

ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech

R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C.W.
Leong, K.Knill, L. Wang
Fondazione Bruno Kessler, Trento, Italy
Educational Testing Service, Princeton, USA
Cambridge University, Cambridge, UK



Présentation du projet

2ème édition du challenge en reconnaissance vocale, special Session durant l'Interspeech 2021

Interspeech : conférence sur les techniques et application de traitement de la parole. Approche interdisciplinaire

But : faire avancer la recherche de la reconnaissance vocale des enfants non-natifs

Pourquoi ? Le cas des enfants effectuant un test de langue intéressant par la production de disfluences, de code-switching... = taux d'erreur important

Manque de datasets publiques

Nouveauté par rapport à l'édition de 2020 : 1 nouveau dataset pour l'anglais, 1 nouvelle langue : l'allemand



Le challenge

La tâche : Amélioration des performances de l'ASR chez les enfants non-natifs parlant allemand et / ou anglais

Les conditions : 11 jours pour soumettre le modèle à raison d'une soumission par jour.

Modèles à développer en open et closed track

Mode d'évaluation : WER

open track : les participants peuvent ajouter les données qui leur semblent pertinentes pour construire leur modèle.

closed track : les participants ne peuvent utiliser que les données fournies par les organisateurs. Ils n'ont pas non plus le droit de prendre les données de l'anglais pour entraîner le modèle allemand.



Les données

How to participate

The resources provided for the challenge are released through the following five compressed packages:

- [ETLT2021_ETS_EN.tgz](#): transcribed audio data in English provided by ETS;
 - [ETLT2021_FBK_EN.tgz](#): transcribed audio data in English, additional text data and lexica provided by FBK;
 - [ETLT2021_CAMBRIDGE_EN_baseline.tgz](#): Kaldi English baseline, provided by Cambridge University;
 - [ETLT2021_FBK_DE.tgz](#): transcribed audio data in German, additional text data and lexica provided by FBK.
 - [ETLT2021_FBK_DE_baseline.tgz](#): Kaldi German baseline, provided by FBK
- Follow [this link](#) to download the user license for getting ETLT2021_ETS_EN.tgz and send the signed license to "amisra001@ets.org".
 - Follow [this link](#) to download the user license for getting both ETLT2021_FBK_EN.tgz and ETLT2021_FBK_DE.tgz and send the signed license to "falavi@fbk.eu".

After signing the licenses you will receive the links to download the data packages, including baselines.

Table 1: *Some statistics, in hours, about the ETLT2021 challenge.*

	English	German
transcribed train	54 h (ETS) + 49 h (FBK)	5 h (FBK)
untranscribed train	–	64 h (FBK)
dev	3 h (ETS)	1 h (FBK)
eval	3 h (ETS)	1 h (FBK)

Table 2: *Some statistics on the speech data in the package; number of utterances, pupils, different questions, running words, total duration. For ETS data (*), we report only the number of different questions for each speaker (4).*

id	#Utt	#Pup	#Q	#Words	Dur
English					
ETStrain	3200	800	4(*)	173770	53:26
TLT1618train	11700	3111	109	136482	40:11
TLT2017train	2299	338	24	22882	09:00
ETSdev	200	50	4(*)	11671	03:20
ETSeval	200	50	4(*)	11864	03:20
German					
deTLT2017train	1445	296	23	8536	04:42
deTLT1618train	10047	2658	124	–	63:55
deTLT2017dev	339	72	23	2244	01:07
deTLT2017eval	329	72	23	2180	01:07

Table 3: *Some statistics on the text data provided in the package: running words, lexicon size, input modality.*

id	Running Words	Lex Size	mode
English			
2016Wtrain	185777	3385	written
2017train	22450	1493	spoken
German			
de16W18Wtrain	234842	3211	written
deTLT2017train	7590	749	spoken



Données pour l'anglais

2 types de ressources :

- enregistrements avec leurs transcriptions
- ressources pour la construction et l'évaluation de modèles de langues

Anglais :

- 2 sources d'enregistrements, une de ETS , une de FBK

ETS : âge 11 ans et +, venant d'Amérique du Sud, de Corée, du Japon, de Turquie, peu de bruits de fond, 4 questions posées = 4 enregistrements / locuteur. Lecture ou parole spontanée.

FBK : âge 9-16, de la région de Trentino en Italie, différents niveaux de langue (A1,A2,B1), enregistrements faits en 2016, 2017 et 2018. Environ 6 000 étudiants de 4 écoles. 40h transcrit manuellement par ETS (enregistrements 2016 et 2018), 9h pour 2017. Le train est le même que le challenge de l'année dernière

- Construction et évaluation du modèle de langue :

Transcriptions manuelles de FBK 2017

Extraction de phrases produites par les élèves en 2016 (FBK)

Lexiques phonétiques basés sur le dictionnaire CMU + transcription automatique des mots inconnus.

2 lexiques pour les mots italiens et allemands, transcription SAMPA.



Données pour l'allemand

- Uniquement les données FBK (TLT), la majorité des élèves passant le test d'anglais faisant aussi de l'allemand => choix d'avoir les locuteurs dans le même groupe

Transcription beaucoup moins importante : 7h en tout

Beaucoup de bruits de fond = diminution de la quantité utile pour la tâche

- Construction et évaluation modèle de langue

transcription manuelle des audios de 2017

Extraction de phrases produites par les élèves en 2016 et 2018

2 lexiques phonétiques :

Lexique 1 : chaque mot = 1 langue et sa transcription correspondante

Lexique 2 : phones italiens et anglais mis en correspondance avec ceux de l'allemand pour n'avoir que des phones allemands.



Baselines

- Point de départ : Kaldi recipe pour les deux langues



Kaldi c'est quoi ?

- Un ensemble d'outils pour l'ASR écrit en C++ sous la licence Apache v2.0.
- Développé par Daniel Povey et Arnab Ghoshal (Johns Hopkins University)
- Pour l'histoire, selon la légende Kaldi est celui qui a découvert le plant de café
- Outils pour les chercheurs / experts “ In general, Kaldi is not a speech recognition toolkit "for dummies." It will allow you to do many kinds of operations that don't make sense.”





The Kaldi Speech Recognition Toolkit

(Povey, Ghoshal, 2011)

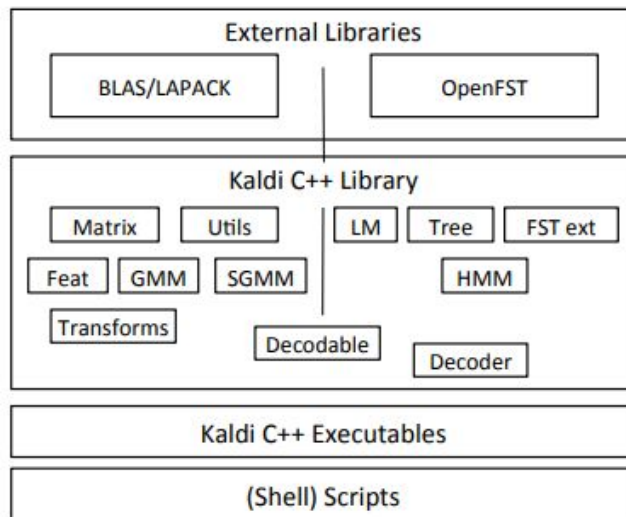


Fig. 1. A simplified view of the different components of Kaldi. The library modules can be grouped into those that depend on linear algebra libraries and those that depend on OpenFst. The *decodable* class bridges these two halves. Modules that are lower down in the schematic depend on one or more modules that are higher up.



Baselines

Baseline anglais	Baseline allemand
<p>AM :</p> <p>Données d'entraînement : 48h du train ETLT_2021_ETS</p> <p>Chain model entraîné avec LF-MMI, composé de 6 couches CNN + 9 couches TDNN-F</p> <p>Données d'entrée : MFCC + i-vecteurs</p>	<p>AM :</p> <p>seed model sur une quantité limitée (-5h de données de réponses transcrites manuellement)</p> <p>Standard chain model avec LF-MMI</p> <p>Utilisation de ce modèle pour modèle final</p> <p>TDNN => données supervisées et non supervisées</p>
<p>LM:</p> <p>4-grammes avec les transcriptions des données d'entraînement de l'AM</p> <p>prononciations phonétiques du CMU avec système G2P entraîné sur Phonetisaurus</p>	<p>LM:</p> <p>n-gramme avec les réponses écrites collectées en 2016 + 2018 et les transcriptions, lexique multilingue pour modéliser les mots qui n'appartiennent pas à l'allemand</p>



Résultats

Table 5: Results achieved by the participants in all tracks of the challenge. The main features of the submitted systems are also listed.

English Closed Track					
Track-Rank	% WER	ASR engine	Acoustic Model	Language Model	System Comb.
EC-1	25.69	Kaldi	TDNNs, CNN-TDNNs	4-grams	YES
EC-2	29.27	Kaldi	CNN-TDNN	word pronunciation	YES
EC-3	29.74	Kaldi	CNN-TDNNs	4-grams + RNNLM resc.	NO
EC-4	31.22	-	-	-	-
EC-5	31.36	-	-	-	-
EC-6	32.21	Kaldi	TDNN + CycleGan augm.	4-grams	NO
EC-7	33.18	Kaldi	-	4-grams	NO
EC-8 baseline	33.21	baseline	baseline	baseline	NO
EC-9	37.05	Kaldi,FAIRSEQ	same as GC-1	n-grams, LM resc.	NO
English Open Track					
Track-Rank	% WER	ASR engine	Acoustic Model, Language Model, System comb.		Additional Data
EO-1	23.98	E2E	transformer, no LM, CTC dec(w=0.5)		native adult
EO-2	29.08	Kaldi	same as EC-2		NO
EO-3	29.58	-	-		-
EO-4	29.63	Kaldi	same as EC-3		NO
EO-5	30.61	FAIRSEQ	WAV2VECT, 4-grams		≈2000h conversational
EO-6 baseline	33.21	baseline	baseline		NO
EO-7	37.05	Kaldi,FAIRSEQ	same as GC-1 (not used all train set)		NO
German Closed Track					
Track-Rank	% WER	ASR engine	Acoustic Model	Language Model	System Comb.
GC-1	23.50	Kaldi,FAIRSEQ	WAV2VECT + transformer	n-grams, LM resc.	NO
GC-2	38.55	Kaldi	CNN+TDNN	4-grams, RNNLM resc.	NO
GC-3	39.98	Kaldi	DNN-BLSTM	4-grams, RNNLM resc.	NO
GC-4	40.04	Kaldi	-	4-grams + graphemic lex.	NO
GC-5	40.51	-	-	-	-
GC-6	40.63	-	-	-	-
GC-7	43.13	Kaldi	same as EC-2	same as EC-2	-
GC-8 baseline	45.21	baseline	baseline	baseline	NO
German Open Track					
Track-Rank	% WER	ASR engine	Acoustic Model, Language Model, System Combination		Additional Data
GO-1	23.50	Kaldi,FAIRSEQ	same as GC-1		NO
GO-2	39.98	Kaldi	same as GC-3		NO
GO-3	40.27	-	-		-
GO-4	40.69	-	-		-
GO-5	40.87	E2E	conformer, no LM, CTC dec(w=0.5)		NO
GO-6 baseline	45.21	baseline	baseline		NO

Performances en open track pour l'anglais légèrement meilleures qu'en closed track. Même score pour l'allemand.



Résultats

Progression de la performance des systèmes soumis comparés à la baseline, particulièrement en allemand (ASR robuste avec peu de données transcrites)

Résultats dans la continuité de ceux du challenge précédent => meilleurs résultats obtenus sans ajout de discours additionnels pour l'entraînement.

Ajout d'une grande quantité de discours d'adultes pas efficace pour cette tâche

Systèmes plus généraux car données en anglais plus diversifiées.


Approche end-to-end qui se révèle efficace, donne les meilleurs résultats pour l'allemand en open et closed track, pour l'anglais en open track

Tous les participants :

- utilisent des contextes temporels longs pour représenter l'audio sous la forme de TDNN / CNN / modèles séquence à séquence
- augmentent les données (pitch, speed, volume perturbation, spectral augmentation)
- (sauf un) utilisent des lexiques et transcriptions automatiques

Certains participants :

- n'ont pas utilisé la totalité des données d'entraînement
- appliquent soit une réévaluation en utilisant un RNNLM, soit un "word lattice combination" ou les deux
- ajoutent des transcriptions qui proviennent de corpus additionnels



The TAL system for the INTERSPEECH2021 Shared Task on Automatic Speech Recognition for Non-Native Children's Speech

Gaopeng Xu, Song Yang, Lu Ma, Chengfei Li, Zhongqin Wu
TAL Education Group, Beijing, China



L'équipe avec le meilleur résultat

TAL education Group, Beijing, China



- Gaopeng Xu, Song Yang, Lu Ma, Chengfei Li, Zhongqin Wu

Objectif :

- Améliorer les performances de l'ASR sur les enfants non-natifs parlant allemand



Challenge

Un discours difficile pour l'ASR en raison de disfluences :

- Mots mals prononcés
- Erreurs grammaticales
- Code-switching
- Mots irréguliers

Différences entre adulte et enfant :

- Langagière : prosodie, acoustique, syntaxe, vocabulaire
- Physiologique (conduit vocal plus court)
- Cognitives : différents niveaux de capacité de cognition langagière



Usages

- Applications pédagogiques
- Médias contenant des voix d'enfants
- Jeux
- Evaluation automatique de l'acquisition de la L2 d'un enfant



Données disponibles

- 64 heures de données non transcrites (TLT1218UNTRStrain)
- 4.6h de données transcrite (TLT2017)
- Augmentation des données par modifications des paramètres :
 - Vitesse
 - Timbre
 - Volume
 - Bruit



Les participants

Groupes	Tranche d'âge	Niveau d'allemand
Groupe 1	9-10 ans	A1
Groupe 2	12-13 ans	A2
Groupe 3	14-16 ans	B1

Tâche :

- Répondre à des questions selon leurs compétences dans la langue



Nettoyage des données

Resegmenter et nettoyer les données d'entraînements :

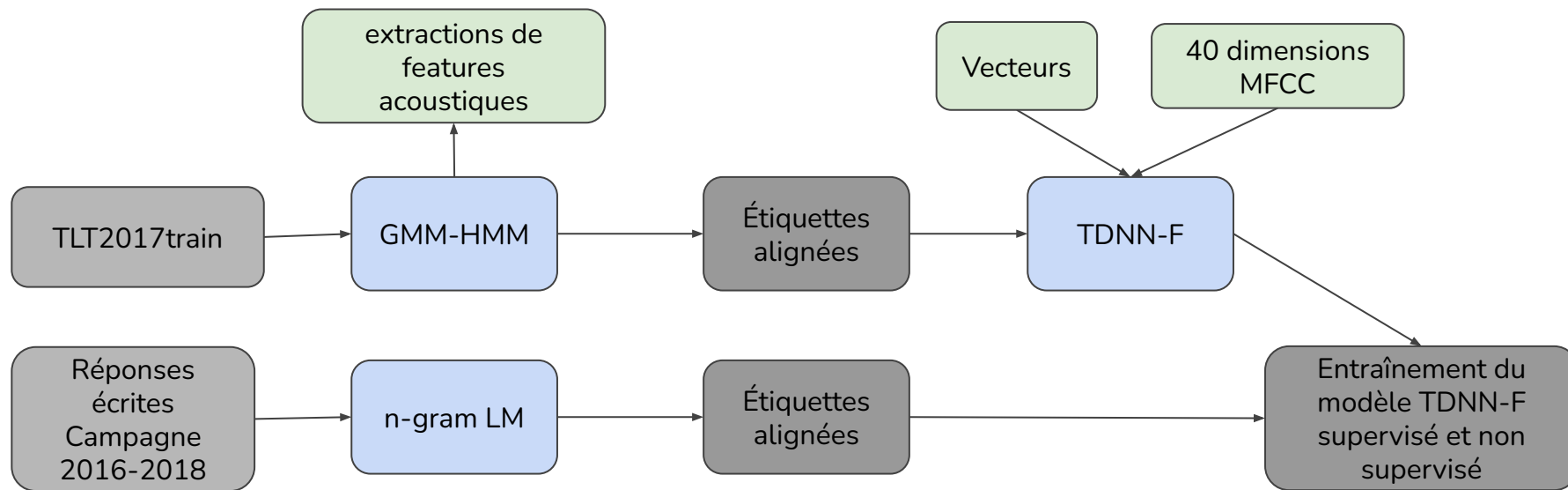
- Pallier aux problèmes de transcriptions
- Supprimer les données ne contenant que du bruits
- Gérer les audios de plus de 100s

Méthode :

- Entraînement d'un modèle chain-based gardant les paires audios transcriptions correctes
- Décodage des données avec un modèle acoustique chain-based pour resegmenter
- VAD (Voice Activity Detection) pour segmenter les données non transcrites



Baseline



GMM-HMM : Gaussian Mixture Model - Hidden Markov Model

TDNN-F : Variation de CNN adapté à la nature séquentielle des données temporelles



Autre baseline

- Wav2vec (non supervisé)
- Remplacement du TDNN-F par un MSA (multi-stream self-attention)
- Transformers



Pré-entraînement

- Utilisation d'un Wav2Vec non supervisé
- BASE model (paramètres : 7 blocs de convolutions, strides (5,2,2,2,2,2,2), kernel widths (10,3,3,3,3,2,2))
- LARGE model → réduction des audios à 30 sec
- optimisation avec Adam



Fine-tuning

- Sur les données transcrites TLT2017train
- Adaptation aux spécificités de l'allemand:
35 tokens utilisés : 30 caractères allemands, 4 symboles d'annotations, 1 token de frontière de mots
- Optimisation du modèle avec Adam

Table 4: % WER of the dev sets for different model with data augmentation (speed perturbation, volume perturbation, reverberation simulation, babble augmentation, music augmentation, noise augmentation and pitch augmentation), data clean-up, VAD segment(vad_seg) and language model

[illegible]



En résumé

- Essai de 6 modèles comme principaux changements :
 - augmentation de données transcrites
 - augmentation de données non transcrites
 - tokenisation avec un VAD
 - utilisation d'un modèle large au lieu du base modèle
- **Modèle le plus performant** : VAD + Large Model transformer avec des 4-gram **WER = 23.50%**
vs WER baseline=43.5%

ASR in German: A Detailed Error Analysis

Johannes Wirth
Research Group System Integration
Hof University of Applied Sciences
Hof, Germany

René Peinl
Research Group System Integration
Hof University of Applied Sciences
Hof, Germany



Objectifs

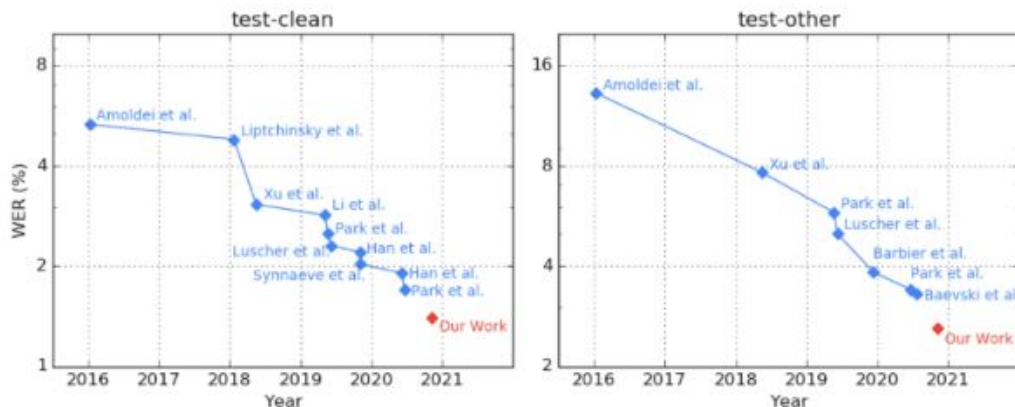
- Évaluer plusieurs modèles d'ASR en allemand et proposer des solutions pour les améliorer en se concentrant sur les types d'erreurs :
 - Lançant les modèles sur un nouveau benchmarks comprenant des données les plus diverses possibles
 - Mettant en place un protocole pour identifier précisément les erreurs



Problème de saturation des benchmarks

Un benchmark est saturé lorsque les performances des modèles sur ce dernier sont si bonnes qu'il est difficile de mesurer les progrès des modèles.

- **Mesures de progrès inefficaces** : une fois qu'un benchmark est saturé, il n'est plus utile pour mesurer ou orienter les progrès des futurs modèles
- **Mesures trompeuses** : Dans un benchmark saturé, même les petites améliorations des performances peuvent sembler importantes, mais elles peuvent ne pas être statistiquement significatives. Cela signifie que ces améliorations pourraient être dues à des variations aléatoires ou à des spécificités du test, plutôt qu'à une véritable avancée dans les capacités du modèle.
- **Non généralisation** : L'optimisation excessive pour des benchmarks spécifiques peut conduire à des modèles qui ne fonctionnent pas aussi bien sur des distributions de données différentes



Évolution du WER sur le benchmarks allemand pour l'ASR Librispeech



Méthodes d'Évaluation traditionnelles pour l'ASR.

- Tous les modèles d'ASR sont généralement évalués selon deux métriques :
 - **WER** (Word Error Rate) qui calcule le nombre d'erreurs dans la transcription textuel (substitutions + suppressions + insertions) / nombre de mots
 - **CER** (Character Error Rate) même chose mais avec les caractères
- C'est pas des métriques assez précises, elles ne disent rien sur la nature ou l'impact de l'erreur ce qui pose problème lorsqu'on veut évaluer/interpréter l'amélioration des modèles



Sélection des modèles

Les modèles retenus dans le cadre de cette étude ont été choisis selon 3 critères principaux :

- **Speaker independent** : Les modèles n'ont pas besoin d'avoir été entraînés sur la voix des locuteurs pour pouvoir fonctionner.
- **Vocabulaire large** : Permet de reconnaître un grand nombre de mots
- **Parole continue et spontanée** : Les modèles sont adaptés aussi bien sur de la parole continue (= préparée, discours/lectures) que sur de la parole spontanée.
- **Disponibilité publique de l'implémentation** : reproductibilité de l'expérience
- Pré-entraînés en allemand



Modèles retenus

- Principalement des modèles fournis par Nvidia qui font partie du NeMo toolkit
- Tous les modèles Nvidia ont été entraînés sur les données en allemand du corpus Mozilla Common Voice, MultiLingual LibriSpeech et VoxPopuli (sauf Quartznet)
- W2V2 et Quartznet ont été pré-entraînés sur de l'anglais et finetunés sur Mozilla Common Voice

TABLE II. ASR MODELS WITH TRAINING DATA EVALUATED

Model	Datasets for training
Citrinet	German MCV 7, MLS, VoxPopuli
ContextNet	German MCV 7, MLS, VoxPopuli
Conformer CTC	German MCV 7, MLS, VoxPopuli
Conformer Transducer	German MCV 7, MLS, VoxPopuli
Wav2Vec 2.0	Pretraining on MLS EN, finetuning on MCV6
Quartznet	Pretraining on 3000h EN, finetuning MCV6



Évaluation à partir du WER

- Pour commencer, les modèles ont été évalués sur différents corpus avec une simple mesure de WER.
- Corpus en tout genre où se mêlent parole spontanée / préparée

TABLE III. DATASETS USED IN EVALUATION

dataset	hours (all test)	speakers	type	len (min/ø/max)
MCV 7.0	965 26.8	15,620	read	1.3/6.1/11.2s
Tuda	127 11.9	147	read	2.5/8.4/33.1s
SWC	285 9.1	363	read	5.0/7.9/24.9s
M-AILabs	237 10.8	29	read	0.4/7.2/24.2s
MLS	3287 14.3	244	read	10/15.2/22s
VoxForge	35 2.7	180	read	1.2/5.1/17.0s
HUI	326 16.3	122	read	5.0/9.0/34.3s
Thorsten	23 1.1	1	read	0.2/3.4/11.5s
VoxPopuli	268 4.9	530	spoken	0.6/9.0/36.4s
Bundestag	604 5.1		spoken	5.0/7.2/35.6s
Merkel	1.0	1	spoken	0.7/6.9/17.1s
TED Talks	16 1.6	71	spoken	0.2/5.1/118s
ALC	95 2.6		read	2.0/12.5/62s
BAS SI100	31 1.8	101	read	2.1/12.7/54.1s

Analyses

- Le meilleur modèle est Conformer Transducer sur tous les corpus et le moins bon est Quartznet
- On peut voir que Citrinet, Conformer CTC et ContextNet ont des performances très similaires (moins de 0,5% de diff entre leurs WER)
- Citrinet et Quartznet ont des écarts moindres entre leur WER ce qui indique une plus grande robustesse
- Pour les données politiques (VoxPopuli, Bundestag et Merkel) les modèles Conformer et Contextnet sont meilleurs ce qui est sûrement dû aux données sur lesquelles ils ont été entraînés

Mais quelles sont les natures de ces erreurs et quels sont leurs impacts ?

TABLE IV. WORD ERROR RATES FOR ALL MODELS AND ALL DATASETS

	Citrinet	Conf. CTC	Conf. T	Contextnet	Wav2Vec 2.0	Quartznet
MCV 7.0	8.78%	8.00%	6.28%	7.33%	10.97%	13.90%
Bundestag	13.25%	13.65%	11.16%	14.44%	21.78%	28.61%
VoxPopuli	10.35%	10.82%	8.98%	10.13%	21.96%	28.34%
Merkel	13.63%	17.17%	13.49%	15.92%	21.81%	27.57%
MLS	5.56%	5.16%	4.11%	4.62%	13.04%	20.34%
MAI-LABS	5.52%	5.56%	4.28%	4.32%	9.94%	18.47%
Voxforge	4.15%	3.95%	3.36%	4.16%	5.64%	7.58%
HUI	2.31%	2.45%	1.89%	2.02%	8.52%	14.66%
Thorsten	6.74%	8.49%	6.20%	9.21%	7.57%	5.95%
Tuda	9.16%	7.81%	5.82%	7.91%	12.69%	20.31%
SWC	10.15%	9.36%	8.04%	9.29%	15.01%	16.49%
German TED	34.53%	35.77%	31.98%	35.58%	41.90%	47.75%
ALC	31.42%	31.30%	25.90%	26.85%	40.94%	45.53%
BAS SL100	23.13%	24.81%	22.82%	22.74%	28.84%	28.94%
Average	12.76%	13.16%	11.02%	12.47%	18.62%	23.17%
Median	9.65%	8.93%	7.16%	9.25%	14.02%	20.33%



Identification des erreurs

- **Création de sets de différences** pour identifier les sources d'erreurs spécifiques à chaque modèle et à son architecture
- **Identification d'erreurs indépendantes aux modèles** : intersections des transcriptions incorrectes prédites pour tous les modèles. Ces erreurs peuvent indiquer des problèmes relatifs aux données et pas aux modèles eux-mêmes
- **Comparaison des vocabulaires** utilisés dans les ensembles de données d'entraînement et des vocabulaires des données de test pour vérifier la capacité des modèles à généraliser à des mots jamais rencontrés (OOV)



Annotation manuelle des erreurs

Parmi les erreurs communes à tous les modèles, 2000 échantillons ont été sélectionnés et ont été étiquetés :

- Erreurs négligeables
- Erreurs mineures
- Erreurs majeures
- Noms propres, emprunts, anglicismes
- Homophones
- Mauvaises transcriptions
- Enregistrements ambigus
- Enregistrements de mauvaise qualité

TABLE VII. ERROR CLASSIFICATION AND PROPORTIONS

Nr	Error Category	Error Proportion
1	Negligible	9,40%
2	Noncontext-Breaking	11,95%
3	Context-Breaking	19,01%
4	Name, Anglicism, Loan Word	19,82%
5	Homophone	2,92%
6	Flawed Ground Truth Transcript	17,85%
7	Ambiguous Audio Input	11,13%
8	Flawed Audio Input	7,91%



Annotation manuelle des erreurs (Cas particuliers)

- **Les erreurs négligeables et les mauvaises transcriptions** (27.25 %) : les modèles seraient bien meilleurs si on en tenait pas compte
- **Noms propres, emprunts, anglicismes** (~ 1/5) : De formes alternatives (noms) jusqu'à des transcriptions incompréhensibles ; une catégorie où l'erreur humaine est elle-même importante
- **Enregistrement ambigu** (11.13 %) : Surtout dû à des locuteurs L2, des dialectes ou des erreurs de prononciation ; dans l'écrasante majorité des cas, les humains n'arrivent pas non plus à les annoter



Annotation manuelle des erreurs

(Erreurs majeures)

- **Problème de normalisation** : Certains chiffres (notamment les années) ont été écrits en chiffre => la normalisation les retranscrits sous une forme particulière (“neunzehnhundertdreiundsechzig” “dix-neuf cent soixante trois”) quand la prononciation les prononcent autrement (“eintausendneunhundertdreiundsechzig” “mille neuf cent soixante trois”)
- **Transcription indirecte** : Certaines transcriptions de l’entraînement contiennent des transcription dans lesquelles un mot a été remplacé par un autre ; ces erreurs peuvent changer le sens de la phrase, elles ne le font pas systématiquement
- **Structure constante** : (Problème du dataset VoxPopuli) La transcription contient une structure (notamment commençant par “Herr Präsident” au début) qui n’est pas prononcée
- **Audio mal découpés** (“Hallucination”) : L’audio commence au milieu d’un mot/d’une phrase => les modèles essayent de prédire le début des phrases, souvent incorrectement



OOV et silence (Thorsten)

- **OOV** : $\approx 50\%$ des OOV sont généralisés correctement
- **Thorsten et problèmes de silences** : Le dataset Thorsten a une particularité, il dispose de peu de silence avant/après ses audio \Rightarrow ajouter 0.3s de silence améliorent les résultats des modèles sur ce dataset

	original	with silence	delta
Citrinet	6.74%	4.09%	-2.65%
Conformer CTC	8.49%	3.96%	-4.53%
Conformer Trans	6.20%	3.61%	-2.59%
Contextnet	9.21%	3.35%	-5.86%
Wav2Vec 2.0	7.57%	5.11%	-2.46%
Quartznet	5.95%	5.00%	-0.95%



Solutions proposées

- **Normalisation** : Vérifier la (correcte) normalisation des datasets
- **Extension du vocabulaire** : Étendre à des domaines spécifiques (en utilisant du TTS)
- **Phonème to caractères** : Mapper des phonèmes aux caractères ; Déterminer les homophones, mieux traiter les audio de locuteurs L2 et des dialectes
- **Prétraitement des données audio** : Dialogues => séparation des locuteurs ; Enregistrements de mauvaise qualité => amélioration automatique des données audio => ajoute du temps de calcul et donc dégrade le temps mis par le modèle ASR pour transcrire



Conclusion

- **But** : Analyser les erreurs de différents modèles d'ASR en allemand => proposer des solutions
- **Erreurs** : Différents types, certaines moins graves (négligeables, problème de transcription dans les datasets) mais aussi des problèmes plus complexes (emprunts et noms, enregistrements de mauvaise qualité, perte de sens)
- **Solutions** : Normalisation ; Extension du vocabulaire (TTS) ; Phonèmes vers caractères ; Prétraitements des audio



Références

<https://vitrinelinguistique.oqlf.gouv.qc.ca/24507/la-prononciation/notions-de-base-en-phonetique/les-phones-et-les-phonemes>

<https://kaldi-asr.org/doc/about.html>

https://www.danielpovey.com/files/2011_asru_kaldi.pdf

<http://themarvinproject.free.fr/final/node4.html>

<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>