



01 WHY ANONYMIZE SPEECH?

Shared speech can expose who is speaking.

Voice identity can enable unauthorized speaker identification. Anonymization must hide identity while retaining the linguistic content and emotional expression that make speech useful.

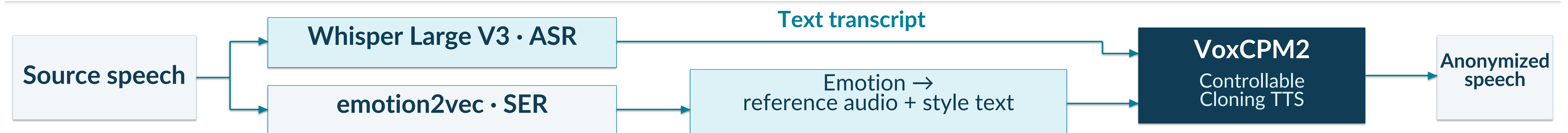
Goal: preserve content + emotion; suppress identity.

02 WHAT MAKES THIS DIFFICULT?

- Signal processing methods mostly based on mathematical rules which can be easily inverted.
- Voice conversion (VC) use unmodified source content features which can leave cues to attacker that has been pretrained on a large VC dataset.
- Cascaded ASR-TTS use transcriptions which mitigate VC problems but cannot recover emotion from the source speech.

Our approach: Follow the cascaded ASR-TTS paradigm, but adopt an emotional TTS using emotion predicted from a Speech Emotion Recognition (SER) model.

03 METHOD

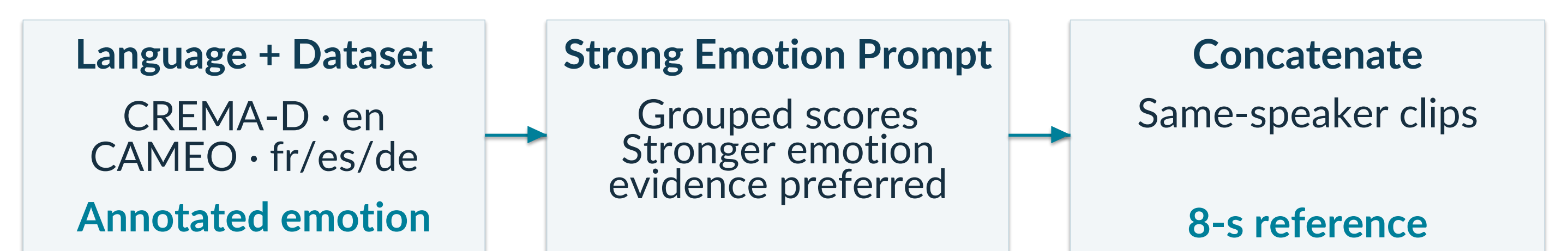


04 HOW IS THE EMOTION CHOSEN?

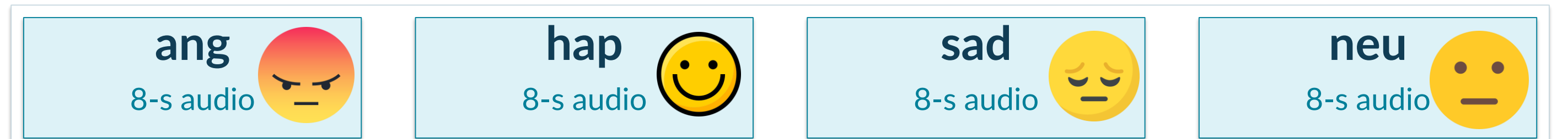
Sum the raw prediction score of emotion2vec within each group, and select the group with the highest sum.

Label	Raw labels	TTS style prompt
ang	angry, disgusted	furious, shouting, intense voice
hap	happy, surprised	very happy, excited, cheerful voice
sad	sad, fearful	deeply sad, sorrowful, trembling voice
neu	neutral, other, unknown	neutral, calm, flat voice

05 HOW IS THE PROMPT POOL BUILT?



Prompt pool: one reference per emotion / language



Source language + predicted emotion → retrieve audio

06 TRACK 1 · ENGLISH

Test (%) · LibriSpeech: EER/WER; IEMOCAP: UAR
Semi-informed ASV attacker

System	WER ↓	UAR ↑	EER ↑
Original	1.84	71.06	3.41
B2	9.11	53.52	3.72
B3	4.26	37.70	14.24
B4	5.90	42.78	11.13
B5	4.53	38.09	17.05
Ours	3.25	55.78	31.08

Privacy under the semi-informed attacker

B5  17.05%

Ours  31.08%

+14.03 percentage points EER vs B5

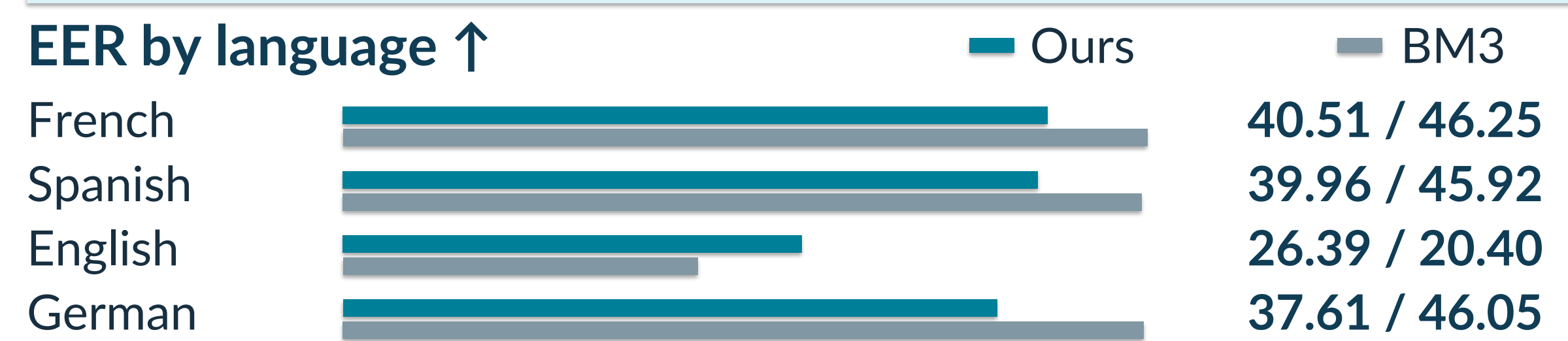
Best test EER, WER and UAR among compared anonymizers.

07 TRACK 2 · MULTILINGUAL

Evaluation (%) · MLS: EER/WER; EmoTrack2: UAR
French / Spanish / English / German

System	WER ↓	UAR ↑	EER ↑
Original	4.93	78.90	4.25
BM1	8.53	42.63	2.10
BM2	27.74	29.04	10.62
BM3	9.98	24.84	39.65
Ours	9.39	62.22	36.12

EER by language ↑



WER / EER: four-language averages. UAR: EmoTrack2.
Highest emotion UAR; average privacy below BM3.

08 INTERMEDIATE ANALYSIS: BEFORE → AFTER ANONYMIZE

ASR WER (%) ↓

Dataset	Original	Anonymized	Change
LS	2.65	3.25	+0.60
MLS-fr	6.43	10.91	+4.48
MLS-es	5.46	7.41	+1.95
MLS-en	6.51	7.66	+1.15
MLS-de	8.98	11.56	+2.58

Emotion UAR (%) ↑

Dataset	Original	Anonymized	Change
IEMOCAP test	84.26	55.78	-28.48
EmoTrack2 eval.	80.40	62.22	-18.18

09 WHAT DOES EMOTION CONDITIONING ADD?

Track 1 test ablation (%) ·

Audio prompt	Text	WER ↓	UAR ↑	EER ↑
Base emotion	Emotion	3.25	55.78	31.08
Large emotion	Emotion	3.75	55.31	32.39
Base emotion	None	3.11	51.16	33.30
All-neutral	Emotion	3.35	36.49	33.58
All-neutral	None	3.03	32.51	33.04

Text conditioning adds 4.62% in UAR (55.78 vs 51.16).

All variants use the attacker trained on the base configuration.

10 DICUSSION

Privacy discussion for TTS approach

Since speech is synthesized solely from text transcriptions, the expected EER should approach 50% (random guess).

Hypothesis: The difference from 50% may stem from ASV fine-tuning overfitting or content leakage.

Evaluating TTS for Speaker Anonymization

TTS based method might be optimal with current benchmark, however TTS discards original prosody in favor of its training distribution, making it unsuitable for prosody-critical domains (e.g., healthcare diagnostics). Therefore, prosody preservation metric should be included depending on application.