ASSESSING EVALUATION METRICS FOR SPEECH-TO-SPEECH TRANSLATION

Elizabeth Salesky[▽] *Julian Mäder*[△] *Severin Klinger*[△]

[▽]Johns Hopkins University, USA
[△]ETH Zürich, Switzerland

ABSTRACT

Speech-to-speech translation combines machine translation with speech synthesis, introducing evaluation challenges not present in either task alone. How to automatically evaluate speech-to-speech translation is an open question which has not previously been explored. Translating to speech rather than to text is often motivated by unwritten languages or languages without standardized orthographies. However, we show that the previously used automatic metric for this task is best equipped for standardized high-resource languages only. In this work, we first evaluate current metrics for speech-to-speech translation, and second assess how translation to dialectal variants rather than to standardized languages impacts various evaluation methods.

***Index Terms*—** evaluation, speech synthesis, speech translation, speech-to-speech, dialects

1. INTRODUCTION

Speech is a more natural modality than text for unwritten languages [1] and predominantly spoken languages or dialects without standardized orthographies, motivating work on speech-to-speech translation [2, 3, 4, 5, 6, 7, 8]. Generating speech rather than text translations may sometimes also be desired for commercial translation applications. For under-resourced speech and language processing tasks, using machine translation (MT) before synthesis can be beneficial to take advantage of larger annotated resources in a related high-resource language.

Evaluating translation combined with speech synthesis is an open question. Speech synthesis is most commonly evaluated through costly human judgments in absence of high-quality automatic metrics; development of direct assessment methods through comparison with reference speech is ongoing but not yet common [9, 10]. For applications combining translation with synthesis such as speech-to-speech (S2S) or text-to-speech translation (T2S), previous work has exclusively transcribed synthesized speech with ASR to evaluate with the text-based metric BLEU [8, 11, 12], in part due to the absence of datasets with parallel speech. The appropriateness of this evaluation approach has not yet been studied.

While combining ASR with BLEU provides a metric to enable development of neural speech-to-speech and text-to-speech translation, it relies on standardized orthographies and sufficiently high-quality ASR models that introduced errors will not change system judgements; this is unlikely to be the case for any but the highest-resource settings. The validity of this approach, and its appropriateness for and robustness to other settings, such as dialects or unwritten languages, has not yet been studied. In this study we address two research questions:

- **RQ1**: How well do different metrics evaluate the combination of translation with speech synthesis? (i.e., how well do they correlate with human judgments)
- **RQ2**: How appropriate are metrics for target languages without standardized orthographies? (here, dialects)

We compare human evaluation to reference-based metrics using reference speech or text. We evaluate translation and synthesis for High German and two Swiss German dialects. Our findings suggest BLEU is not appropriate for evaluating speech-to-speech translation for high-resource languages or non-standardized dialects, and while character-based metrics are better, improved metrics are still needed.

2. METRICS

2.1. Human Evaluation: Mean Opinion Score (MOS)

Mean opinion score (MOS) is an aggregate subjective quality measure, and is the most common metric used to evaluate speech synthesis in the absence of high-quality automatic metrics. To evaluate MOS, we randomly selected a fixed held-out set of 20 samples which cover the full character-level vocabulary of each dataset. Each sample is rated for its overall quality by three native speakers along a 5-point Likert scale, where higher values are better.

While for TTS, the *naturalness* of the synthesized voice is the most salient point to evaluate, when synthesizing translations, the translation quality is also significant. The intended meaning may not be understandable due to translation errors and also pronunciation mistakes, each of which influence our perception of the other. To attempt to isolate the individual

dimensions of performance for this task, we ask annotators for subjective measures for three specific categories [13]:

2.1.1. Adequacy

In this category we asked annotators to evaluate how well the meaning of the source sentence was conveyed by the synthesized translation. If information was lost, added, or distorted in some way, this would be assessed here. When translating to speech, errors in pronunciation may also affect the ability to understand the meaning of the sentence.

2.1.2. Fluency

To evaluate fluency, we asked annotators to focus on the cohesion or flow of the synthesized translation. Errors in, for example, grammatical correctness or use of unnatural or archaic word choices would factor here.

2.1.3. Naturalness

For naturalness we asked annotators to focus on the quality of the synthetic voice and the appropriateness of pronunciation for the particular language and dialect. This evaluation is designed to be independent of the correctness of the translation; for example, an incorrectly translated word synthesized with a natural voice should be rated higher here than a correctly translated word synthesized unnaturally or artificially.

2.2. Reference Text: ASR Transcription with MT Metrics

Machine translation metrics compare discrete text representations against references. To evaluate synthesized speech translations with standard automatic MT metrics, previous work on neural speech-to-speech translation [8, 11, 12] has utilized large ASR models trained on hundreds of hours of external corpora in the target language or commercial transcription services to transcribe synthesized samples for comparison against text references. The use of high-quality external models is to prevent the introduction of ASR errors which may impact the downstream MT metric.

Previous work has evaluated using ASR and BLEU only [14] and have experiments with high-resource languages with standardized orthographies only; however, language dialects often have non-standardized orthographies which we show disproportionately affect word-level metrics like BLEU. With this in mind, we also compare two character-level MT metrics. chrF [15] computes F1-score of character n-grams, while character-level BLEU (charBLEU) computes BLEU on character rather than word sequences. We use SacreBLEU [16] to calculate both BLEU and chrF scores.

2.3. Reference Speech: Mel-Cepstral Distortion (MCD)

Mel-Cepstral Distortion [9] is an objective metric for evaluating synthesized speech, given reference audio, which computes the mean distance between two sequences of cepstral features. To account for differences in phone timing and sequence length, dynamic time warping (DTW) [17] is used to align the two sequences. Alternatively, segmental approaches may synthesize test utterances using 'gold' phone durations from the original speech, such that the audio does not need to be aligned in time. In this study we use MCD with DTW.

3. EXPERIMENTS

3.1. Data

3.1.1. German.

For our German (DE) audio corpus, we used the German subset of CSS10 [18], a collection of single-speaker speech datasets for each of ten languages. The German data is composed of 16 hours of short audio clips from LibriVox audiobooks [19] with aligned transcripts.

3.1.2. Swiss German.

For our Swiss German corpus we used SwissDial [20], a collection of single speaker recordings across 8 Swiss German dialects. Each dialect has 2-4 hours of speech with aligned transcripts and sentence-level text translations to High German. This enables us to train both Swiss German synthesis models and translation models to and from German. In this study we focus on Berndeutsch (CH-BE) and Zürichdeutsch (CH-ZH).

3.2. Audio Format

Audio for all experiments used a sampling rate of 22050kHz with pre-emphasis of 0.97. Audio spectrograms were computed with Hann window function with frames of size computed every 12.5ms. MCD was computed using 34 Mel-cepstral coefficients extracted with SPTK [21].

3.3. Models

3.3.1. Machine Translation

We train Transformer [22] models for machine translation using FAIRSEQ [23] following recommended hyperparameter settings from previous work [22] for IWSLT'14 En–De settings. We additionally train multilingual many-to-many models to translate to and from the 8 Swiss German dialects in SwissDial and High German by prepending language ID tags to each sentence and force-decoding the first token [24]. Initial experiments showed character-level models were more

appropriate for these language pairs than subwords, so our final models use character-level tokenization via SentencePiece [25]. We decode with length-normalized beam search with a beam size of 5, and applied early stopping during training using validation sets constructed from 10% of the data; model parameters for evaluation were taken from the checkpoint with the best validation set BLEU.

3.3.2. Speech Synthesis

For speech synthesis we compare two different model architectures. First, we combine Tacotron [26] and WaveNet [27] for a neural 'end-to-end' approach. Tacotron is an end-to-end model which uses a combination of convolutional, fully connected, and recurrent layers to generate (mel) spectrograms from text. The WaveNet vocoder is then responsible for waveform synthesis to generate playable audio. Given the lower-resource nature of this task, we additionally compare a segmental speech synthesis model as implemented by SlowSoft [28, 29] to the 'end-to-end' approach.

3.3.3. Speech Recognition

To transcribe synthesized samples for evaluation with MT metrics we compared two commercial systems, Recapp and Amazon Transcribe, both of which support German and Swiss German.[1] We observed better results with Recapp, which we use throughout. Note that the limited size of the SwissDial corpus would be insufficient to train a high-quality ASR system for Swiss German, exhibiting a common tradeoff between domain and dialect specificity on one hand, and better trained but less suited models on the other.

4. RESULTS

We first show results in § 4.1 on the individual subtasks machine translation (MT) and speech synthesis (TTS), and provide context for our particular languages and metrics. We then present our text-to-speech (T2S) translation results in § 4.2, where we discuss our two research questions: how well different metrics capture model quality (§ 4.2.1) and whether we observe differences across dialectal variants (§ 4.2.2).

4.1. Individual Subtasks

4.1.1. Machine Translation.

We compare machine translation to a high-resource language (German: DE) and to dialectal variants (Swiss German: CH-BE and CH-ZH). Table 1 shows performance on our three MT metrics: BLEU, character-level BLEU, and chrF. We see similar performance across all three languages and metrics for machine translation in isolation. We note SwissDial transcripts were produced by one annotator per dialect: for this reason, dialectal spelling and orthography is more consistent than is typical across speakers and datasets for Swiss German, a common challenge when evaluating dialectal models.

		CH→**DE**	DE→**CH-BE**	DE→**CH-ZH**
Text (↑)	*chrF*	0.8	0.8	0.8
Text (↑)	*charBLEU*	81.2	82.5	82.7
Text (↑)	*BLEU*	45.3	40.2	44.5

Table 1: Machine translation (MT): automatic evaluation metrics for text-to-text baselines.

4.1.2. Speech Synthesis.

We evaluate our TTS models with the same metrics for T2S models to compare the effects of combination with machine translation. Table 2 shows our neural TTS models evaluated with human judgments (MOS), speech-reference metrics (MCD), and text-reference metrics (ASR-chrF, ASR-charBLEU, ASR-BLEU). The SwissDial dataset does not have reference speech for High German, only text, so we cannot compute MCD for DE.

Human judgments for all categories are shown in Figure 1. We see proportionally higher ratings for fluency with the neural models than the segmental models, a well-known phenomena [30]. The segmental model performed most similarly to the neural model in adequacy, the category which synthesis errors are least likely to affect. The neural synthesis model was consistently significantly better than the segmental, despite less training data, and so we conduct our analysis on our neural models.

TTS models for both dialects exhibit similar trends across MOS categories, with consistently slightly higher average judgments for CH-ZH, approaching those of DE. Better performance on CH-ZH may reflect the slightly closer relationship between Zürichdeutsch (CH-ZH) and High German (DE); concretely, this relationship may enable better transfer through pretraining on High German.

Table 2 shows similar MCD scores for both dialects, despite higher MOS for CH-ZH. Computing 95% confidence intervals through bootstrap resampling [31] shows that MCD scores on CH-ZH are slightly more stable than CH-BE.

Text-based metrics depend on ASR transcription to apply. German is an ideal case for these metrics: it is a standardized high-resource language, and the ASR model used is of commercial quality. As such, we treat it as an upper bound for these metrics. For German, the synthesis and ASR roundtrip yields near-perfect[2] character-level metrics (chrF and charBLEU), while occasional errors in synthesis or transcription more greatly affect word-level BLEU.

[1] Our experiments suggest Amazon Transcribe's Swiss German model may support German as spoken in Switzerland, rather than Swiss German.

[2] It has been shown [32] that ASR applied to synthesized speech has higher performance than on natural speech.

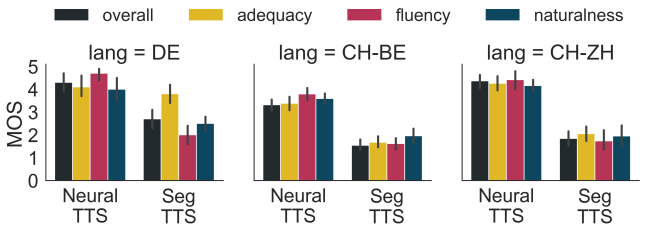

Fig. 1: **MOS: Speech Synthesis.**

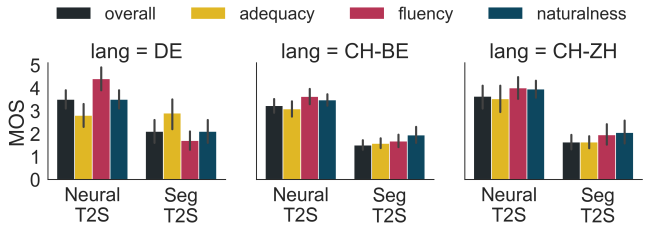

Fig. 2: **MOS: Translation with Synthesis.**

		DE	**CH-BE**	**CH-ZH**
Human (↑)	*MOS*	3.8 ± 0.1	3.3 ± 0.26	3.7 ± 0.26
Speech (↓)	*MCD*	—	6.1 ± 0.42	6.1 ± 0.35
Text (↑)	*ASR-chrF*	0.9	0.4	0.5
Text (↑)	*ASR-charBLEU*	94.9	50.3	58.2
Text (↑)	*ASR-BLEU*	75.2	1.4	3.6

Table 2: Speech synthesis (TTS): evaluation metrics for speech synthesis baselines.

		CH→**DE**	DE→**CH-BE**	DE→**CH-ZH**
Human (↑)	*MOS*	3.8 ± 0.15	3.2 ± 0.31	3.3 ± 0.29
Speech (↓)	*MCD*	—	6.3 ± 0.45	6.2 ± 0.34
Text (↑)	*ASR-chrF*	0.7	0.4	0.5
Text (↑)	*ASR-charBLEU*	78.1	46.5	55.8
Text (↑)	*ASR-BLEU*	41.5	0.9	3.7

Table 3: Translation and synthesis (T2S): evaluation metrics for combined text-to-speech models.

The text-reference metrics show a greater difference between both dialects and between the dialects and High German than either MOS or MCD do. The question we address in this work is whether the differences shown by some metrics reflect better sensitivity to model differences we care about, or, if those metrics are less appropriate for this task. Lower text-reference scores for dialects scores in part reflect slightly lower-quality synthesis models (seen in lower average MOS scores than DE); however, while the character-level metrics chrF and charBLEU are weakly correlated with MOS judgments for the two dialects, BLEU itself is not (see Figure 6). BLEU scores decrease from 75.2 for DE to 1.4 and 3.6 for CH-BE and CH-ZH—which do not reflect the similar human judgments for adequacy across all three language variants.

While we use commercial ASR systems for Swiss German, one of which has been developed specifically for Swiss dialects (Recapp), there is no standardized spelling or orthographic convention for Swiss German [33]. For example, in our dataset, *Scheintüren* may be written as *Schiitüre*, *Schiintüürä*, *Schiporte*, among others. Due to this, the ASR step to apply text-reference metrics introduces a model which has likely been trained on different spelling conventions than in the SwissDial dataset, and can cause greater divergence from text references. We discuss this issue further in § 4.2.2.

4.2. Translation with Synthesis

Human judgments for the text-to-speech translation with synthesis task (T2S) across all categories are shown in Figure 2, and scores for all metrics are shown in Table 3. Our analysis is primarily conducted using the neural synthesis models. Correlations between metrics and human judgments can be found in Figure 6.

We expect to see differences from the individual tasks when combining translation with synthesis. Interestingly, though, human judgments suggest the combination with machine translation does not significantly impact overall synthesis quality (Figure 3). When compared to synthesis alone (Figure 1) we see only small degradations across MOS categories between TTS and T2S, with more similar results for both dialects. The category most affected by the introduction of translation is adequacy, which assesses how well the intended meaning was conveyed, where MT is most likely to introduce impactful errors. Fluency assesses the coherence of the language, and naturalness the quality of the voice and appropriateness of pronunciation. Both categories may be affected by character substitutions introduced by MT which do not fit the phonotactics of the language, and so more greatly affect synthesis than e.g., rare sequences.

Below we assess the ability of the automatic metrics to capture human-perceptible differences in quality for high-resource languages and dialectal variants.

4.2.1. RQ1: Do metrics capture human quality judgments?

Character-level metrics (chrF, charBLEU) have stronger correlations with human judgments than BLEU. The text-reference metrics show the biggest degradation between TTS and T2S. As shown in Figure 4 and Figure 5, ASR-chrF has stronger correlations with overall human judgments for all languages than ASR-BLEU. We also find correlations of *r=0.98* between chrF and charBLEU after ASR, with slightly higher correlations with MOS for chrF, suggesting character-level metrics are more robust to the synthesis-ASR roundtrip. These metrics also reflect more specifically targeted MOS adequacy ratings, which were highly correlated with overall

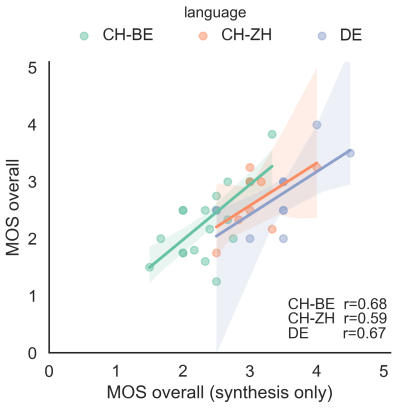

Fig. 3: **MOS T2S—MOS TTS.**

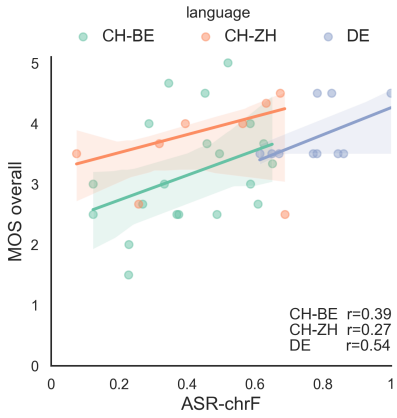

Fig. 4: **MOS overall—ASR-chrF.**

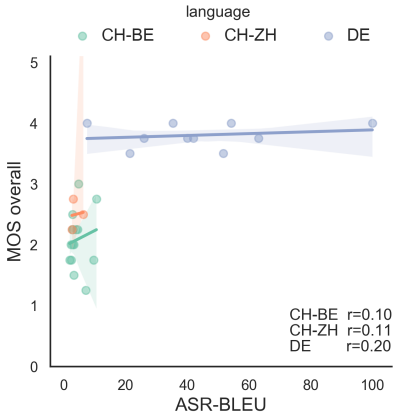

Fig. 5: **MOS overall—ASR-BLEU.**

Correlations with Human Judgments (MOS).

ratings. BLEU had only weak correlations with human judgments (overall and adequacy: *r=0.1-0.2*). This reflects that sentences which had large variation in BLEU scores received similar (high) human judgments. This follows comparisons in [15], where chrF had higher correlations than its word-based equivalent for segments with higher human ratings. We discuss language-specific metric effects in § 4.2.2.

MCD reflects naturalness, which text-based metrics cannot directly assess. By transcribing synthesized speech to text, we lose the ability to assess key speech-specific characteristics such as naturalness. MCD directly compares synthesized to reference speech, and moderately reflects naturalness (correlations for neural models in Figure 6); correlations are negative as these metrics have inverted scales. Reference-text metrics can assess naturalness only implicitly. While character-based metrics in particular also correlate moderately, this instead reflects other confounding factors. When we compare the segmental and neural synthesis models, the ranges of the human naturalness judgments are completely disjoint, as are the MCD scores'; however, the reference-text metrics are unable to discriminate between the models.

Character-level metrics have stronger correlations with MCD. MT models with better character-level performance have better MCD scores. Character-level metrics for MT correlate moderately with MCD↓: we see Pearson's *r=-0.43* (chrF↑) and *r=-0.48* (charBLEU↑); negative correlations are expected given the polarity of the metrics. As an edit distance metric which uses DTW to match segments of approximately phones, which are more similar in granularity to characters than words, this finding is unsurprising but encouraging for the use of implicit character-level metrics if only text rather than speech references are available.

4.2.2. RQ2: Are metrics equally able to evaluate dialects?

Table 3 shows that some metrics exhibit greater differences between TTS and T2S than others. In some cases, this may reflect useful and complementary information: for example, that sample adequacy has been more affected by translation. However, for many metrics, we find utility to be language-dependent.

Character-level metrics are more robust to dialectal targets. Most metrics do not reflect human judgments for dialects. Figure 7 and Figure 8 show correlations between MT metrics applied before and after synthesis (and ASR). While both BLEU and chrF are strongly correlated (*r>0.8*) between both conditions for DE, there is a stark contrast for the dialects. Where chrF shows moderate correlations, BLEU exhibits weak-to-negative correlations for both dialects. Given that human judgments remain correlated with and without translation, this suggests a failure of the metrics to appropriately reflect the differences between models.

ASR disproportionately affects metrics for dialects. Few dialects have standardized spelling conventions; one technique to train models for dialects is to create a normalized representation [34, 33] to remove such variation, which has been created for these ASR models, or even to 'transcribe' to a standardized variant such as High German [35]. In many cases, this is less transcription than translation, as there can also be variation in word order and grammar in addition to pronunciation between language variants. Further, many dialects do not have normalization standards, so normalized forms may have been created on an individual level and so still differ between models and corpora. While normalization may improve model performance for some tasks, then, it also creates a diglossia where the text may not reflect the speech, which may degrade others (like evaluation metrics).

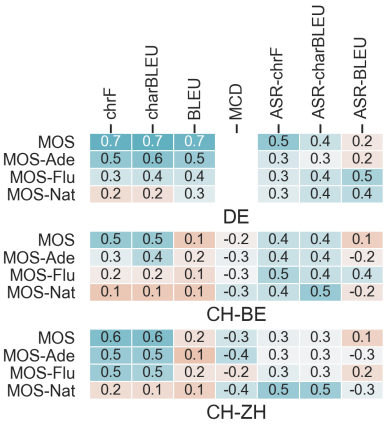

Fig. 6: **Pearson's *r* correlations.**

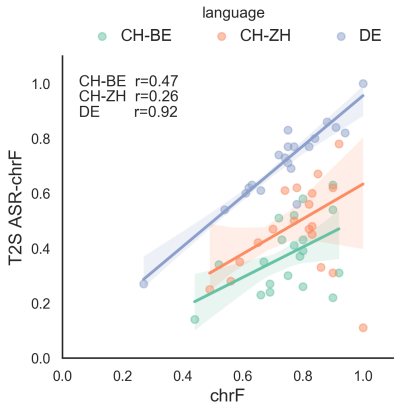

Fig. 7: **chrF—ASR-chrF.**

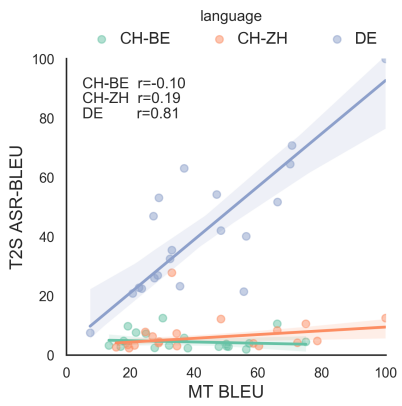

Fig. 8: **BLEU—ASR-BLEU.**

Correlations of text-based metrics before synthesis, and after synthesis and transcription (ASR).

The use of ASR for evaluation can introduce a model which has been trained on different spelling and orthographic conventions than the translation and synthesis models to be evaluated. This means that ASR will yield different spelling standards than the reference text, artificially reducing text-metric scores. Our test sets are parallel between all languages, which enables us to assess this effect quantitatively; using BLEU as a distance metric between the two dialects, there are $\sim 2\times$ higher scores between CH-ZH and CH-BE after ASR (23.6) than between the reference texts (12.7), due to normalizing effects from ASR. Past work on ASR [36, 33] has explored WER-variants where words with the same normalized form are deemed correct. In place of a normalization dictionary [37] or morphological analyzer, given vocabulary mismatches, we train WFST grapheme-to-phoneme models [34] using Phonetisaurus [38] to map transcripts and references to a 'normalized' phonetic form. When subsequently applying text metrics to phonetic forms, we see improvements of 10-20% in reference-text metrics for both dialects. However, the discrepancy between German and the dialects remains for all metrics, and this step introduces one more intermediate model before evaluation.

ASR can introduce errors but also corrections. High-performance ASR models are integrated with MT metrics in order to reduce errors potentially introduced by the transcription needed to evaluate with reference-text metrics. However, we find large language models and well-trained decoders can also *correct* errors in speech synthesis, biasing the evaluation of the synthesized speech to appear better. These corrections can range from small errors in pronunciation by the synthesis model where the correct transcript is recovered by ASR beam search (Figure 9: *Pronunciation*), to reorderings in the ASR transcript from the word order present in the synthesized speech (Figure 9: *Ordering*).

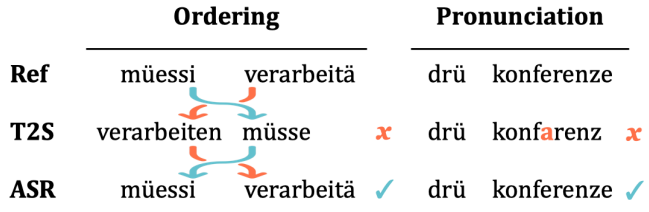

Fig. 9: Selected corrections introduced by ASR.

5. CONCLUSION

We evaluated the current metrics for translation with synthesis, and how translation to dialectal variants rather than to standardized languages impacts various evaluation methods. We found that many available metrics do not represent translation performance well: specifically, word-based text metrics (BLEU) and the speech metric MCD misrepresent text-to-speech performance for dialects and do not correlate well with human judgments. The character-based text MT metrics chrF and character-level BLEU are more robust to dialectal variation and transcription quality in evaluation, but correlations with human judgments remain moderate. Character-based MT metrics correlate better with human judgments than BLEU or MCD for our high-resource language, suggesting they may be the best currently available metrics for speech-to-speech translation and related tasks, but our findings suggest better metrics are still needed.

Acknowledgments. This project is supported by Ringier, TX Group, NZZ, SRG, VSM, Viscom, and the ETH Zürich Foundation. We would also like to thank our annotators; Chan Young Park for assistance with our annotation interface; and Matt Post, Nathaniel Weir, and Carlos Aguirre for feedback.

6. REFERENCES

[1] Steven Hillis, Anushree Prasanna Kumar, and Alan W Black, "Unsupervised phonetic and word level discovery for speech to speech translation for unwritten languages," in *INTERSPEECH*, 2019, pp. 1138–1142.

[2] Alex Waibel, "Interactive translation of conversational speech," *Computer*, vol. 29, no. 7, pp. 41–48, 1996.

[3] Enrique Vidal, "Finite-state speech-to-speech translation," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1997, vol. 1, pp. 111–114.

[4] Laurent Besacier, Hervé Blanchon, Yannick Fouquet, Jean-Philippe Guilbaud, Stéphane Helme, Sylviane Mazenot, Daniel Moraru, and Dominique Vaufreydaz, "Speech translation for French in the Nespole! European project," in *Eurospeech'01*, 2001, pp. pp–1291.

[5] Florian Metze, John McDonough, Hagen Soltau, Alex Waibel, Alon Lavie, Susanne Burger, Chad Langley, Lori Levin, Tanja Schultz, Fabio Pianesi, et al., "The NESPOLE! Speech to Speech Translation System," in *Human Language Technologies 2002*, 2002, pp. 6–pages.

[6] Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto, "The atr multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 365–376, 2006.

[7] Wolfgang Wahlster, *Verbmobil: foundations of speech-to-speech translation*, Springer Science & Business Media, 2013.

[8] Y. Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, M. Johnson, Z. Chen, and Yonghui Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *INTERSPEECH*, 2019.

[9] John Kominek, Tanja Schultz, and Alan W Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion," in *Spoken Languages Technologies for Under-Resourced Languages*, 2008.

[10] Ron J Weiss, RJ Skerry-Ryan, Eric Battenberg, Soroosh Mariooryad, and Diederik P Kingma, "Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis," *arXiv preprint arXiv:2011.03568*, 2020.

[11] Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu, "UWSpeech: Speech to Speech Translation for Unwritten Languages," *arXiv preprint arXiv:2006.07926*, 2020.

[12] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 958–965.

[13] Kei Hashimoto, Junichi Yamagishi, William Byrne, Simon King, and Keiichi Tokuda, "Impacts of machine translation and speech synthesis on speech-to-speech translation," *Speech Communication*, vol. 54, no. 7, pp. 857–866, 2012.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 311–318.

[15] Maja Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, Sept. 2015, pp. 392–395, Association for Computational Linguistics.

[16] Matt Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Oct. 2018, pp. 186–191.

[17] Meinard Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007.

[18] Kyubyong Park and Thomas Mulc, "Css10: A collection of single speaker speech datasets for 10 languages," *Interspeech*, 2019.

[19] "LibriVox," `https://librivox.org/`, 2018.

[20] Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann, "SwissDial: Parallel Multidialectal Corpus of Spoken Swiss German," 2021.

[21] "SPTK," `http://sp-tk.sourceforge.net/`, 2017.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.

[23] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019, pp. 48–53.

[24] Thanh-Le Ha, Jan Niehues, and Alexander Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," *Proceedings of IWSLT*, 2016.

[25] Taku Kudo and John Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Brussels, Belgium, Nov. 2018, pp. 66–71, Association for Computational Linguistics.

[26] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *ICASSP*, 2018.

[27] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[28] Christof Traber, Schamai Safra, Bleicke Holm, Dominic Schnyder, and Philipp Lichtenberg, "Text-to-Speech (TTS) for Seven Swiss German Dialects," SwissText, 6 2019.

[29] Christof Traber, Schamai Safra, Bleicke Holm, Dominic Schnyder, and Philipp Lichtenberg, "Text-to-Speech (TTS) for Seven Swiss German Dialects," SwissText, 6 2020.

[30] Philipp Koehn and Rebecca Knowles, "Six challenges for neural machine translation," *Proceedings of the First Workshop on Neural Machine Translation*, 2017.

[31] Bradley Efron and Robert J Tibshirani, *An introduction to the bootstrap*, CRC press, 1994.

[32] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 996–1002.

[33] Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic, "ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German," in *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 2020, pp. 15–24.

[34] Larissa Schmidt, Lucy Linder, Sandra Djambazovska, Alexandros Lazaridis, Tanja Samardžić, and Claudiu Musat, "A Swiss German dictionary: Variation in speech and writing," in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 2720–2725, European Language Resources Association.

[35] M. Stadtschnitzer and C. Schmidt, "Data-driven pronunciation modeling of swiss german dialectal speech for automatic speech recognition," in *LREC*, 2018.

[36] Ahmed Ali, Salam Khalifa, and Nizar Habash, "Towards Variability Resistant Dialectal Speech Evaluation," in *INTERSPEECH*, 2019, pp. 336–340.

[37] Tanja Samardžić, Yves Scherrer, and Elvira Glaser, "ArchiMob - a corpus of spoken Swiss German," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May 2016, pp. 4061–4066, European Language Resources Association (ELRA).

[38] Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2016.