie auf MINDS-14 und SLURP, zwei SLU-Datensätzen, evaluiert
- SLURP beinhaltet ca. 72.000 englischsprachige Audioaufnahmen mit etwas weiteren 69.000 synthetischen
- MINDS-14 ist ein multilingualer Datensatz basierend auf 14 Sprachen a 600 individuellen Aufzeichnungen
- Neben der Genauigkeit von GMA-SLU wurden ebenfalls andere Teilstrategien untersucht:
  - Reduced Strategy
  - Different Acoustic Tokens
  - Data Filtration
  - Two-Stage Training Strategy

Ergebnisse

Model	Pre-trained Model	Slot (SLU-F1)	Intent (Acc)
HuBERT SLU	HuBERT	78.92	89.38
CIF-PT	-	78.67	89.60
CIF-PT	Data2vec	82.63	91.32
GMA-SLU w/o Reduced Strategy	HuBERT	79.68	90.39
GMA-SLU w/o Different Acoustic Tokens	HuBERT	79.45	90.13
GMA-SLU w/o Data Filtration	HuBERT	79.57	90.24
GMA-SLU w/o Two-Stage Training Strategy	HuBERT	79.82	90.55
GMA-SLU (base)	HuBERT	**80.93**	**90.97**
GMA-SLU	SpeechTokenizer	81.25	91.23
GMA-SLU	Llama 2	83.46	92.51

Ergebnisse

Model	Pre-trained Model	Slot (SLU-F1)	Intent (Acc)
HuBERT SLU	HuBERT	78.92	89.38
CIF-PT	-	78.67	89.60
CIF-PT	Data2vec	82.63	91.32
GMA-SLU w/o Reduced Strategy	HuBERT	79.68	90.39
GMA-SLU w/o Different Acoustic Tokens	HuBERT	79.45	90.13
GMA-SLU w/o Data Filtration	HuBERT	79.57	90.24
GMA-SLU w/o Two-Stage Training Strategy	HuBERT	79.82	90.55
GMA-SLU (base)	HuBERT	**80.93**	**90.97**
GMA-SLU	SpeechTokenizer	81.25	91.23
GMA-SLU	Llama 2	83.46	92.51

Ergebnisse

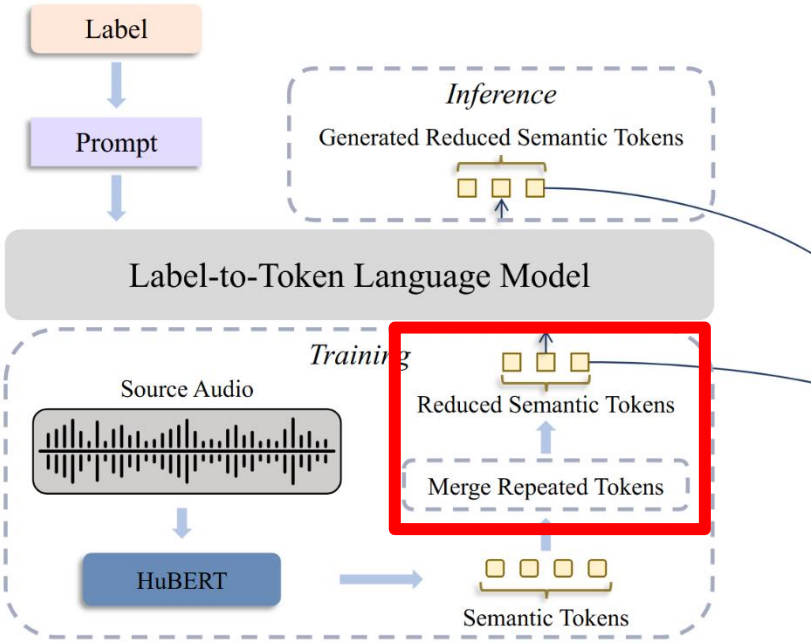
Reduced Strategy

Model	Pre-trained Model	Slot (SLU-F1)	Intent (Acc)
HuBERT SLU	HuBERT	78.92	89.38
CIF-PT	-	78.67	89.60
CIF-PT	Data2vec	82.63	91.32
GMA-SLU w/o Reduced Strategy	HuBERT	79.68	90.39
GMA-SLU w/o Different Acoustic Tokens	HuBERT	79.45	90.13
GMA-SLU w/o Data Filtration	HuBERT	79.57	90.24
GMA-SLU w/o Two-Stage Training Strategy	HuBERT	79.82	90.55
GMA-SLU (base)	HuBERT	**80.93**	**90.97**
GMA-SLU	SpeechTokenizer	81.25	91.23
GMA-SLU	Llama 2	83.46	92.51

Ergebnisse

Reduced Strategy

- Reduced Strategy bezieht sich auf die Zusammenführung akustische Tokens
- Falls dieser Schritt nicht erfolgt, verschlechtert sich die Genauigkeit um ca. 15%



Ergebnisse

Different Acoustic Tokens

Model	Pre-trained Model	Slot (SLU-F1)	Intent (Acc)
HuBERT SLU	HuBERT	78.92	89.38
CIF-PT	-	78.67	89.60
CIF-PT	Data2vec	82.63	91.32
GMA-SLU w/o Reduced Strategy	HuBERT	79.68	90.39
GMA-SLU w/o Different Acoustic Tokens	HuBERT	79.45	90.13
GMA-SLU w/o Data Filtration	HuBERT	79.57	90.24
GMA-SLU w/o Two-Stage Training Strategy	HuBERT	79.82	90.55
GMA-SLU (base)	HuBERT	**80.93**	**90.97**
GMA-SLU	SpeechTokenizer	81.25	91.23
GMA-SLU	Llama 2	83.46	92.51

Ergebnisse

Model	Pre-trained Model	Slot (SLU-F1)	Intent (Acc)
HuBERT SLU	HuBERT	78.92	89.38
CIF-PT	-	78.67	89.60
CIF-PT	Data2vec	82.63	91.32
GMA-SLU w/o Reduced Strategy	HuBERT	79.68	90.39
GMA-SLU w/o Different Acoustic Tokens	HuBERT	79.45	90.13
GMA-SLU w/o Data Filtration	HuBERT	79.57	90.24
GMA-SLU w/o Two-Stage Training Strategy	HuBERT	79.82	90.55
GMA-SLU (base)	HuBERT	**80.93**	**90.97**
GMA-SLU	SpeechTokenizer	81.25	91.23
GMA-SLU	Llama 2	83.46	92.51

Ergebnisse

- Ebenfalls wurde analysiert wie gut Sprachmodelle an sich die Aufgabe eines SLU Modelles erfüllen können
- Hierzu wurden Transkripte erstellt, welche dann an ChatGPT und SpeechGPT weitergegeben wurden um ID durchzuführen

Ergebnisse

Model	en-US	fr-FR	pl-PL	ko-KR
LaBSE [Gerz *et al.*, 2021]	95.1	93.1	89.2	91.4
XLSR [Lozhkov, 2022]	93.3	94.4	91.5	86.5
ChatGPT (0-shot)	95.4	97.4	90.0	89.2
ChatGPT (1-shot)	97.9	99.3	96.1	90.5
GMA-SLU w/o RS	96.6 (1.6↓)	98.6 (0.9↓)	95.2 (1.1↓)	91.3 (0.5↓)
GMA-SLU w/o DAT	96.1 (2.1↓)	97.8 (1.7↓)	94.4 (1.9↓)	90.5 (1.3↓)
GMA-SLU w/o DF	96.3 (1.9↓)	98.2 (1.3↓)	94.7 (1.6↓)	90.9 (0.9↓)
GMA-SLU w/o TSTS	96.8 (1.4↓)	98.9 (0.6↓)	95.5 (0.8↓)	91.5 (0.3↓)
GMA-SLU (ours)	**98.2**†	**99.5**†	**96.3**†	**91.8**†

Ergebnisse

Model	en-US	fr-FR	pl-PL	ko-KR
LaBSE [Gerz *et al.*, 2021]	95.1	93.1	89.2	91.4
XLSR [Lozhkov, 2022]	93.3	94.4	91.5	86.5
ChatGPT (0-shot)	95.4	97.4	90.0	89.2
ChatGPT (1-shot)	97.9	99.3	96.1	90.5
GMA-SLU w/o RS	96.6 (1.6↓)	98.6 (0.9↓)	95.2 (1.1↓)	91.3 (0.5↓)
GMA-SLU w/o DAT	96.1 (2.1↓)	97.8 (1.7↓)	94.4 (1.9↓)	90.5 (1.3↓)
GMA-SLU w/o DF	96.3 (1.9↓)	98.2 (1.3↓)	94.7 (1.6↓)	90.9 (0.9↓)
GMA-SLU w/o TSTS	96.8 (1.4↓)	98.9 (0.6↓)	95.5 (0.8↓)	91.5 (0.3↓)
GMA-SLU (ours)	**98.2**†	**99.5**†	**96.3**†	**91.8**†

Agenda

1. Einleitung
2. Transcription, ASR, SLU
3. Spoken Language Understanding
4. GMA-SLU
5. Ergebnisse
6. Fazit

Fazit

- GMA-SLU scheint ein vielversprechender Vorgang zu sein um End-to-End SLU Modelle, durch größere Datensätze zu verbessern
- Verschiedene Sprachmodelle, sowie Methoden haben ebenfalls zu einer Verbesserung der Ergebnisse geführt was die Robustheit des Systems bestärkt
- Es gibt jedoch noch viele offene Fragen vor allem im Bezug auf den Impakt der Sprachmodelle selbst
- Alleine durch diesen Ansatz konnte die Lücke zwischen End-to-End SLU Modellen und traditionellen Verfahren nicht geschlossen werden

- Multimodale Sprachmodelle wurden nicht mit einbezogen in der Evaluation

Quellenverzeichnis

- [1] Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee,Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., de Rosa,G., Saarikivi, O., Salim, A., Shah, S., Wang, X., Ward, R., Wu, Y., Yu, D.,Zhang, C., and Zhang, Y. (2024). Phi-4 technical report.
- [2] Cheng, X. and Zou, Y. (2024). Generating more audios for end-to-end spoken language understanding. In Larson, K., editor, Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24,pages 6234–6242. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- [3] Harper, E., Majumdar, S., Kuchaiev, O., Jason, L., Zhang, Y., Bakhturina, E., Noroozi, V., Subramanian, S., Nithin, K., Jocelyn, H., Jia, F., Balam, J., Yang, X., Livne, M., Dong, Y., Naren, S., and Ginsburg, B. NeMo: a toolkit for Conversational AI and Large Language Models.
- [4] He, M. and Garner, P. N. (2023). Can chatgpt detect intent? Evaluating large language models for spoken language understanding.

Quellenverzeichnis

- [5] He, Z., Wang, Z., Wei, W., Feng, S., Mao, X., and Jiang, S. (2020). A survey on recent advances in sequence labeling from deep learning models.
- [6] Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In Third Workshop on Very Large Corpora.
- [7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. CoRR, abs/1706.03762.
- [8] Wang, Y.-Y., Deng, L., and Acero, A. (2005). Spoken language understanding. IEEE Signal Processing Magazine, 22(5):16–31.