Speech is More Than Words: Do Speech-to-Text Translation Systems Leverage Prosody?

Ioannis Tsiamas[◇*] **Matthias Sperber**[†] **Andrew Finch**[†] **Sarthak Garg**[†]
[◇]Universitat Politècnica de Catalunya [†]Apple
ioannis.tsiamas@upc.edu, sperber@apple.com

Abstract

The prosody of a spoken utterance, including features like stress, intonation and rhythm, can significantly affect the underlying semantics, and as a consequence can also affect its textual translation. Nevertheless, prosody is rarely studied within the context of speech-to-text translation (S2TT) systems. In particular, end-to-end (E2E) systems have been proposed as well-suited for prosody-aware translation because they have direct access to the speech signal when making translation decisions, but the understanding of whether this is successful in practice is still limited. A main challenge is the difficulty of evaluating prosody awareness in translation. To address this challenge, we introduce an evaluation methodology and a focused benchmark (named CONTRAPROST) aimed at capturing a wide range of prosodic phenomena. Our methodology uses large language models and controllable text-to-speech (TTS) to generate contrastive examples. Through experiments in translating English speech into German, Spanish, and Japanese, we find that (a) S2TT models possess some internal representation of prosody, but the prosody signal is often not strong enough to affect the translations, (b) E2E systems outperform cascades of speech recognition and text translation systems, confirming their theoretical advantage in this regard, and (c) certain cascaded systems also capture prosodic information in the translation, but only to a lesser extent that depends on the particulars of the transcript's surface form.[1]

1 Introduction

Prosody, which includes features like stress, intonation, and rhythm, is crucial for conveying meaning in spoken language beyond the literal words used (Ladd, 1980; Bolinger, 1989). Among others, prosody can direct focus and clarify meaning (Bolinger, 1961; Halliday, 1967), disambiguate syntax and sentence structure (Bolinger, 1989), convey the emotional state of the speaker (Banse and Scherer, 1996), and provide useful cues that make communication more effective (Shriberg et al., 1998). For example, the phrase "*Really?*" can express surprise, genuine interest or disbelief, depending on the intonation with which is spoken.

Example: *These are German teachers.*		
A	Prosody	These are GERMAN teachers.
	Explanation	Teachers from Germany
	Translation	Dies sind Deutschlehrer.
B	Prosody	These are German TEACHERS.
	Explanation	Teachers that teach German
	Translation	Dies sind deutsche Lehrer.
Example: *John laughed at the Party.*		
A	Prosody	John LAUGHED (pause) at the Party.
	Explanation	Laughed while at the party (literal)
	Translation	John lachte während der Party.
B	Prosody	John LAUGHED AT (pause) the Party.
	Explanation	Ridiculed the party (idiomatic)
	Translation	John lachte über die Party.

Table 1: Examples of prosody-aware Speech Translation from English to German.

Table 1 illustrates the importance of considering prosody when generating translations in S2TT. Sperber and Paulik (2020) suggest that E2E S2TT systems may have an inherent advantage over cascaded systems in this regard, because only the former have access to the speech signal when making translation decisions. However, our understanding of whether prosody informs translation choices in practice is currently still limited, as prior research on this topic either shows only anecdotal evidence (Huang et al., 2023b), focuses on only a small subset of prosodic phenomena (Zhou et al., 2024; Chen et al., 2024), or considers how prosody informs target-side speech with regards to generated prosody but not lexical choice (§6).

[*] Work done during an internship at Apple.
[1] github.com/apple/ml-speech-is-more-than-words

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 1235–1257
November 15–16, 2024.

In this paper, we take steps toward a reliable and comprehensive evaluation methodology, which is one of the most important prerequisites for achieving prosody-aware S2TT. We identify three central challenges that must be addressed: (1) Existing S2TT benchmarks often do not include prosody-rich spontaneous speech and/or do not include translations that are informed by the audio, limiting the extent to which reference translations are influenced by source-side prosody. (2) General-purpose evaluation methods like BLEU (Papineni et al., 2002) and COMET (Guerreiro et al., 2023) are insensitive to the often subtle changes in translation caused by input prosody. (3) Existing prosody-centric benchmarks are difficult to scale to broader coverage of languages and prosodic phenomena, which hinders comprehensive analysis.

To address these challenges, we take inspiration from prior work on behavioral testing (Ribeiro et al., 2020; Ferrando et al., 2023) and contrastive evaluation (Sennrich, 2017). We address the first challenge by synthesizing prosody-rich data that covers a wide range of prosodic phenomena through the use of large language models (LLMs) and controllable TTS (cTTS). We tackle the second challenge by developing a double-contrastive evaluation approach, i.e. a directional behavioral test that relies on minimal pairs (differing only in prosody) to evaluate prosody-awareness in S2TT in isolation. The resulting benchmark, CONTRAPROST (Contrastive Prosody ST), covers a variety of language pairs and prosodic phenomena. Since it is mostly automated, it can be further extended, thus addressing also the third challenge.

To investigate how well current state-of-the-art models understand and leverage prosody, we evaluate S2TT models of various sizes and types, including both E2E and cascaded systems. We find indications that S2TT models represent prosody internally, but this knowledge is often not manifested in the translations. We observe that while tested cascaded systems perform better on traditional evaluation (COMET), E2E models outperform cascaded models on CONTRAPROST. We also find indications that some amount of prosody is carried through transcripts in cascaded setups, but this depends on the particulars of the transcriptions. The most important implication of our findings is the need for exploring improvements of S2TT regarding prosody-awareness, e.g. through auxiliary losses or finetuning on prosody-rich data.

2 The CONTRAPROST Benchmark

CONTRAPROST is composed of double-contrastive examples (see Table 1), where each example is composed of a sentence in English that could be semantically ambiguous, along with two different pairs of <speech, translation> that capture contrastive cases of prosody.

As it would be expensive and practically difficult to collect such test data manually, we employ an automatic data generation process, illustrated in Fig. 1. First, we identify several relevant categories where prosody influences sentence semantics in important ways, and construct illustrative examples that reflect the respective phenomena of each category, while highlighting differences in prosody-induced meaning (§2.1). We then prompt GPT-4[2] (OpenAI, 2024) to generate sentences similar to the examples for each subcategory using in-context learning, grounding the generation on different text domains to increase diversity (§2.2). Next, GPT-4 is prompted to translate each prosodic case, while also being given access to the prosodies, meanings and general information of the category, thus acting as a prosody- and context-aware oracle translator (§2.3). Finally, we use the OpenAI TTS API[3] to synthesize the prosodic speech of each case (§2.4). Each generation stage is coupled with filtering and quality assessment to ensure the data are of high quality.

2.1 Categorization of Prosodic Phenomena

Below, we summarize the examined prosodic categories. Details and examples are available in the Appendices A and B.

(1) Sentence Stress. This is usually manifested through increased loudness, vowel length or higher pitch (Fry, 1955), invoking emphasis on certain words within a sentence, potentially changing the semantics by shifting focus (Wagner, 2020). We further categorize prosodic stress in four subcategories according to the purpose of the stress or its use in disambiguation of linguistic phenomena (see Appendix A.1).

(2) Prosodic Breaks. Here we consider the existence or placement of longer breaks in the flow of speech, primarily associated with tempo, that create different phrasal boundaries and help disambiguate syntax and sentence structure (Bolinger, 1989). We follow Hirschberg (2017) and use the

[2] GPT-4O-2024-05-13

[3] TTS-1-HD, platform.openai.com/docs/models/tts

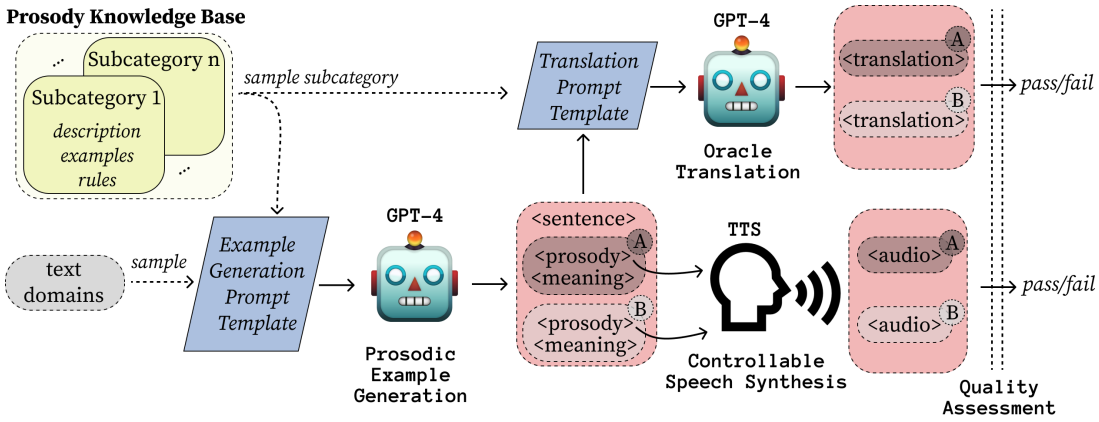

Figure 1: The Data Generation process for CONTRAPROST.

subcategories outlined in Appendix A.2.

(3) Intonation Patterns. This concerns the modality of the sentence, specifically whether it is a statement (falling tone), or a declarative question (rising tone) (Gunlogson, 2002).

(4) Emotional Prosody. A different emotional tone can indicate a speaker's emotional state and thus affect the semantics of the utterance (Banse and Scherer, 1996). Emotional tone is usually manifested through changes in pitch, tempo, and loudness. For example, happiness is associated with higher values in pitch and tempo, while sadness exhibits lower values for pitch, tempo, and loudness (Larrouy-Maestri et al., 2024). Here, we focus on the seven *basic* emotions: happy, sad, angry, disgust, surprisal, fear, and neutral (Ekman and Friesen, 1971; Ekman, 1992), based on which we construct all possible pairs, thus having 21 subcategories.

(5) Politeness. The level of politeness can be conveyed by non-verbal cues, and influences the pragmatic context of a conversation. A polite tone is associated with a higher pitch and a smooth rhythm, while an impolite tone is manifested through low pitch, irregular rhythm and very high or low loudness levels (Culpeper et al., 2003; Culpeper, 2011).

2.2 Prosodic Example Generation

For each category, we prompt GPT-4 to generate sentences based on hand-crafted category-specific examples. More specifically, we have the LLM generate English sentences, each with two different textual prosodic annotations and respective meanings/interpretations to guide subsequent translation (§2.3). The generated annotations include rich text that indicates different levels of emphasis, pause tags, and special punctuation such as ellipsis, exclamation, or interobang (!?). The sentence itself is generated to be as simple as possible, ending with a full stop or question mark.

Prompt 1: Prosodic Example Generation

You are a helpful assistant with expert knowledge in linguistics, speech, and prosody. Your task is to come up with examples of English sentences where different prosody would change the meaning of the sentence significantly.(1)
{Details for Category & Subcategory}(2)
Here are some examples to guide you:
{List of Examples}(3)
Strictly follow these rules:
{List of Rules}(4)
Provide a rating of how significant is the difference between the two meanings.(5)
Generate {n} such examples, with rating as high as possible,(6) in the domain of {domain}.(7)

The general prompt template is displayed in Prompt 1. It starts with some general information about the task, see superscript (1). The prompt then continues with details describing the current category/subcategory (2). The next part refers to in-context learning (Brown et al., 2020), where we provide a list of illustrative, hand-crafted examples for the LLM to follow (3). In certain subcategories, due to repeated mistakes observed in preliminary explorations, we also provide examples to avoid. In (4) we provide a list of rules for the LLM to adhere to, indicating the desired structure of the sentence and how to use prosodic notation, which might not be obvious from the examples (3). Examples of such rules are "do not include prosodic annotations in the sentence," or "stress different noun-phrases in each prosodic case." We further-

more use *self-criticism* (Huang et al., 2023a) by instructing the model to rate its own generations, according to how different the two prosodic interpretations are (5). Then we instruct the LLM to generate examples that have high scores after self-reflection (6). These scores are also used later during filtering. Finally, to avoid repetitive examples and enhance diversity, we condition the generation on specific text domains (7) (Chung et al., 2023). The list of domains is also generated by GPT-4 based on the context that its subcategory would naturally occur (e.g. *legal testimonies*). For each text domain in the subcategory the LLM then generates n candidate examples. We use several hand-crafted text-based filtering steps to ensure that the examples generated by the LLM at this stage comply with the instructions specified in (4).

2.3 Oracle Translation

Recent research on the emerging capabilities of LLM-based MT (Vilar et al., 2023; Alves et al., 2023; Zhang et al., 2023) has shown that LLMs can attain very high translation quality, especially for high-resource languages (Robinson et al., 2023) and including translation factors such as emotions (Brazier and Rouas, 2024), suggesting the possibility that LLMs can be leveraged for prosodic translation synthesis. To obtain the translations of the prosodic cases, we thus utilize GPT-4 as a prosody- and context-aware oracle translator. The LLM is prompted to translate, while having access to the sentence, the textual prosodic annotations (prosody-awareness), and the semantic interpretations (context-awareness). The template prompt is shown in Prompt 2. We provide a list of contraints to the LLM with several goals in mind: (i) avoid generating prosodic annotations in the translations; (ii) avoid translating the interpretations rather than the sentences; (iii) encourage the model to generate different translations for each case; (iv) ensure that differences in the translations are only due to the difference in the prosodies.

Although prosody variants substantially influence sentence semantics, this does not always imply that the ideal translations must differ. In particular, sometimes a translation that leaves semantics ambiguous may be preferred as the most natural translation.[4] As a consequence, constraint (iii) is sometimes overly strict and even in conflict with constraint (iv), leading to changes in the translations

[4]This is essentially an instance of the fluency-accuracy trade-off (Lim et al., 2024).

Prompt 2: Oracle Translation

You are a helpful assistant with expert knowledge in speech, prosody, linguistics and translation, particularly in English and {Target Lang}.
You will be provided with a sentence in English (S) and two different prosodic variations (S_A, S_B), focused on {Category}, which correspond to two different semantic interpretations.
Your task is to translate S, S_A and S_B into {Target Lang}, as T, T_A, and T_B.
Carry out the translation in these steps:
(1) Translate S into T.
(2) Translate S_A to T_A and S_B to T_B, by focusing on how T should change in order to reflect the additional information from the prosodies.
The following constraints should be applied: {List of Constraints}
The sentence S is: {sentence}
The two different prosodic variations are:
S_A. {$prosody_A$} ({$meaning_A$})
S_B. {$prosody_B$} ({$meaning_B$})

that do not stem from the prosodies, that are not idiomatic. To account for that, we include a post-editing step, where GPT-4 is instructed to choose the most fitting translation among $\{T, T_A, T_B\}$ for each prosodic case, independently from the other prosodic cases, while having access only the prosody information (Prompt 3). We prompt the LLM to first provide an explanation, before selecting the most appropriate translation, in order to induce *chain-of-thought* reasoning effect (Kojima et al., 2024).

Prompt 3: Translation Post-editing

You are a helpful assistant and an expert translator. You will be provided with a sentence in English and different possible translations in {Target Lang}. The English sentence can contain rich prosodic text with {Category-specific information}, that affects the meaning of the sentence. Your task is to select the most appropriate and prosody-aware translation. First provide a brief explanation of your reasoning and then the index of the selected translation.
The sentence S to be translated is {sentence} and the candidate translations are: [T, T_A, T_B]}

After post-editing we remove all examples where the prosodic cases have identical translations, i.e. (T_A=T_B). As an extra measure, we also remove examples where the word length-ratio of the non-

prosodic translation T and one of the prosodic translations T_A, T_B is not within (0.75, 1.25)[5]. This aims to remove translations that are overly explanatory, including new bits of information that can be due to the prosody, but are making the translation unnatural (see Table 10 in App. D.2 for examples.).

2.4 Controllable Speech Synthesis

We use the OpenAI TTS which can synthesize very natural speech with high-quality audio, offering six different voice profiles. While there are no clear guidelines[6] on how to control prosody, we identified some effective prompting strategies to control the TTS output through trial-and-error (Table 2).

Effect	TTS Prompting
Strong Emphasis	*WORD*
Normal Emphasis	*word*
Slight Emphasis	_word_
Pause	<pause>
Statement Intonation	Prepend <statement>
Question Intonation	Prepend <question> & Append ????
Emotional/Polite Tone	Prepend & Append Emojis

Table 2: OpenAI TTS prompting strategies.

To ensure that the generated audio follows the correct wording and exhibits the intended prosodic characteristics we use the following process: First, we generate six candidates (one per voice) for each prosody, discarding invalid candidates (WER$\neq$0) using an ASR model. Then we estimate prosody quality using category-specific tests in order to rank or filter examples. These tests employ techniques such as forced alignment (Kürzinger et al., 2020), signal processing, punctuation probability, and speech emotion classification. They are explained in detail in Appendix C.

3 Contrastive Evaluation

General-purpose MT metrics like BLEU and COMET may be insensitive to subtle changes caused by prosody, and do not allow disentangling prosody awareness from overall translation quality. Thus, to assess how well an S2TT model can handle prosody specifically, we develop a contrastive evaluation framework (Sennrich, 2017). Note that previous work on contrastive evaluation uses a single source and two or more targets (Sennrich, 2017; Vamvas and Sennrich, 2021; Zhou et al., 2024) of which only one is correct. The model likelihood is then estimated for each target, and models are preferred that assign a better score to the correct example than to the foil(s). Here, we generalize this approach to leverage ContraProST's *double-contrastive* pairs, i.e. two sources and two targets (Fig. 1).

Formally, each double-contrastive pair has two cases $\{X^a, Z, Y^a\}$ and $\{X^b, Z, Y^b\}$, where X^a, X^b are the two different prosodic speech signals, Z is the source text (same for both cases), and Y^a, Y^b are the different translated texts for each case. Thus, each example has two correct pairs (X^a, Y^a), (X^b, Y^b) and two incorrect ones (X^a, Y^b), (X^b, Y^a). We propose the following conditions to assess whether the S2TT model can correctly solve the contrastive example, and to what degree:

$$C_{\mathcal{G}} = \mathbf{1}\Big[f(Y^a \mid X^a;\theta) - f(Y^b \mid X^a;\theta) > 0 \text{ and } f(Y^b \mid X^b;\theta) - f(Y^a \mid X^b;\theta) > 0\Big]$$

$$C_{\mathcal{D}} = \mathbf{1}\Big[f(Y^a \mid X^a;\theta) - f(Y^b \mid X^a;\theta) + f(Y^b \mid X^b;\theta) - f(Y^a \mid X^b;\theta) > 0\Big]$$

Here, $\mathbf{1}[\cdot]$ is the indicator function, and $f(\cdot) > 0$ is a function that measures the agreement between audio input X and target translation Y under the S2TT model with parameters θ. $C_{\mathcal{G}}$ is a *global* condition, requiring the model to prefer both of the correct pairs versus the incorrect ones according to f. $C_{\mathcal{D}}$ is a *directional* condition (Ribeiro et al., 2020) where we require a net positive directional movement for the two comparisons. We expect a model to have a strong internal representation of prosody if it can solve the global condition, and weak representation if it can only solve the directional one.[7]

We consider two different functions f to measure the agreement of X and Y.

3.1 Contrastive Likelihood

Similar to prior work on contrastive evaluation (Sennrich, 2017; Vamvas and Sennrich, 2021; Zhou et al., 2024) we use the model likelihood to measure the level of agreement between input audio and target text. We obtain the model likelihood $\mathcal{L} \in \mathbb{R}^+$ for a reference $Y = (y_1, \ldots, y_{|Y|})$, given a speech signal $X \in \mathbb{R}^k$ and an E2E S2TT model with parameters θ_{E2E}. It is defined as the product

[5] We use character-based length-ratio for Japanese.

[6] platform.openai.com/docs/guides/text-to-speech

[7] Note that $C_{\mathcal{G}}$ is a sufficient condition for $C_{\mathcal{D}}$.

of the conditional probabilities, normalized by the length of the reference. Formally:

$$\mathcal{L}(Y \mid X; \theta_{\text{E2E}}) = \frac{1}{|Y|} \prod_{i=1}^{|Y|} p_{\theta_{\text{E2E}}}(y_i \mid X, y_{<i})$$

For a cascaded S2TT model we approximate the true likelihood by considering the top-n ASR hypotheses $\mathcal{Z} = \{Z^{(1)}, \ldots, Z^{(n)}\}$. Assuming the lengths of the $\mathcal{Z}$ are generally similar, we get:

$$\mathcal{L}(Y \mid X; \theta_{\text{casc}}) \approx \mathcal{L}(Y \mid \mathcal{Z}; \theta_{\text{MT}}) \mathcal{L}(\mathcal{Z} \mid X; \theta_{\text{ASR}})$$
$$\approx \frac{\sum_{j=1}^{n} \left[\mathcal{L}(Y \mid Z^{(i)}; \theta_{\text{MT}}) \cdot \mathcal{L}(Z^{(i)} \mid X; \theta_{\text{ASR}})\right]}{\sum_{j=1}^{n} \mathcal{L}(Z^{(i)} \mid X; \theta_{\text{ASR}})}$$

Furthermore, to remove a potential bias of the model against rare translations, we normalize by the unconditioned decoder likelihood of the reference:[8]

$$f_{\overline{\mathcal{L}}}(Y \mid X; \theta) = \frac{\mathcal{L}(Y \mid X; \theta)}{\mathcal{L}(Y \mid \theta)} \quad (1)$$

3.2 Contrastive Translation Quality

A common criticism of using model likelihoods is that they do not assess whether the correct output is actually generated in practice, due to teacher forcing. To address this, we propose another function that leverages translation quality estimation (QE) to compare unconstrained autoregressively generated model outputs. We obtain the hypothesis $\hat{Y}$ of input X by generating with the S2TT model $\mathcal{M}_\theta$, and use XCOMET (Guerreiro et al., 2023) to measure the quality of the translation. Thus:

$$f_{\mathcal{Q}}(Y \mid X; \theta) = \mathcal{Q}\big(Y, \mathcal{M}_\theta(X)\big) = \mathcal{Q}(Y, \hat{Y}) \quad (2)$$

The contrastive metrics using $f_{\mathcal{Q}}$ are expected to give us a better insight into how influential prosody is when translating with S2TT models, as compared to using $f_{\mathcal{L}}$ (Eq. 1), since they consider autoregressive generation and beam search.

4 Experimental Setup

4.1 Data Generation

For prosodic example generation with GPT-4 (§2.2) we used a temperature of 1, and 20 text domains per subcategory. The model was prompted to generate 10 examples[9] for each pair of (subcategory, domain). The total number of subcategories is 27 (more details in App. A), amounting to 5.5k examples of English sentences with pairs of prosodies and meanings created initially. Then we generated the candidates for the six voices with the TTS (5.5k×6×2 = 66k) and choose the 11k best candidates as described in §2.4. After quality assessment we end up with 2.8k examples with good prosody quality in the generated audio. Then we separately translated each one to the three target languages German (De), Spanish (Es), and Japanese (Ja). After post-editing and filtering we obtained 1.3k–1.4k full examples for each language pair (Table 3).

Category	En-De	En-Es	En-Ja
Emotional prosody	373	379	376
Sentence stress	277	279	342
Prosodic breaks	276	252	289
Politeness	212	193	206
Intonation patterns	173	173	173
Total	1,311	1,294	1,386

Table 3: Number of examples for each language pair in CONTRAPROST. More details are in Appendix D.1.

4.2 Speech-to-text Translation Models

We evaluated S2TT models that fall under these three categories:

- E2E, where inference is done without an intermediate transcription step. The decoder of this model has full access to the prosody of the input.
- AED-based cascade, which is composed of an attentional encoder-decoder (AED) (Vaswani et al., 2017) ASR model and an MT model. We expect the decoder of the MT model to have limited access to prosody, unless the ASR model is able to encode it in the transcription. This is possible mainly though punctuation, but also when the ASR model is acting more interpretative (i.e. generating synonyms that better fit the prosody rather than the spoken words).
- CTC-based cascade, which uses a CTC encoder (Graves et al., 2006) for the ASR part. The decoder of the MT model is expected to have almost no access to prosody since CTC model outputs are not punctuated and cannot be interpretative.

[8] Estimated by using an empty audio for E2E case and empty source text in the MT model for the cascade.

[9] We generated 15/20 examples for intonation patterns/politeness, respectively.

We are evaluating the following S2TT models:

- SEAMLESSM4T (Seamless Communication, 2023b) is a multilingual and multimodal encoder-decoder. It is trained with multi-task learning on ASR, MT, S2TT and also on speech-to-speech translation (S2ST), and can thus be used in either E2E or cascaded (AED) mode.
- XLS-R (Babu et al., 2021) is a multilingual E2E model, of which the encoder is based on WAV2VEC2.0 and its decoder on MBART50 (Tang et al., 2020).
- ZEROSWOT (Tsiamas et al., 2024) is a zero-shot E2E model that connects a WAV2VEC 2.0 CTC encoder and NLLB (NLLB Team, 2022).
- SALMONN (Tang et al., 2024) is an audio LLM that connects WHISPER (Radford et al., 2022) and BEATs (Chen et al., 2023) to the Vicuna LLM (Peng et al., 2023), and can be used as an E2E S2TT model.
- WHISPER & NLLB (AED-based cascade).
- CTC & NLLB (CTC-based cascade) with WAV2VEC 2.0 or HUBERT (Hsu et al., 2021).

We considered different versions of these 6 models, thus evaluating in total 31 S2TT model variants of different sizes and capabilities (App. E).

4.3 Metrics

We used beam search with beam size 5 to generate hypotheses. For estimating the conditional likelihood of the cascade (§3.1) we used the top-5 ASR hypotheses. For the contrastive translation quality (§3.2) we used XCOMET-XL[10] (Guerreiro et al., 2023), which is a state-of-the-art neural quality estimation metric based on XLM-R (Conneau et al., 2020). For all evaluated models we present their *contrastive likelihood* and *contrastive translation quality* scores, both *global* and *directional* versions, as a percentage of solved examples. We also evaluate them on standard QE using XCOMET-XL, by using the 2 correct pairs of each example (2.6k samples). For statistical significance testing we used bootstrap resampling (Efron, 1979) with 10k resamples and a 95% confidence interval.

5 Experimental Results

In Table 4 we present the results of evaluating a selection of large and recent model versions all three language pairs. We find that most S2TT models have at least some internal representation of prosody, enabling them to outperform the random baseline of 50% for the directional contrastive likelihood. On the other hand, when we consider autoregressive generation, we observe that the scores for the directional contrastive quality are relatively low[11], indicating that prosody is often not prominent enough in the internal representations of the models for it to be manifested in the generated translations. Furthermore, we find that the task of correctly solving both sub-cases of each example (global agreement) is very challenging for all models, with scores ranging around 10% for both contrastive metrics. We observe that even though the best performing model according to standard evaluation (XCOMET) is a cascade system, it falls behind the best E2E models when considering the contrastive evaluation on CONTRAPROST. This finding illustrates why it is beneficial to separate prosody evaluation from general accuracy evaluation to study the phenomenon, which is further supported by our observation that the prosody and general accuracy metrics are only moderate correlated (see Fig. 5 in App. F).

Are model type and model size important for prosody-awareness? We evaluate all 31 S2TT models using *global contrastive quality*, and run a regression analysis with the model type (E2E/AED-cascade/CTC-cascade) and model size as inputs. We use a mixed effects model (Pinheiro and Bates, 2006) to group together each model family, and thus account for random effects, such as the training data and hyperparameters. Specifically:

$$y_{ij} = \beta_0 + \beta_1 \mathrm{S}_{ij} + \beta_2 \mathrm{AED}_{ij} + \beta_3 \mathrm{CTC}_{ij} + u_j + \epsilon_{ij},$$

where y_{ij} is the score of i-th model variant of the j-th model family, β_0 is the intercept, S is the log of the model size, AED and CTC are binary variables, u_j is the random effect for j-th model family, and ϵ_{ij} is a residual error term. All scores are available in Table 11 in App. F. In Figure 2 we confirm with statistical significance that the E2E models outperform the cascades in all three language directions.[12] There is also a statistically significant negative impact on prosody-awareness when the cascade is based on a CTC ASR model that may be explained by the absence of punctuation in CTC

[10] hf.co/Unbabel/XCOMET-XL

[11] Assuming XCOMET is 0 for randomly generated text, the baseline scores are also 0.

[12] Note that results are borderline non-significant for En-Ja against the AED-cascade.

Model Name	Contrastive Likelihood		Contrastive Quality		xCOMET
	Directional	Global	Directional	Global	
English → German					
SeamlessM4T-v2-Large	61.2	**13.5**	37.4	14.5	0.988
XLS-R 2B	59.3	4.6	31.1	7.3	0.980
ZeroSwot-Large	60.6	9.7	29.2	8.7	0.990
SALMONN-13B	**62.8**	7.2	**43.2**	**15.9**	0.975
SeamlessM4T-v2-Large	60.2	12.9	31.1	10.4	0.991
Whisper-v3-Large & NLLB-3.3B	60.7	5.8	23.1	5.5	**0.992**
HuBERT-XL & NLLB-3.3B	39.4	0.5	20.5	2.6	0.979
English → Spanish					
SeamlessM4T-v2-Large	**64.9**	**13.4**	37.9	11.0	0.982
XLS-R 2B	57.6	5.6	32.0	8.4	0.930
ZeroSwot-Large	57.5	9.2	31.1	5.6	0.948
SALMONN-13B	61.3	3.6	**39.6**	**12.3**	0.967
SeamlessM4T-v2-Large	61.3	11.7	29.5	7.6	0.984
Whisper-v3-Large & NLLB-3.3B	63.2	2.9	25.4	4.8	**0.987**
HuBERT-XL & NLLB-3.3B	41.8	0.2	20.8	2.4	0.968
English → Japanese					
SeamlessM4T-v2-Large	59.4	**12.4**	40.3	13.8	0.956
XLS-R 2B	60.0	4.6	27.4	7.0	0.950
ZeroSwot-Large	58.8	7.9	23.6	7.9	**0.970**
SALMONN-13B	**60.4**	10.8	**46.1**	**16.1**	0.859
SeamlessM4T-v2-Large	59.4	9.1	31.0	8.7	0.961
Whisper-v3-Large & NLLB-3.3B	59.8	4.9	21.5	5.3	0.960
HuBERT-XL & NLLB-3.3B	40.4	0.8	15.7	2.5	0.922
Average					
SeamlessM4T-v2-Large	**61.8**	**13.1**	38.5	13.1	0.975
XLS-R 2B	59.0	4.9	30.2	7.6	0.953
ZeroSwot-Large	59.0	8.9	28.0	8.1	0.969
SALMONN-13B	61.5	7.2	**42.9**	**14.8**	0.933
SeamlessM4T-v2-Large	60.3	11.2	30.5	8.9	0.979
Whisper-v3-Large & NLLB-3.3B	61.2	4.5	23.3	5.2	**0.980**
HuBERT-XL & NLLB-3.3B	40.5	0.5	19.0	2.5	0.956

Table 4: Contrastive Evaluation of S2TT models on ContraProST. Grey background indicates a cascaded system.

transcripts, which if present can at least approximately signal some prosodic phenomena. Finally, although there is some evidence that larger models are more prosody-aware, results are not statistically significant. We speculate that larger models have more capacity to encode prosody in the weights, but since prosody is perhaps not sufficiently represented in the training data, this effect is limited.

How do results compare across categories and models? In Figure 3 we present results across individual prosodic categories for four different English-German models, and perform pairwise model comparisons via bootstrap resampling[13]. The only category models are able to solve consistently is *intonation patterns*, which can also be solved by cascaded models due to the presence of punctuation in the transcription. The comparably lower scores in the other four categories further demonstrate the inability of current state-of-the-art models to use prosody, with *sentence stress* being the most challenging. Through the pairwise comparisons, we find that an LLM-based model (SALMONN) is not statistically different from a more standard S2TT model, like SeamlessM4T. Next, comparing the SeamlessM4T model in both E2E and cascade allows us to control for parameters such as training data and architecture, in order to observe the effect of model type, giving more clarity of our results on the theoretical advantage of E2E models. Finally, we observe a clear performance gain by using the SeamlessM4T cascade over the Whisper & NLLB one. We hypothesize this advantage is due to the multitasking nature of SeamlessM4T, which makes its ASR

[13]English-Spanish/Japanese are available at Figures 6, 7 in App. F.

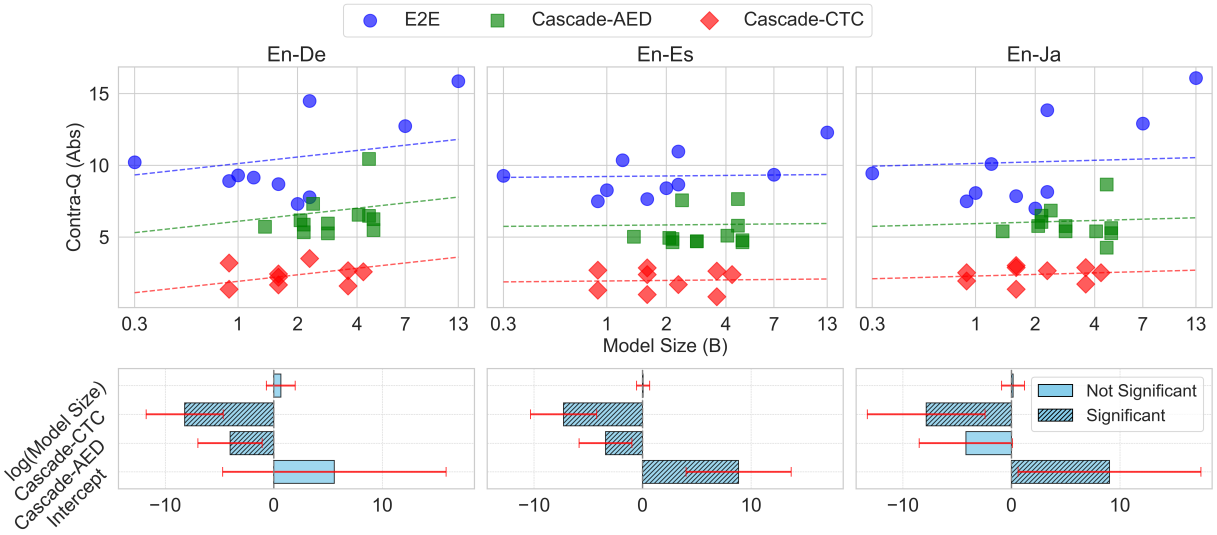

Figure 2: Regression Analysis of model types and model sizes per language pair.

mode more interpretative than standard ASR models. This allows the ASR part of the cascade to escape the word-by-word paradigm, and use more fitting words in the transcription (such as synonyms) that fit better the prosody of the audio. Supporting this hypothesis. we observe a worse WER score for SEAMLESSM4T (11%) compared to WHISPER (4%).

Is the level of prosody-awareness language-dependant? In Figure 4 we carry out a similar regression analysis as in Figure 2, but with the language pair as an independent categorical variable. Interestingly, we observe that there are differences between the three language pairs, and also significant for Spanish vs. German, which indicates that prosody-awareness in S2TT could be language-dependant. We hypothesize that the expressivity of the target language might be a relevant factor, since more expressive languages might be able to easier encode the prosody of the source speech into text.

6 Related Work

Prosody has traditionally been an important topic for TTS research (Kohler, 1991), either for transferring (Skerry-Ryan et al., 2018) or encoding it (Pamisetty and Sri Rama Murty, 2022) in the synthesized speech. Furthermore, Torresquintero et al. (2021) created a dataset for evaluating prosody transfer in TTS models, which contains several categories, similar to our study here. Naturally prosody has also been the focus of S2ST systems, in order to translate in a more expressive way (Aguero et al., 2006; Do et al., 2017; Communication et al., 2023). The topic has received less attention in the context of S2TT. Chen et al. (2024) present a dataset for emotional prosody based on speech and translations from TV series, and show that finetuning with emotion labels, can improve translation quality. Zhou et al. (2024) studied the prosody-awareness of WHISPER in E2E and cascade mode, in translating Korean *wh-phrases* using contrastive likelihood, and find evidence of the E2E model outperforming the cascade. Here we contribute a broader study of prosody in S2TT, by proposing a double-contrastive benchmark that covers several prosodic categories, the use of more generative-like contrastive evaluation, and evaluating a plethora of S2TT models. Finally, de Seyssel et al. (2023) present a benchmark for evaluating prosody-awareness in self-supervised acoustic representations. Similarly to our study they present evidence of prosody awareness in the representations. Contrary to our results, they conclude that size has a positive effect on prosody awareness.

7 Conclusions

We presented CONTRAPROST, a benchmark based on double-contrastive examples for evaluating prosody-awareness in S2TT models, covering several categories and languages. In addition to standard contrastive evaluation based on model likelihoods, we proposed a generative contrastive metric based on quality estimation. We evaluated a plethora of models, and found that they exhibit some signs of prosody-awareness, but the effect is often not strong enough to influence the translations. We also confirmed the previously hypothesized inherent advantage of E2E models com-

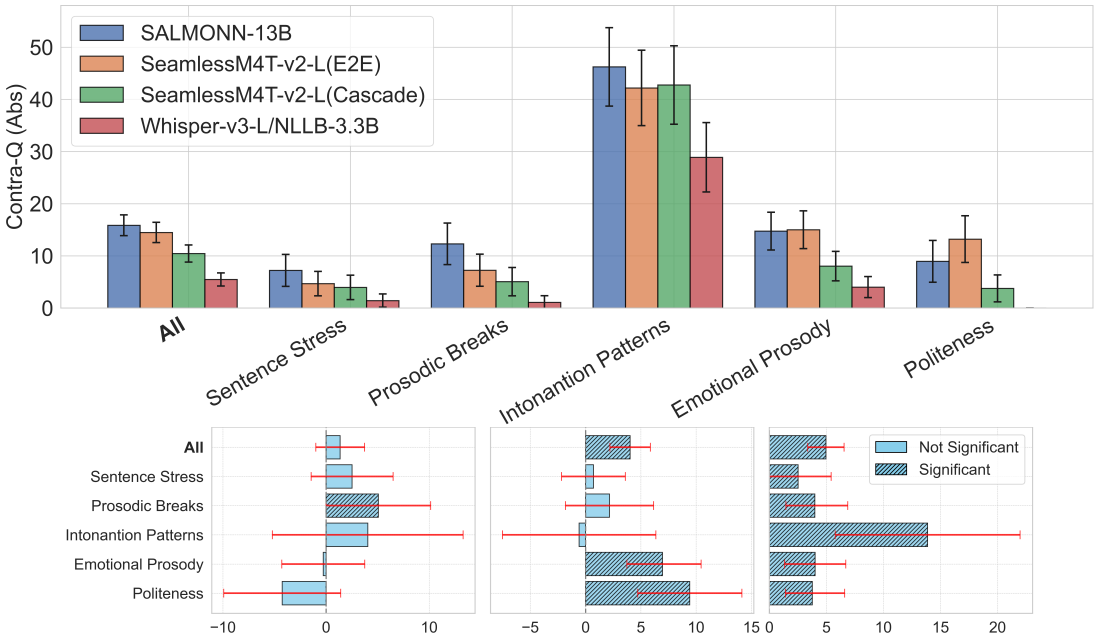

Figure 3: Upper: Model performance per category (En-De). Lower Model performance comparisons (En-De), (a): SALMONN-13B vs. SEAMLESSM4T-V2-LARGE, (b) SEAMLESSM4T-V2-LARGE(E2E) vs. SEAMLESSM4T-V2-LARGE(cascade), (c) SEAMLESSM4T-V2-LARGE(cascade) vs. WHISPER-V3-LARGE/NLLB-3.3B.

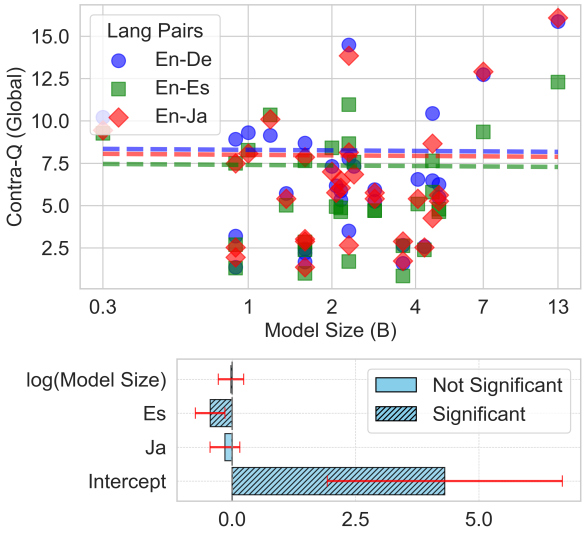

Figure 4: Regression Analysis of language pairs.

pared to cascaded models. We hope that our benchmark and findings will motivate more research into prosody-aware S2TT in the future, enabling us to better understand it and improve it.

Limitations

For creating CONTRAPROST we relied on an almost entirely automated data generation process. This allowed us to create a comprehensive dataset covering several prosodic phenomena and three language pairs, in a fast and cost-effective way. It would also enable expanding the coverage of languages and prosodic phenomena relatively easy in the future. Nevertheless, despite our best efforts regarding filtering and quality assessment (§2 and App. C), the data is not perfect and includes a certain amount of noise. We observed the following sources of noise in order of decreasing importance: (1) prosody not prominent in the generated speech; (2) translations overly explanatory or not encoding prosody; (3) semantic interpretations of the two cases rather similar. We do not expect these issues to be so frequent as to alter the findings of this work in a systematic way, but additional human annotation or verification would be a valuable step for future work. Furthermore, as the landscape of available generative models, in particular controllable TTS, is changing quickly, the quality of results using our data generation process would expectantly become less of a concern in future iterations.

Our study follows a contrastive evaluation methodology in order to isolate prosody-related behavior. As a consequence, our study does not allow drawing conclusions on how much prosody matters in real life data, and in what domains it is especially important. In addition, we hypothesize that some prosodic phenomena could be correctly translated by having access to the broader context of the conversation (context-aware S2TT), which we leave for future research.

References

P.D. Aguero, J. Adell, and A. Bonafonte. 2006. Prosody Generation for Speech-to-Speech Translation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *Preprint*, arXiv:2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Rainer Banse and Klaus R Scherer. 1996. Acoustic Profiles in Vocal Emotion Expression. *Journal of personality and social psychology*, 70(3):614.

Dwight Bolinger. 1989. *Intonation and Its Uses*. Stanford University Press, Redwood City.

Dwight L. Bolinger. 1961. Contrastive Accent and Contrastive Stress. *Language*, 37(1):83–96.

Charles Brazier and Jean-Luc Rouas. 2024. Conditioning LLMs with Emotion in Neural Machine Translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 33–38, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio Pre-training with Acoustic Tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Sirou Chen, Sakiko Yahata, Shuichiro Shimizu, Zhengdong Yang, Yihang Li, Chenhui Chu, and Sadao Kurohashi. 2024. MELD-ST: An emotion-aware speech translation dataset. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10118–10126, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual Expres-

sive and Streaming Speech Translation. *Preprint*, arXiv:2312.05187.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jonathan Culpeper. 2011. *"It's not what you said, it's how you said it!" Prosody and Impoliteness*, pages 57–84. De Gruyter Mouton, Berlin, New York.

Jonathan Culpeper, Derek Bousfield, and Anne Wichmann. 2003. Impoliteness Revisited: With Special Reference to Dynamic and Prosodic Aspects. *Journal of Pragmatics*, 35:1545–1579.

Maureen de Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, and Emmanuel Dupoux. 2023. ProsAudit, a prosodic benchmark for self-supervised speech models. In *Proc. INTERSPEECH 2023*, pages 2963–2967.

Nicole Dehé. 2014. *Parentheticals in Spoken English : The Syntax-Prosody Relation*. Cambridge [u.a.] : Cambridge University Press.

Quoc Truong Do, Sakriani Sakti, and Satoshi Nakamura. 2017. Toward Expressive Speech Translation: A Unified Sequence-to-Sequence LSTMs Approach for Translating Words and Emphasis. In *Proc. Interspeech 2017*, pages 2640–2644.

B. Efron. 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26.

Paul Ekman. 1992. Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science*, 3(1):34–38.

Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Javier Ferrando, Matthias Sperber, Hendra Setiawan, Dominic Telaar, and Saša Hasan. 2023. Automating Behavioral Testing in Machine Translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1014–1030, Singapore. Association for Computational Linguistics.

D. B. Fry. 1955. Duration and Intensity as Physical Correlates of Linguistic Stress. *The Journal of the Acoustical Society of America*, 27(4):765–768.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Preprint*, arXiv:2310.10482.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Christine Gunlogson. 2002. Declarative questions. In *Proceedings of Semantics and Linguistic Theory (SALT) XII*, pages 124–143, Ithaca, NY. CLC Publications.

M. A. K. Halliday. 1967. Notes on transitivity and theme in English. Part 1 and 2. *Journal of Linguistics*, 3:199–244.

Julia Hirschberg. 2017. Pragmatics and Prosody (Chapter 28). In *The Oxford Handbook of Pragmatics*. Oxford University Press.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023a. Large Language Models Can Self-Improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Zhichao Huang, Rong Ye, Tom Ko, Qianqian Dong, Shanbo Cheng, Mingxuan Wang, and Hang Li. 2023b. Speech Translation with Large Language Models: An Industrial Practice. *arXiv preprint arXiv:2312.13585*.

Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino. 2023. UnitY: Two-pass direct speech-to-speech translation with discrete units. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15655–15680, Toronto, Canada. Association for Computational Linguistics.

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, Massachusetts.

Wouter Jansen, Michelle L. Gregory, and Jason M. Brenier. 2001. Prosodic correlates of directly reported speech: Evidence from conversational speech. In *Proc. ITRW on Prosody in Speech Recognition and Understanding*, page paper 14.

J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

G. Klewitz and E. Couper-Kuhlen. 1999. Quote-unquote? The role of prosody in the contextualization of reported speech sequences. *Pragmatics*, 9(4):459–485.

K.J. Kohler. 1991. Prosody in speech synthesis: the interplay between basic research and TTS application. *Journal of Phonetics*, 19(1):121–138. Speech Synthesis and Phonetics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large Language Models are Zero-shot Reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. CTC-Segmentation of Large Corpora for German End-to-End Speech Recognition. In *Speech and Computer*, pages 267–278, Cham. Springer International Publishing.

J.D.R. Ladd. 1980. *The Structure of Intonational Meaning: Evidence from English*. Indiana University Press, Bloomington.

Pauline Larrouy-Maestri, David Poeppel, and Marc D. Pell. 2024. The Sound of Emotional Prosody: Nearly 3 Decades of Research and Future Directions. *Perspectives on Psychological Science*, 0(0):17456916231217722. PMID: 38232303.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Mark Y. Liberman and Richard Sproat. 1992. The Stress and Structure of Modified Noun Phrases in English. In *Lexical Matters*.

Zheng Wei Lim, Ekaterina Vylomova, Trevor Cohn, and Charles Kemp. 2024. Simpson's paradox and the accuracy-fluency tradeoff in translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–103, Bangkok, Thailand. Association for Computational Linguistics.

Steven R. Livingstone and Frank A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):1–35.

Marina Nespor and Irene Vogel. 1986. Prosodic Phonology. *Phonology*, 5(1):161–168.

Giridhar Pamisetty and K. Sri Rama Murty. 2022. Prosody-TTS: An End-to-End Speech Synthesis System with Prosody Control. *Circuits Syst. Signal Process.*, 42(1):361–384.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction Tuning with GPT-4. *Preprint*, arXiv:2304.03277.

Gabriel Peyré and Marco Cuturi. 2019. *Computational Optimal Transport: With Applications to Data Science*. Now Foundations and Trends.

Jose Pinheiro and Douglas M. Bates. 2006. *Mixed-effects Models in S and S-PLUS*. Statistics and Computing. Springer Science & Business Media, New York.

Patti Price, Mari Ostendorf, Stefanie Shattuck-Hufnagel, and Cynthia Fong. 1991. The Use of Prosody in Syntactic Disambiguation. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, page 372–377, USA. Association for Computational Linguistics.

Joel Pynte. 1996. Prosodic Breaks and Attachment Decisions in Sentence Parsing. *Language and Cognitive Processes*, 11(1-2):165–192.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Rico Sennrich. 2017. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol van Ess-Dykema. 1998. Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech? *Language and Speech*, 41(3-4):443–492. PMID: 10746366.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4693–4702. PMLR.

Matthias Sperber and Matthias Paulik. 2020. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *Preprint*, arXiv:2008.00401.

Alexandra Torresquintero, Tian Huey Teh, Christopher G.R. Wallis, Marlene Staib, Devang S. Ram Mohan, Vivian Hu, Lorenzo Foglianti, Jiameng Gao, and Simon King. 2021. ADEPT: A Dataset for Evaluating Prosody Transfer. In *Proc. Interspeech 2021*, pages 3880–3884.

Ioannis Tsiamas, Gerard Gállego, José Fonollosa, and Marta Costa-jussà. 2024. Pushing the Limits of Zero-shot End-to-End Speech Translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14245–14267, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. On the Limits of Minimal Pairs in Contrastive Evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Michael Wagner. 2020. Prosodic Focus. In Daniel Gutzmann, Lisa Matthewson, Cecilia Meier, Hotze Rullmann, and Thomas E. Zimmermann, editors, *The Wiley Blackwell Companion to Semantics*. Wiley–Blackwell.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert,

Gabriel Synnaeve, and Michael Auli. 2020. Self-training and Pre-training are Complementary for Speech Recognition. *Preprint*, arXiv:2010.11430.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Giulio Zhou, Tsz Kin Lam, Alexandra Birch, and Barry Haddow. 2024. Prosody in Cascade and Direct Speech-to-Text Translation: a case study on Korean Wh-Phrases. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 674–683, St. Julian's, Malta. Association for Computational Linguistics.

A Prosodic Subcategories

Here we expand the categorization of §2.1, and discuss the identified subcategories for sentence stress and prosodic breaks, which are 4 and 6 respectively. Intonation patterns and Politeness do not have subcategories. For emotional prosody we have 15 emotion pairs[14], thus having 15 subcategories. Examples are available at Tables 5 and 6.

A.1 Sentence Stress Subcategories

(1.1) *Contrastive Stress*, which highlights differences or corrects previous statements, emphasizing contrasts between elements (Bolinger, 1961).
(1.2) *New vs. Given Information*, which differentiates between new and given information, emphasizing what is considered new (Halliday, 1967).
(1.3) *Relational vs. Descriptive Adjectives*, where stressing the adjective or the noun can differentiate between the relational and descriptive uses of attributive adjectives (Liberman and Sproat, 1992).
(1.4) *Focus-Sensitive Operators*, where stress indicates the focus of adverbs of quantification (*only*, *just*, etc), shifting the meaning of the sentence accordingly (Halliday, 1967; Jackendoff, 1972).

A.2 Prosodic Break Subcategories

(2.1) *Direct vs. Indirect Statements*, where a prosodic break can indicate whether a phrase is a direct or an indirect quote (Klewitz and Couper-Kuhlen, 1999; Jansen et al., 2001).
(2.2) *Restrictive vs. Non-Restrictive Clauses*, which involves the use of prosodic breaks to differentiate between essential and non-essential information, impacting the specificity of the noun being described (Nespor and Vogel, 1986).
(2.3) *VP vs. NP Attachment*, where a trailing phrase can be attached either to the verb-phrase or the noun-phrase, depending on the existence of a prominent prosodic break (Pynte, 1996).
(2.4) *Particle vs. Preposition*, where a prosodic break can disambiguate between the literal and idiomatic meaning of phrasal verbs, by grouping the preposition with or without it (Price et al., 1991).
(2.5) *Broad vs. Narrow Scope*, where the existence of a prosodic break can signal that a modifier (adjective) has narrow scope, and refers only to one of two nouns that follow it (Hirschberg, 2017).
(2.6) *Complementizer vs. Parenthetical*, where the location of a prosodic break indicates whether an intermediate phrase acts as a complementizer or simply parenthetical to the main one (Dehé, 2014).

[14] Removed *fearful* emotion due to issues with the TTS.

B Examples for In-context Learning

In Tables 5, 6 and 7 we present some of the examples used for in-context learning when generating new examples with GPT-4 (§2.2).

C Quality Assessment for TTS candidates

Here we present the objectives we defined for assessing the quality of the generated speech candidates for each contrastive example. The objective is applied only to candidates that had WER $= 0$ using WHISPER. If all candidates are invalid for a prosodic case, the whole example is removed. We also defined some threshold levels for the objectives after trial-and-error, in order to remove examples where the best candidate was below it.

Sentence Stress. We use forced-alignment with WAV2VEC 2.0 (Baevski et al., 2020) to obtain the segment for each word in the signal, and extract their loudness, pitch and duration features. Then we define the stress level *stress* for a word w as the weighted sum of these three features. Finally we select the best candidate according to a simple objective obj_{stress} that has three goals: (1) maximize the stress of the target word ($stress_{tgt}$), (2) minimize the stress of the target word of the contrastive case ($stress_{foil}$), and (3) minimize the average stress of the rest.

$$stress_w = \lambda_1 loud_w + \lambda_2 pitch_w + \lambda_3 dur_w$$

$$obj_{stress} = 2 \cdot stress_{tgt} - stress_{foil} - \frac{1}{n-1} \sum_{w \neq tgt} stress_w,$$

where we used $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.2$. Note that in the sentence stress examples, there is

1.1 Contrastive Stress	
Sentence	She didn't give the book to John.
Prosody_A	She didn't give the *BOOK* to John.
Meaning_A	Something else was given to John.
Prosody_B	She didn't give the book to *JOHN*.
Meaning_B	The book was given to someone else.
1.2 New vs. Given Information	
Sentence	The committee decided to postpone the meeting.
Prosody_A	The *COMMITTEE* decided to postpone the meeting.
Meaning_A	Given: Someone decided to postpone the meeting; New: It was the committee who decided.
Prosody_B	The committee decided to *POSTPONE* the meeting.
Meaning_B	Given: The committee decided something; New: The decision was to postpone it.
1.3 Relational vs. Descriptive Adjectives	
Sentence	They are German teachers.
Prosody_A	They are *GERMAN* teachers.
Meaning_A	Teachers who teach the German language. (Relational)
Prosody_B	They are German *TEACHERS*.
Meaning_B	Teachers who are German. (Descriptive)
1.4 Focus-Sensitive Operators	
Sentence	I only introduced John to Maria at yesterday's party.
Prosody_A	I only introduced *JOHN* to Maria at yesterday's party.
Meaning_A	John was the only person I introduced to Maria.
Prosody_B	I only introduced John to *MARIA* at yesterday's party.
Meaning_B	Maria was the only person I introduced John to.

Table 5: Examples in the category *Sentence Stress* that were used for in-context learning.

always exactly 1 target word in each contrastive prosodic case.

Prosodic Breaks. Likewise, after forced-alignment, we measure the duration *dur* of each gap l between the words in the utterance, and define a similar objective obj_{break} as:

$$obj_{break} = 2\frac{1}{|tgt|}\sum_{l \in tgt} dur_l - \frac{1}{|foil|}\sum_{l \in foil} dur_l - \frac{1}{n - |tgt|}\sum_{l \notin tgt} dur_l$$

In this category, there can be 0 to 2 breaks in each prosodic case, which could be shared between the two prosodic cases. In the objective we consider only the ones that are not common in the two cases.

Intonation Patterns. We use teacher-forcing with WHISPER to extract the punctuation probabilities given the transcription text without the ending punctuation. The probability of the sentence to be a statement is the sum of the probabilities of the tokens "." and "!", while the probability of a question is the probability of the token "?". Thus the objective obj_{inton} for a statement is defined as:

$$obj_{inton} = p(. \mid X, Z_{<n}) + p(! \mid X, Z_{<n}) - p(? \mid X, Z_{<n}),$$

where X is the speech signal and $Z_{<n}$ are the tokens of the transcription, excluding the final one, which corresponds in all cases of this category. to the punctuation. The negative objective $-obj_{inton}$ is used for a case that is a question.

Emotional Prosody. We employ an emotion classifier[15] which is a based on a finetuned WAV2VEC 2.0 on the RAVDESS dataset (Livingstone and Russo, 2018), and define the objective as:

$$obj_{emo} = p(e_{tgt} \mid X) - p(e_{foil} \mid X),$$

where θ are the parameters of the classifier, e_{tgt} is the target emotion label and e_{foil} is the emotion label of the other prosodic case.

Pragmatic Prosody. To the best of our knowledge there is no open-sourced audio classifier to detect politeness levels, thus we re-purpose the emotion classifier and define the probabilities of politeness

[15] hf.co/ehcalabres/wav2vec2-lg-XLS-R-en-speech-emotion-recognition

2.1 Direct vs. Indirect Statements	
Sentence	Alex announced Jamie will meet the manager.
Prosody_A	Alex *ANNOUNCED* | Jamie will meet the manager.
Meaning_A	(Direct Statement)
Prosody_B	Alex announced Jamie will meet the manager.
Meaning_B	(Indirect Statement)
2.2 Restrictive vs. Non-Restrictive Phrases	
Sentence	The students who were talking were sent out.
Prosody_A	The students who were *TALKING* | were sent out.
Meaning_A	Only the students who were talking were actually sent out. (Restrictive)
Prosody_B	The *STUDENTS* | who were talking | were sent out.
Meaning_B	All students were sent out, and the fact they were talking is additional information. (Non-restrictive)
2.3 Verb-phrase vs. Noun-phrase Attachment	
Sentence	Paula phoned her friend from Alabama.
Prosody_A	Paula phoned her friend | from *ALABAMA*.
Meaning_A	Paula called her friend while she was in Alabama. (VP Attachment)
Prosody_B	Paula phoned | her *FRIEND* from Alabama.
Meaning_B	Paula phoned her friend who is from Alabama. (NP Attachment)
2.4 Phrasal Verbs	
Sentence	John laughed at the party.
Prosody_A	John *LAUGHED* | at the party.
Meaning_A	John laughed while he was at the party. (Literal)
Prosody_B	John *LAUGHED AT* | the party.
Meaning_B	John made fun of the party. (Idiomatic)
2.5 Complementizer vs. Parenthetical	
Sentence	We only suspected they all knew that a burglary had been committed.
Prosody_A	We only *SUSPECTED* | they all knew that a burglary had been committed.
Meaning_A	The suspicion was that they all knew about the burglary. (Complementizer)
Prosody_B	We only suspected | they all *KNEW* | that a burglary had been committed.
Meaning_B	They all knew that we only suspected that a burglary had been committed. (Parenthetical)
2.6 Modifier Scope	
Sentence	This collar is dangerous to younger dogs and cats.
Prosody_A	This collar is dangerous to *YOUNGER* dogs and cats.
Meaning_A	Younger refers to both dogs and cats. (Broad Scope)
Prosody_B	This collar is dangerous to *YOUNGER* dogs | and *CATS*.
Meaning_B	Younger refers only to dogs. (Narrow Scope)

Table 6: Examples in the category *Prosodic Breaks* that were used for in-context learning.

and impoliteness as a weighted sum of the 8 available emotion classes.

$$p(\textit{polite}) = \frac{\sum_e w_e p(e \mid X)}{\sum_e w_e},$$

and similarly for impolite. We used the weighted scheme displayed in Table 8, which was obtained by prompting GPT-4.

D Data

D.1 Data Statistics

In Table 9 we provide the analytic data statistics for each category/subcategory, throughout the generation process stages. The poor quality of the cTTS, where prosody was not always encoded in the speech, led us to remove a large percentage of the examples before translating them. Also many examples where removed because the oracle translations for both cases were the same.

D.2 Overly Explanatory Examples

In Table 10 we present two examples where GPT-4 acting as an oracle translator (§2.3) proposed overly explanatory translations in the emotional prosody category. Both are inline with the emotion of the speaker, but they contain new bits of information, not initially there. These were removed in filtering due to excessive word-length ratio between the two cases.

	3. Intonation Patterns
Sentence	You can solve this problem
Prosody_A	You *CAN* solve this problem.
Meaning_A	Encouraging or asserting the person's ability to solve this problem.
Prosody_B	You _can_ solve this problem?
Meaning_B	Questioning the person's ability to solve this problem.
	4. Emotional Prosody (Happy/Sad)
Sentence	The surgery went as expected.
Prosody_A	<happy> The surgery went *AS EXPECTED*!
Meaning_A	The surgery's successful outcome aligns with hopes and predictions, leading to joy and relief.
Prosody_B	<sad> The surgery went _as expected_ ...
Meaning_B	The expected outcome was not favorable, leading to a somber tone.
	4. Emotion Prosody (Fearful/Angry)
Sentence	Can we talk about this later?
Prosody_A	<fearful> Can we... talk about this... later?
Meaning_A	Indicates hesitation or fear about the topic, or the situation in general.
Prosody_B	<angry> Can we *TALK* about this later!?
Meaning_B	Implies urgency or frustration, and a demand for immediate attention.
	5. Politeness
Sentence	Can you move your car?
Prosody_A	<polite> Can you _move_ your car?
Meaning_A	A polite request to move the car.
Prosody_B	<impolite> Can you *MOVE* your *CAR*?!
Meaning_B	A rude demand to move the car, with an aggressive tone.

Table 7: Examples in the categories *Intonation Patterns*, *Emotional Prosody*, *Politeness*, and that were used for in-context learning.

Emotion	Politeness	Impoliteness
Happy	0.3	-0.1
Calm	0.3	-0.2
Neutral	0.2	0.1
Surprised	0.1	0.1
Sad	0.0	0.2
Disgust	-0.1	0.3
Angry	-0.2	0.4
Fearful	-0.1	0.0

Table 8: Weighting scheme for Politeness and Impoliteness labels based on the emotion classifier.

E Evaluated Speech Translation Models

Here we describe in more detail the model families and the specific versions used. We evaluated in total 31 S2TT model variants. All models are available in the Transformers Huggingface Library (Wolf et al., 2020). For inference we used the default generation parameters and a beam search of 5.

1. SEAMLESSM4T (Seamless Communication, 2023a) and its updated version v2 (Seamless Communication, 2023b) is a recently proposed family of unified encoder-decoder models that are both multilingual (many-to-many, 100 languages) and multimodal (speech/text input or output), meaning they can carry out the tasks of ASR, TTS, MT, S2TT, and also S2ST. The architecture is composed of a text encoder, text decoder, speech encoder, and speech decoder, and different parts are active depending on the input/output modalities. The text encoder-decoder is based on NLLB (NLLB Team, 2022), the speech encoder on a newly proposed conformer (Gulati et al., 2020) W2V-BERT (Chung et al., 2021), and the speech decoder on a unit decoder (Inaguma et al., 2023) and a HiFi-GAN vocoder (Kong et al., 2020). The original version has a medium (1.2B)[16] and a large (2.3B)[17] variant, while the updated v2 has a large variant (2.3B)[18]. For cascade S2TT we first use the model in ASR mode, and then the

[16] hf.co/facebook/seamless-m4t-medium
[17] hf.co/facebook/seamless-m4t-large
[18] hf.co/facebook/seamless-m4t-v2-large

Category / Subcategory	Initial	Generated	Synthesised	Translated		
				De	Es	Ja
Contrastive Stress (General)	200	199	183	87	76	97
Relational/Descriptive Adjectives	200	199	147	42	33	51
Contrastive Stress (Noun-Phrase)	200	199	124	37	36	39
New/Given Information	200	197	146	51	65	91
Focus-sensitive Operators	200	181	118	60	42	64
Sentence Stress	1000	975	718	277	252	342
Complementizer/Parenthetical	200	200	171	59	46	73
VP/NP Attachment	200	200	66	23	18	20
Modifier Scope	200	200	200	83	107	81
Restrictive/Nonrestrictive	200	199	177	65	82	40
Direct/Indirect	200	198	154	41	25	70
Phrasal Verbs	42	42	17	5	1	5
Prosodic Breaks	1042	1039	785	276	279	289
Intonation Patterns	300	263	174	173	173	173
Sad-Happy	200	200	1	1	1	1
Neutral-Angry	200	199	185	123	111	119
Neutral-Happy	200	198	161	81	97	81
Disgust-Angry	200	198	18	4	5	3
Disgust-Sad	200	198	-	-	-	-
Neutral-Surprised	200	198	43	33	35	30
Disgust-Neutral	200	197	7	2	5	5
Happy-Angry	200	197	138	50	65	72
Sad-Surprised	200	197	3	2	2	2
Sad-Neutral	200	196	4	3	2	2
Sad-Angry	200	196	5	1	4	4
Disgust-Surprised	200	196	4	2	2	1
Disgust-Happy	200	195	10	5	7	6
Happy-Surprised	200	195	52	34	27	21
Angry-Surprised	200	193	68	32	34	30
Emotional Prosody	3000	2953	699	433	418	377
Politeness	400	375	387	212	193	206
Total	**5742**	**5605**	**2763**	**1311**	**1294**	**1386**

Table 9: Number of Examples by Category and Subcategory

same model is MT mode.

2. XLS-R (Babu et al., 2021) is a multilingual E2E S2TT model that is based on a multilingual WAV2VEC 2.0 (Baevski et al., 2020) trained with self-supervised learning on a large speech corpus on 128 languages. For S2TT, the encoder is coupled with the decoder from MBART50 (Tang et al., 2020), and finetuned on paired speech-translation data. We use the folowing versions that are finetuned on English-to-15 on CoVoST2 (Wang et al., 2021): 300M[19], 1B[20], and 2B[21].

3. ZEROSWOT is a zero-shot E2E S2TT model that softly connects a WAV2VEC 2.0 encoder and an NLLB model, by compressing the speech representation into subword units and Optimal Transport (Peyré and Cuturi, 2019) alignment, using only ASR data. The versions used here are based on NLLB that were finetuned on the text data of CoVoST2, and the ZEROSWOT model was trained on Com-

[19] hf.co/facebook/wav2vec2-xls-r-300m-en-to-15
[20] hf.co/facebook/wav2vec2-xls-r-1b-en-to-15
[21] hf.co/facebook/wav2vec2-xls-r-2b-en-to-15

Example 1: *This will only take a minute.*		
A	(neutral)	Das dauert nur eine Minute. (This will only take a minute.)
B	(angry)	Das dauert nur eine Minute, also machen Sie keinen Aufstand. (This will only take a minute so don't make a fuzz about it.)
Example 2: *Our case was dismissed.*		
A	(neutral)	Unser Fall wurde abgewiesen. (Our case was dismissed.)
B	(sad)	Unser Fall wurde abgewiesen und das macht mich fassungslos. (Our case was dismissed which is just perplexing.)

Table 10: Examples of overly explanatory translations proposed by GPT-4.

monVoice (Ardila et al., 2020). The MEDIUM version[22] has 1B parameters and the LARGE version[23] has 1.7B parameters.

4. SALMONN (Tang et al., 2024) is a general-purpose audio LLM that is capable of several speech- and audio-related tasks, including S2TT. It is build on top of the Vicuna LLM (Peng et al., 2023), and uses two encoders, one from WHISPER and one from BEATs (Chen et al., 2023). The concatenated output representations from the two encoders are processed by a Q-former (Li et al., 2023) and fed to the LLM which is finetuned with LoRA (Hu et al., 2022). There is a 7B version[24] and a 13B version[25]. To translate speech into a target language we use the recommended prompt from the paper: "Listen to the speech and translate it into {Target Language}".

5. WHISPER & NLLB is an AED-based cascade. WHISPER (Radford et al., 2022) is an encoder-decoder ASR and many-to-en S2TT model. We use three different versions for this casdade, namely the WHISPER-MEDIUM[26], the WHISPER-LARGE[27], and the latest v3 large version[28]. We primarily present results with the WHISPER-LARGE-V3, but since it was also used for filtering we also discuss v1 in order to avoid biasing our results. NLLB (NLLB Team, 2022) is a massively multilingual many-to-many MT model with access to 200 languages. We used the two distilled versions from the 54B MoE model, namely the distilled-600M[29] and the distilled-1.3B[30], as well as the 3.3B model[31]. We evaluated all possible combinations, thus having 9 cascade variants with these models.

6. CTC & NLLB is a CTC-based cascade. We use three different CTC encoders for the cascades. The first one is the Large version (300M) of WAV2VEC 2.0[32] which is finetuned on Libri-Light (Kahn et al., 2020) and Librispeech (Panayotov et al., 2015), additionally using self-training (Xu et al., 2020). The second is the Large version (300M) of HUBERT[33] (Hsu et al., 2021), finetuned on Librispeech. The third is also based on HUBERT, more specifically to the XL version[34] with 1B parameters. We use the same three versions of NLLB, as we did for the AED-based cascade, thus having in total 9 variants of the CTC-based cascade.

F Supplementary Results

In Figure 5 we present the Spearman rank correlation for the four contrastive metrics and the standard evaluation metric XCOMET. They were computed by evaluating all 31 models (§E) for all 3 language pairs, thus having a total of 93 observations.

In Table 11 we present the global contrastive quality scores for all 31 S2TT models for the 3 language pairs, which were used for the analysis of Figure 2 in §5 of the main text.

In Figures 6 and 7 we present the comparisons of the 4 models for Spanish and Japanese, similar to what we did in Figure 3 for German in the main text. In general, the findings and observations here coincide with those for German.

[22] hf.co/johntsi/ZeroSwot-Medium-cv-covost2-en-to-15
[23] hf.co/johntsi/ZeroSwot-Large-cv-covost2-en-to-15
[24] hf.co/tsinghua-ee/SALMONN-7B
[25] hf.co/tsinghua-ee/SALMONN
[26] hf.co/openai/whisper-medium
[27] hf.co/openai/whisper-large
[28] hf.co/openai/whisper-large-v3
[29] hf.co/facebook/nllb-200-distilled-600M
[30] hf.co/facebook/nllb-200-distilled-1.3B
[31] hf.co/facebook/nllb-200-3.3B
[32] hf.co/facebook/wav2vec2-large-960h-lv60-self
[33] hf.co/facebook/hubert-large-ls960-ft
[34] hf.co/facebook/hubert-xlarge-ls960-ft

Model	Model Type	Model Size (B)	Contrastive Quality (Global)			
			En-De	En-Es	En-Ja	Average
SeamlessM4T-v1-Medium	E2E	1.2	9.1	10.4	10.1	9.9
SeamlessM4T-v1-Large	E2E	2.3	7.8	8.7	8.2	8.2
SeamlessM4T-v2-Large	E2E	2.3	14.5	11.0	13.9	13.1
XLS-R 300M	E2E	0.3	10.2	9.3	9.5	9.6
XLS-R 1B	E2E	1.0	9.3	8.3	8.1	8.6
XLS-R 2B	E2E	2.0	7.3	8.4	7.0	7.6
ZeroSwot-Medium	E2E	0.9	8.9	7.5	7.5	8.0
ZeroSwot-Large	E2E	0.9	8.7	7.7	7.9	8.1
SALMONN-7B	E2E	7.0	12.7	9.4	12.9	11.7
SALMONN-13B	E2E	13.0	**15.9**	**12.3**	**16.1**	**14.8**
SeamlessM4T-v1-Medium	Cascade-AED	2.4	7.3	7.6	6.9	7.2
SeamlessM4T-v1-Large	Cascade-AED	4.6	6.5	5.8	4.3	5.5
SeamlessM4T-v2-Large	Cascade-AED	4.6	10.5	7.7	8.7	8.9
Whisper-v1-Medium & NLLB-600M	Cascade-AED	1.4	5.7	5.0	5.4	5.4
Whisper-v1-Medium & NLLB-1.3B	Cascade-AED	2.1	6.2	5.0	5.8	5.6
Whisper-v1-Medium & NLLB-3.3B	Cascade-AED	4.1	6.6	5.1	5.4	5.7
Whisper-v1-Large & NLLB-600M	Cascade-AED	2.2	5.9	4.9	6.1	5.6
Whisper-v1-Large & NLLB-1.3B	Cascade-AED	2.9	6.0	4.7	5.8	5.5
Whisper-v1-Large & NLLB-3.3B	Cascade-AED	4.9	6.3	4.6	5.6	5.5
Whisper-v3-Large & NLLB-600M	Cascade-AED	2.2	5.3	4.6	6.5	5.5
Whisper-v3-Large & NLLB-1.3B	Cascade-AED	2.9	5.3	4.7	5.4	5.1
Whisper-v3-Large & NLLB-3.3B	Cascade-AED	4.9	5.5	4.8	5.3	5.2
wav2vec 2.0 & NLLB-600M	Cascade-CTC	0.9	1.4	1.3	2.0	1.5
wav2vec 2.0 & NLLB-1.3B	Cascade-CTC	1.6	1.7	1.0	1.4	1.3
wav2vec 2.0 & NLLB-3.3B	Cascade-CTC	3.6	1.6	0.9	1.7	1.4
HuBERT & NLLB-600M	Cascade-CTC	0.9	3.2	2.7	2.5	2.8
HuBERT & NLLB-1.3B	Cascade-CTC	1.6	2.2	2.4	2.9	2.5
HuBERT & NLLB-3.3B	Cascade-CTC	3.6	2.7	2.6	2.9	2.7
HuBERT-xl & NLLB-600M	Cascade-CTC	1.6	2.4	2.9	3.0	2.8
HuBERT-xl & NLLB-1.3B	Cascade-CTC	2.3	3.5	1.7	2.7	2.6
HuBERT-xl & NLLB-3.3B	Cascade-CTC	4.3	2.6	2.4	2.5	2.5

Table 11: Contrastive Quality (Global) scores for English-German, English-Spanish, and English-Japanese, including their averages.

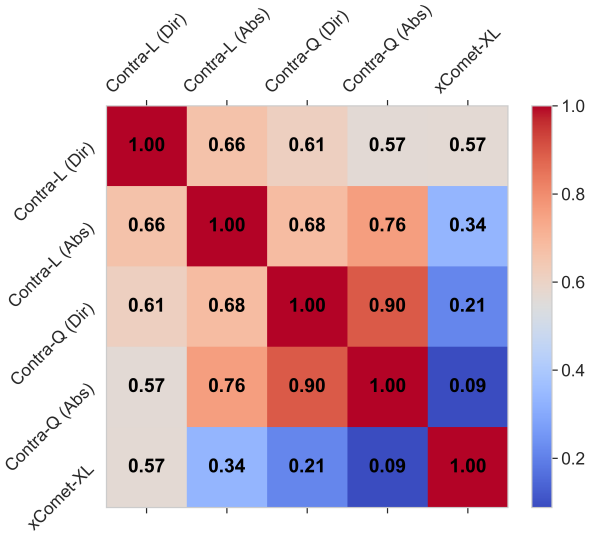

Figure 5: Correlation Matrix of the metrics across all language pairs and models.

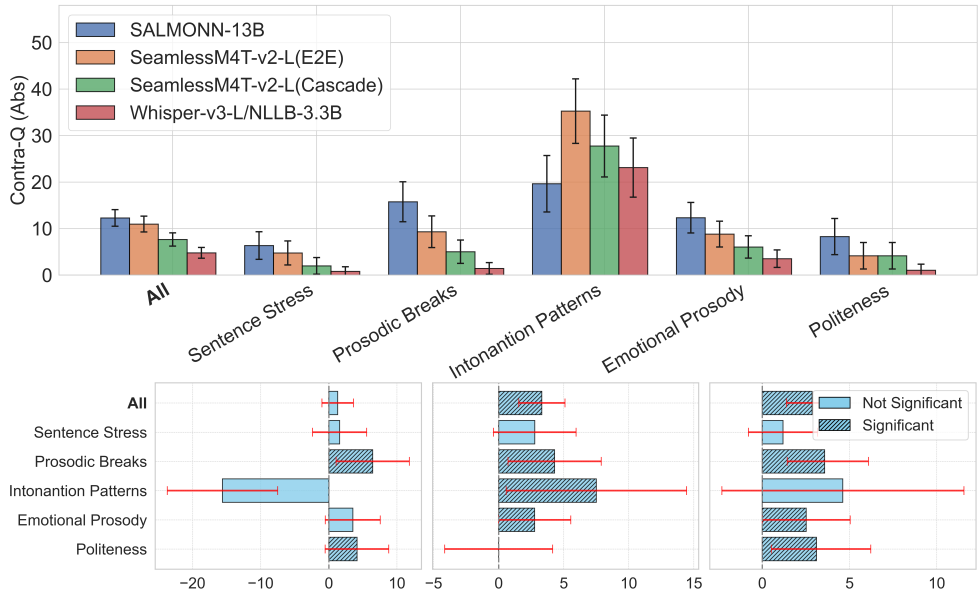

Figure 6: Upper: Model performance per category (En-Es). Lower: Model performance comparisons (En-Es), (a): SALMONN-13B vs. SEAMLESSM4T-V2-LARGE, (b) SEAMLESSM4T-V2-LARGE(E2E) vs. SEAMLESSM4T-V2-LARGE(cascade), (c) SEAMLESSM4T-V2-LARGE(cascade) vs. WHISPER-V3-LARGE/NLLB-3.3B.

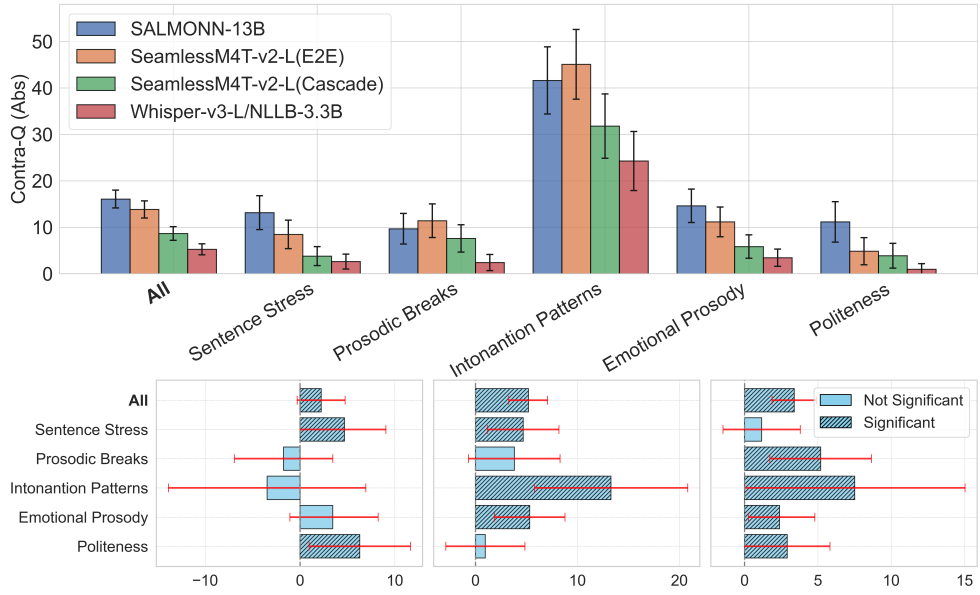

Figure 7: Upper: Model performance per category (En-Ja). Lower: Model performance comparisons (En-Ja), (a): SALMONN-13B vs. SEAMLESSM4T-V2-LARGE, (b) SEAMLESSM4T-V2-LARGE(E2E) vs. SEAMLESSM4T-V2-LARGE(cascade), (c) SEAMLESSM4T-V2-LARGE(cascade) vs. WHISPER-V3-LARGE/NLLB-3.3B.