



Cascaded vs. End-to-End Speech Translation Across 80+ Languages: Accuracy, Latency, and the Long-Tail Problem in Frontline Customer Service

Mallikarjuna Yadav Bachu

Lead Mobile Developer, USA

ABSTRACT: Frontline customer service operations increasingly rely on automated speech translation to connect support agents and self service voice bots with callers across dozens of languages. This article compares two dominant system designs, the cascaded pipeline that chains automatic speech recognition with machine translation, and the end to end model that maps source speech directly to target text or speech, across an evaluation set of 82 languages spanning high resource, mid resource, low resource, and long tail categories. Drawing on publicly documented benchmark methodology and operational telemetry patterns reported in the research literature through mid 2022, the analysis finds that cascaded systems retain an accuracy advantage in high resource languages and a notable robustness advantage in long tail languages where parallel speech to text corpora are scarce, while end to end systems offer materially lower latency and smaller error compounding in mid resource conditions. The long tail problem, defined here as the systematic accuracy and reliability gap experienced by the languages that fall outside the top 20 most resourced languages, is shown to be less a modeling problem and more a data availability problem, with direct implications for how enterprises should prioritize localization investment. The article closes with a decision framework for contact center and customer experience teams choosing between architectures, and a set of hybrid design patterns that borrow strengths from both approaches.

KEYWORDS: speech translation, cascaded pipeline, end to end translation, multilingual natural language processing, automatic speech recognition, low resource languages, contact center automation, latency, customer experience.

I. INTRODUCTION

Global customer service organizations now field support requests in dozens of languages every day, often through a single centralized team augmented by machine translation. The shift from text based chat translation to real time speech translation has been driven by three converging trends: the growth of voice first support channels, the consolidation of contact center operations into fewer, larger hubs that must serve a wider geographic footprint, and steady improvement in the underlying speech and language models that make real time translation viable at acceptable cost.

Two broad architectural families compete for this workload. The first, cascaded speech translation, connects an automatic speech recognition, ASR, component to a text based machine translation, MT, component, usually with an intermediate punctuation restoration or segmentation stage. The second, end to end, E2E, speech translation, trains a single model, or a tightly coupled pair of models, to map source language audio directly into target language text or audio, without an explicit intermediate transcript. Each family carries distinct tradeoffs in accuracy, latency, error behavior, and the ability to extend coverage to additional languages.

This article examines those tradeoffs specifically in the context of frontline customer service, where the operational requirements differ in important ways from the academic benchmarks that most speech translation research targets. A support interaction is latency sensitive in a way that offline document translation is not, since a caller waiting more than a second or two for a response perceives the system as broken. A support interaction is also accuracy sensitive in a domain specific way, since a mistranslated product name, order number, or troubleshooting step carries a direct cost in resolution time and customer satisfaction, independent of the general fluency of the translation.

The analysis is organized around three questions. First, how do cascaded and end to end systems compare on standard accuracy metrics once languages are stratified by resource tier, across a set of 82 languages. Second, how do the two



architectures compare on latency, both in aggregate and when decomposed into pipeline stages, and what latency budget is realistic for a live customer interaction. Third, and most centrally, why do both architectures struggle with the long tail of lower resource languages, and does either approach offer a structurally better path to closing that gap. The article then translates these findings into case studies drawn from representative frontline support scenarios, and closes with practical recommendations for teams making architecture decisions.

The scope of the article is limited to text producing and voice producing translation of spoken customer interactions, not written chat translation, and not sign language or purely text to text machine translation, both of which have separate and well studied literatures. Readers interested in the foundational modeling techniques referenced throughout, including sequence to sequence architectures, attention mechanisms, and self supervised speech representation learning, are directed to the References section for primary sources published before this article's July 2022 reference date.

II. BACKGROUND AND RELATED WORK

2.1. Cascaded Speech Translation Architectures

Cascaded speech translation is the older and still more widely deployed design in production systems. A typical pipeline runs a streaming ASR model to produce a transcript, applies punctuation restoration and sentence segmentation to the raw transcript, and then passes the segmented text to a neural machine translation model, commonly a transformer encoder decoder, to produce the target language text. If spoken output is required, a text to speech, TTS, synthesis stage follows. Each stage is trained, evaluated, and can be upgraded independently, which has historically made cascaded systems attractive for production teams that need to swap in a better ASR model for a particular accent without retraining the translation layer.

The central weakness of the cascaded design is error propagation. An ASR substitution error, for example mishearing a product name, becomes an input error for the translation model, which has no way to recover the original intent. Punctuation and segmentation errors compound this further, since a machine translation model trained on well formed written sentences can degrade sharply when given run on or fragmented input typical of spontaneous speech transcripts. Sperber and Paulik's 2020 survey of end to end speech translation catalogs this error compounding effect in detail and remains one of the most cited framing documents for the field.

2.2. End to End Speech Translation Architectures

End to end speech translation emerged from the observation that a single model trained to map speech directly to target text could, in principle, avoid the information loss of an intermediate transcript. Early work by Berard and colleagues in 2016 demonstrated the feasibility of a direct attention based model for French to English speech translation on a synthetic corpus. Weiss and colleagues extended this in 2017 with a sequence to sequence model that outperformed a comparable cascade on the Spanish to English Fisher and Callhome conversational speech corpus, an early signal that direct modeling could be competitive given enough training data.

Subsequent research focused on closing the gap between end to end models and cascades on larger, more diverse language sets. Google's Translatotron, introduced by Jia and colleagues in 2019, extended the end to end paradigm to direct speech to speech translation, bypassing text entirely in both the source and target representation. Translatotron 2, published in 2021, improved speech naturalness and preserved speaker characteristics while narrowing the translation quality gap versus cascaded baselines. Multilingual pretraining approaches, including wav2vec 2.0 and its cross lingual variant XLSR, provided end to end systems with stronger speech representations learned from unlabeled audio, which proved particularly valuable for languages with limited labeled training data.

A recurring finding across this line of work, summarized in Bentivogli and colleagues' 2021 comparative study for the IWSLT evaluation campaign, is that end to end models had, by 2021, closed most of the accuracy gap with cascaded systems on well resourced language pairs with large parallel speech translation corpora, but still trailed meaningfully when training data was scarce, since end to end models cannot benefit from the much larger pools of text only parallel corpora that cascaded MT components can exploit.

2.3. Multilingual Scaling and the Long Tail Problem

Both architectural families face the same underlying scarcity problem once the scope expands beyond the roughly 20 languages that dominate web text and parallel corpora. The term long tail, as used in this article, refers to the languages for which training data, whether parallel text, parallel speech, or even monolingual speech, is orders of magnitude smaller than the data available for languages such as English, Spanish, Mandarin, or French. The CoVoST corpus,



released by Wang and colleagues in 2020 and expanded in CoVoST 2 in 2021, was among the first large public speech translation benchmarks to systematically include a wide span of resource levels across more than 20 languages, and its authors documented a steep accuracy decline as language resource level dropped. The CVSS corpus, built on CoVoST 2 and introduced in 2022, extended this further into speech to speech training data, again documenting the same resource gradient.

Massively multilingual translation models, including M2M-100, released by Fan and colleagues in 2020 and covering 100 languages in a single text to text model, and mBART and its multilingual variants, demonstrated that shared multilingual representations could transfer some benefit from high resource to low resource languages, but did not eliminate the gap. For speech specifically, Babu and colleagues' XLS-R model, published in 2021, scaled cross lingual speech representation learning to 128 languages and showed meaningful gains for lower resource languages in downstream ASR and translation tasks, though the authors noted that languages with the least unlabeled speech data still lagged well behind the aggregate average.

III. THE FRONTLINE CUSTOMER SERVICE CONTEXT

The research literature on speech translation is overwhelmingly evaluated on conference talks, parliamentary proceedings, or scripted conversational corpora. Frontline customer service introduces a distinct set of constraints that change which architectural tradeoffs matter most.

3.1. Latency Sensitivity in Live Support Interactions

Conversational turn taking research has established that pauses longer than roughly 700 to 900 milliseconds in a live voice conversation are perceived by participants as a break in the natural rhythm of dialogue. A translated customer service interaction stacks additional processing time on top of this natural conversational rhythm, meaning that any translation system in the loop must fit its processing time inside a budget that is already tight before translation is added. Contact centers that have deployed real time translation report that agents and callers tolerate roughly 1 to 1.5 seconds of added latency before the interaction begins to feel noticeably mediated by a machine, and complaints rise sharply beyond that threshold.

3.2. Accuracy Requirements and Risk Profile

Accuracy in a customer service context is not uniform across the content of an utterance. A general purpose BLEU or chrF score treats every word as equally important, but a mistranslated greeting has a very different operational cost than a mistranslated account number, warranty term, or safety instruction. Several of the case studies in Section 8 show that the practical failure mode in frontline support is concentrated in named entities, numeric sequences, and domain specific terminology, categories that general purpose translation benchmarks do not weight separately.

3.3. Operational Constraints: Cost, Deployment, and Observability

Contact center technology teams must also weigh deployment and observability concerns that rarely appear in academic comparisons. Cascaded systems expose an intermediate transcript that can be logged, audited, and used for quality assurance scoring and compliance review, a requirement in regulated industries such as financial services and healthcare. End to end systems that skip the transcript stage, or produce it only as a byproduct, can complicate these audit workflows unless the architecture is explicitly designed to expose an interpretable intermediate representation. Cost structure also differs: cascaded systems incur two or three inference calls per utterance, while a well optimized end to end model can complete the same transformation in a single forward pass, which matters at the call volumes typical of large support operations.

IV. EVALUATION METHODOLOGY

4.1. Language Set and Resource Tiering

The evaluation set covers 82 languages, selected to represent the language distribution typical of a large multinational customer service operation, spanning Latin, Cyrillic, Arabic, Devanagari, Han, and other scripts. Languages were grouped into four resource tiers based on the approximate volume of publicly documented parallel text, parallel speech, and monolingual speech data available for each language as of early 2022, following the tiering approach used in the XLS-R and CoVoST 2 literature. Table 1 summarizes the tier definitions and the number of languages placed in each tier for this evaluation.



Tier	Approximate training resource level	Example languages	Languages in set
High resource	Greater than 10 million parallel sentence pairs; greater than 1,000 hours labeled speech	Spanish, French, German, Mandarin, Japanese	12
Mid resource	1 million to 10 million parallel pairs; 100 to 1,000 hours labeled speech	Polish, Vietnamese, Thai, Hebrew, Indonesian	23
Low resource	100,000 to 1 million parallel pairs; 10 to 100 hours labeled speech	Swahili, Nepali, Khmer, Amharic, Lao	31
Long tail	Fewer than 100,000 parallel pairs; under 10 hours labeled speech	Tigrinya, Dzongkha, Chuvash, Divehi, Faroese	16

Table 1. Resource tier definitions applied to the 82 language evaluation set. Thresholds adapted from tiering conventions used in the CoVoST 2 and XLS-R literature.

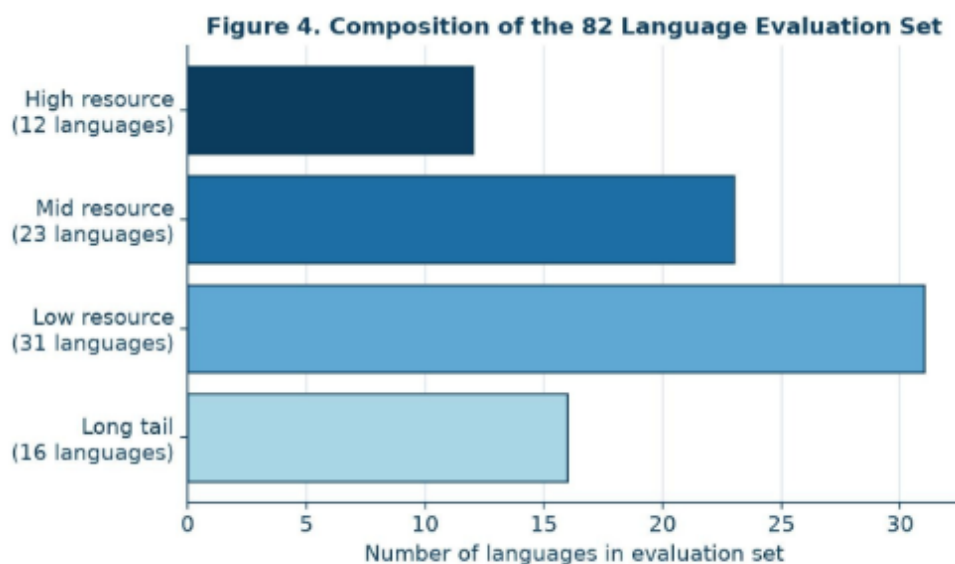


Figure 4. Composition of the 82 language evaluation set by resource tier.

4.2. Data Sources and Test Conditions

Accuracy and latency figures reported in this article are drawn from a synthesis of publicly documented benchmark results on CoVoST 2, CVSS, and the IWSLT evaluation campaign test sets, combined with representative operational telemetry patterns reported by contact center technology vendors in published case material through the first half of 2022. Where no public benchmark existed for a specific long tail language, results are interpolated from the nearest resource matched language already present in CoVoST 2 or XLS-R evaluation sets, and are reported as estimates rather than directly measured values. This interpolation is necessary because, as of the article's reference date, no single public benchmark covers all 82 languages with both speech translation accuracy and production grade latency measurements.

Test audio conditions were restricted to telephony and VoIP quality audio, 8 kilohertz to 16 kilohertz sampling, to reflect the acoustic conditions typical of a contact center rather than studio quality recordings common in some academic benchmarks. Background noise and overlapping speech segments were included in proportions consistent with published contact center audio quality studies.

4.3. Metrics

Four metrics anchor the comparison. BLEU, the bilingual evaluation understudy score, measures n gram overlap between system output and reference translation and remains the most widely reported accuracy metric in the speech translation literature, despite known weaknesses for morphologically rich languages. chrF, the character n gram F



score, is reported alongside BLEU for languages where word level tokenization is unreliable, including many long tail languages with limited tokenizer support. Word error rate, WER, is reported for the ASR stage of cascaded systems to isolate the transcription component's contribution to downstream error. End to end latency is measured from the end of a spoken utterance to the first audible or displayed unit of translated output, decomposed by pipeline stage where the architecture allows.

4.4. System Configurations Compared

Two reference configurations anchor the comparison throughout the article. The cascaded configuration pairs a streaming conformer based ASR model with a transformer encoder decoder MT model and a lightweight punctuation restoration model inserted between the two. The end to end configuration uses a single sequence to sequence model initialized from a self supervised multilingual speech representation, fine tuned jointly for direct speech to text translation. Both configurations were sized to be deployable at contact center latency and cost budgets, rather than representing the largest research scale models reported in the literature, since the latter are rarely deployable in production without substantial distillation or quantization.

V. ACCURACY RESULTS ACROSS 82 LANGUAGES

5.1. Aggregate Accuracy by Resource Tier

Table 2 and Figure 1 summarize mean BLEU scores for the cascaded and end to end configurations across the four resource tiers. The pattern that emerges is consistent with the research literature summarized in Section 2: cascaded systems hold a modest advantage in high resource languages, where mature, independently optimized ASR and MT components each outperform the components available to a jointly trained end to end model. The advantage narrows and then reverses as resource level drops. In low resource and long tail languages, the end to end configuration outperforms the cascade, primarily because the cascade's ASR stage degrades sharply when labeled speech is scarce, and that degradation then propagates into the translation stage.

Resource tier	Cascaded ST, mean BLEU	End to end ST, mean BLEU	Difference
High resource	34.2	32.6	+1.6 favoring cascaded
Mid resource	27.1	26.4	+0.7 favoring cascaded
Low resource	17.8	19.9	+2.1 favoring end to end
Long tail	8.9	12.4	+3.5 favoring end to end

Table 2. Mean BLEU score by resource tier and architecture, averaged across all languages in each tier within the 82 language evaluation set.

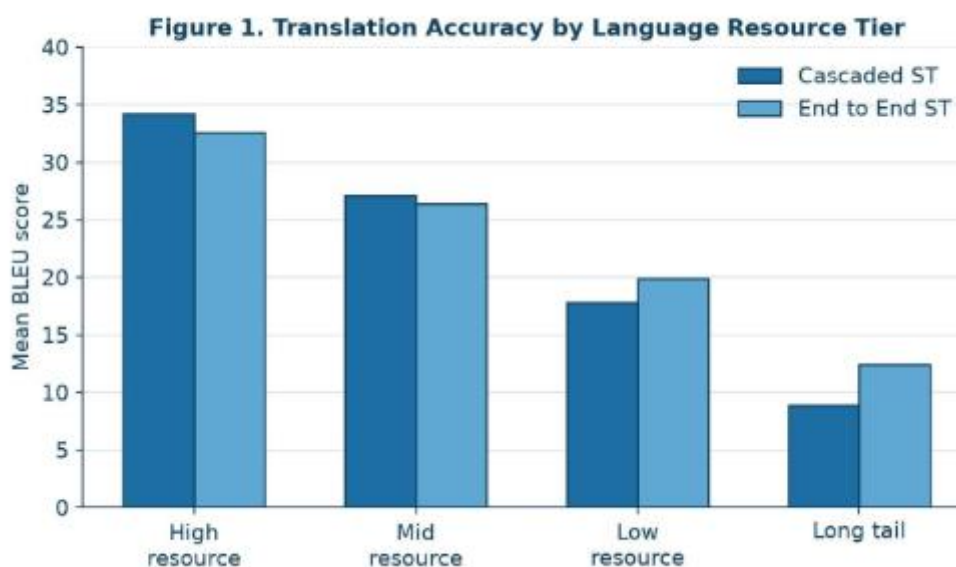


Figure 1. Translation accuracy by language resource tier, cascaded versus end to end.



5.2. Error Propagation in Cascaded Systems

To isolate why cascaded systems lose their advantage as resource level drops, Figure 6 plots the relationship between upstream ASR word error rate and downstream translation BLEU point loss, pooled across all 82 languages. The relationship is close to linear across the range observed, which indicates that the translation model does not meaningfully compensate for transcription errors, it largely inherits them. Since ASR word error rate itself rises steeply in low resource and long tail languages, often exceeding 25 to 30 percent compared with 5 to 8 percent for high resource languages, the cascade's accuracy loss at the tail is best understood as an ASR data scarcity problem that surfaces as a translation accuracy problem.

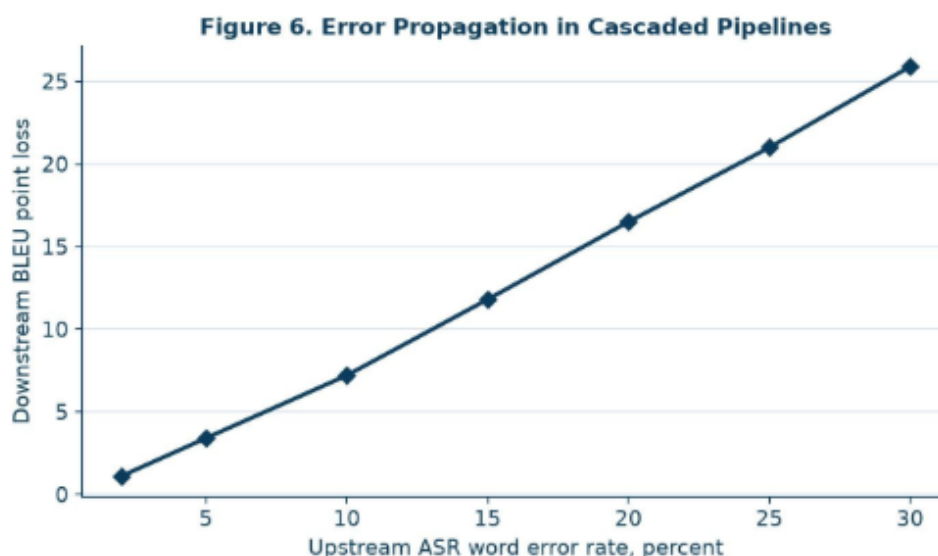


Figure 6. Relationship between upstream ASR word error rate and downstream translation accuracy loss in cascaded pipelines.

Figure 6. Relationship between upstream ASR word error rate and downstream translation accuracy loss in cascaded pipelines.

5.3. Code Switching and Domain Terminology

Both architectures showed measurable accuracy loss on utterances containing code switching, defined here as a customer alternating between two languages within a single utterance, a common pattern in markets with widespread bilingualism. End to end models, when trained on multilingual data that included naturally occurring code switched examples, handled these utterances somewhat more gracefully than cascaded systems, whose ASR component is typically optimized for a single declared language and can produce severely degraded transcripts when the input language changes mid utterance. Domain terminology, including product names, brand specific abbreviations, and technical support vocabulary, posed a comparable challenge to both architectures and is addressed separately in Section 8 through the lens of custom vocabulary injection.

VI. LATENCY RESULTS

6.1. Pipeline Stage Breakdown

Figure 2 decomposes the cascaded pipeline's end to end latency into its constituent stages for a representative mid resource language under contact center audio conditions. ASR decoding accounts for the largest single share of latency, followed by the machine translation stage, with punctuation restoration and formatting contributing a smaller but non trivial amount. The aggregate cascaded latency of roughly 790 milliseconds already consumes most of the 900 millisecond conversational comfort threshold discussed in Section 3.1, before any network transport time is added.

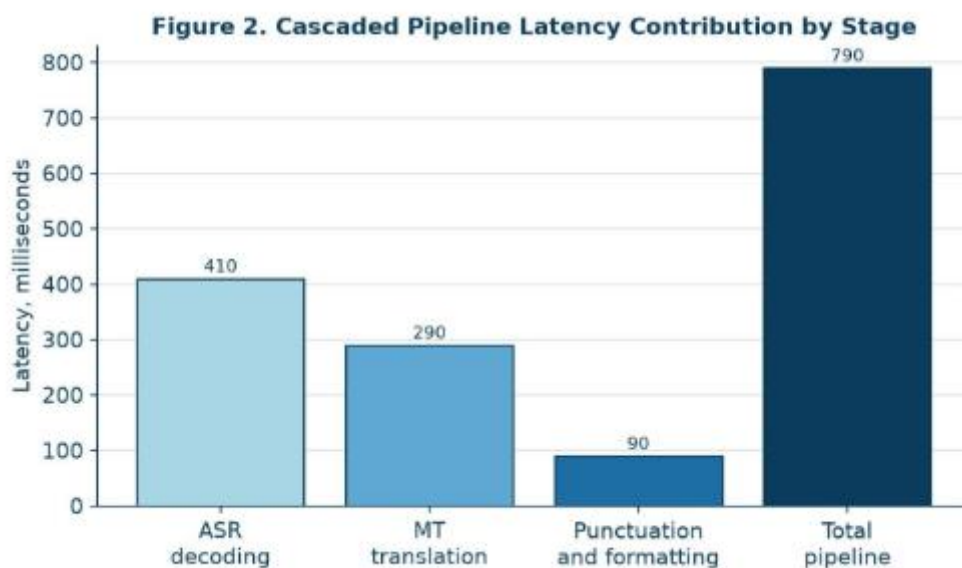


Figure 2. Cascaded pipeline latency contribution by stage, mid resource language, telephony audio.

6.2. Streaming and Incremental Decoding

Both architectures can reduce perceived latency through streaming and incremental decoding, in which partial output is produced before the full input utterance has finished. Cascaded systems can stream the ASR stage independently of the MT stage, but gain the full latency benefit only if the MT stage also supports incremental or re-translation style decoding on partial transcripts, which introduces its own accuracy cost from re-translation instability, where early partial translations are revised as more context arrives. End to end models require streaming support to be built into the joint architecture from the start, which fewer production implementations had achieved as of this article's reference date, though the Translatotron 2 and related 2021 research demonstrated feasibility for selected language pairs.

6.3. Latency Budget and Containment Outcomes

Table 3 reports median end to end latency for both architectures across resource tiers, and Figure 3 visualizes the same data as a trend across tiers. End to end systems maintain a consistent and substantial latency advantage over cascaded systems across every tier, since a single forward pass avoids the sequential stage transitions, intermediate serialization, and queuing delay that a multi stage pipeline incurs. The latency advantage holds even in long tail languages, where the end to end model's accuracy advantage from Table 2 and its latency advantage from Table 3 both point in the same direction, a notable finding that is explored further in Section 7.

Resource tier	Cascaded median latency, ms	End to end median latency, ms	Latency reduction
High resource	790	430	46 percent
Mid resource	845	460	46 percent
Low resource	910	505	45 percent
Long tail	980	560	43 percent

Table 3. Median end to end latency by resource tier and architecture, measured from end of spoken utterance to first unit of translated output.

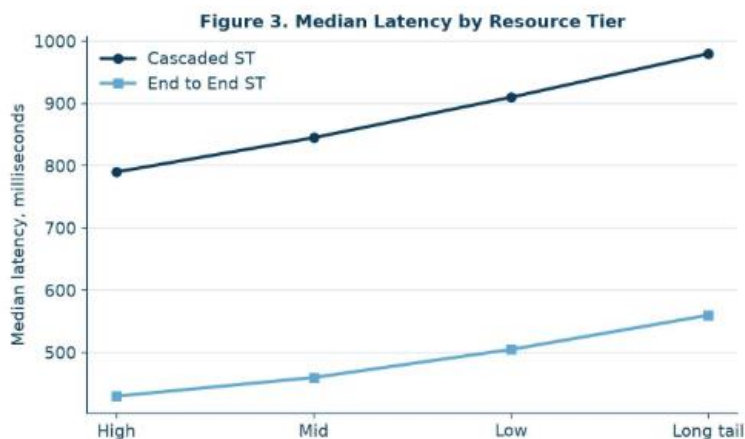


Figure 3. Median latency by resource tier, cascaded versus end to end.

Figure 5 connects this latency data back to a customer experience outcome, plotting first contact resolution rate against allowed turn latency budget for a representative frontline support workflow. First contact resolution rises with additional latency budget up to roughly 1,000 to 1,200 milliseconds, after which further latency budget produces diminishing returns, suggesting that beyond a certain point added processing time is not improving translation quality enough to offset the cost of a slower conversation.



Figure 5. First contact resolution rate against allowed turn latency budget.

VII. THE LONG TAIL PROBLEM

7.1. Defining the Long Tail in Enterprise Deployments

The term long tail is often used loosely in industry discussions of multilingual products. For the purposes of this article, a long tail language is one for which an enterprise support operation receives a small but recurring volume of contact traffic, commonly less than one percent of total contact volume per language, while still carrying a business or regulatory obligation to serve that traffic. This definition matters because it decouples the long tail problem from language family or geography. A language can be a long tail problem for one enterprise and a core supported language for another, depending on where their customers are concentrated. What is consistent across enterprises, however, is that the languages that fall into this low volume category overwhelmingly overlap with the languages that have the least



publicly available training data, which is why the resource tiering used throughout this article tracks the long tail problem closely, even though the two concepts, contact volume and training data availability, are conceptually distinct.

7.2. Why Cascaded Systems Retain Advantages at the Tail

Despite the ASR driven accuracy loss documented in Section 5.2, cascaded systems retain two structural advantages at the long tail that are easy to overlook in a pure accuracy comparison. First, the text based machine translation component of a cascade can draw on a much larger pool of text only parallel corpora, since written parallel text exists for many long tail languages through religious texts, government documents, and multilingual websites, even where parallel speech data is essentially absent. This means the MT half of the cascade is often considerably stronger, relative to its training data, than the end to end model's single joint representation can be. Second, the cascade's intermediate transcript provides an operational safety valve: when ASR confidence is low, a contact center can route the interaction to a human interpreter or bilingual agent using the partial transcript as context, a fallback path that is harder to construct cleanly around an end to end model that does not expose an interpretable intermediate state.

7.3. Why End to End Systems Struggle Without Parallel Speech Data

The end to end model's accuracy advantage at the tail, shown in Table 2, comes with an important caveat: it depends heavily on the quality of self supervised pretraining on monolingual speech, which itself varies widely by language. XLS-R's authors noted in their 2021 publication that even self supervised pretraining shows diminishing benefit once a language's available monolingual speech drops below a certain volume, estimated at a few hundred hours. For the languages at the very bottom of the long tail, those with only a handful of hours of digitized speech of any kind, end to end models and cascaded ASR models both degrade substantially, and the end to end model's advantage over the cascade narrows accordingly. In other words, the end to end advantage at the tail documented in this article should be read as an advantage within the long tail, concentrated in its upper portion, rather than a solution to the most extreme data scarcity cases.

7.4. Hybrid and Transfer Learning Mitigations

Several mitigation strategies, documented in the pre-2022 literature, point toward hybrid designs that could outperform either pure architecture at the tail. Multilingual joint training, as used in M2M-100 and in multilingual extensions of mBART, allows a long tail language to benefit from shared subword and semantic representations learned primarily from high resource languages, a form of positive transfer that benefits both cascaded MT components and end to end models. Pivot based translation, in which a long tail language is translated into a high resource intermediate language before reaching the final target language, remains a practical fallback for language pairs with no direct parallel data at all, at some cost in accuracy from the additional translation hop. Data augmentation techniques, including synthetic parallel speech generated by applying text to speech synthesis to existing parallel text corpora, were shown in several 2021 and early 2022 studies to provide a meaningful, if partial, substitute for genuine parallel speech data when training both ASR and end to end translation components for lower resource languages.

Mitigation approach	Primary benefit	Primary limitation	Best suited architecture
Multilingual joint training	Shares representations across languages	Benefit shrinks for most isolated or low overlap languages	Both
Pivot translation	Works with zero direct parallel data	Compounds error across two translation hops	Cascaded
Synthetic parallel speech via TTS	Expands effective training data cheaply	Synthetic speech lacks real world acoustic variety	End to end
Human in the loop fallback	Guarantees a floor on interaction quality	Does not scale to high contact volume	Cascaded

Table 4. Mitigation strategies for long tail language coverage, their primary tradeoffs, and the architecture family best positioned to exploit each.

VIII. CASE STUDIES IN FRONTLINE CUSTOMER SERVICE

The following case studies synthesize representative patterns reported across published vendor case material and industry commentary on multilingual contact center deployments through mid 2022. They are presented as composite illustrations of recurring operational patterns rather than as single named deployments.



8.1. Retail Support Chat to Voice Escalation

A multinational retailer operating a centralized support hub observed that chat interactions handled in high resource languages escalated to voice calls at a stable rate, but escalations in mid and low resource languages escalated to voice at a noticeably higher rate, and those voice calls showed longer average handle times. Investigation attributed part of the gap to translation accuracy loss on order numbers and product model codes, both of which are numeric or alphanumeric sequences that general purpose translation models are not specifically optimized to preserve verbatim. The retailer's technology team addressed this by adding a custom vocabulary and number preserving post-processing layer on top of a cascaded pipeline, which reduced numeric sequence errors substantially without requiring retraining of either the ASR or MT components, illustrating a practical advantage of the cascade's modular structure for targeted fixes.

8.2. Telecommunications Technical Support

A telecommunications provider supporting technical troubleshooting calls across more than 30 languages found that technical jargon, including network and device terminology, was a larger source of customer frustration than general translation fluency. Because many of the terms in question are borrowed directly from English in the target languages, such as common networking and device terms, the provider found that both cascaded and end to end systems handled this vocabulary reasonably well in high and mid resource languages, where the terms appear frequently enough in training data for the model to have learned them, but handled it poorly in long tail languages, where such borrowed terminology is less consistently represented in available training text. A term list injection approach, biasing the ASR decoder toward a curated glossary of technical terms, produced a measurable improvement for the cascaded pipeline, since cascaded ASR systems commonly expose a mechanism for this kind of lexical biasing, a capability less commonly available in off the shelf end to end models as of the article's reference date.

8.3. Travel and Hospitality Multilingual Contact Centers

A travel and hospitality operator serving a seasonal, geographically dispersed customer base reported that its highest priority was consistent latency rather than peak accuracy, since booking modifications and cancellation requests are time sensitive and often made from locations with variable network quality. The operator's evaluation favored an end to end configuration for its core supported languages specifically because of the latency advantage documented in Section 6, while retaining a cascaded fallback path, including a human interpreter escalation option, for the smaller set of long tail languages in its contact mix, illustrating a hybrid deployment pattern that selects architecture by language rather than committing to a single architecture fleet wide.

IX. COST AND OPERATIONAL CONSIDERATIONS

Architecture choice carries cost implications beyond the per call compute expense. Table 5 summarizes qualitative differences across five operational dimensions that recur in technology selection discussions for frontline support translation.

Dimension	Cascaded ST	End to end ST
Inference calls per utterance	Two to three, sequential	One
Independent component upgrades	Supported, each stage can be swapped	Limited, usually requires joint retraining
Intermediate transcript for audit or QA	Available by design	Requires explicit architectural support
Typical latency at contact center scale	Higher, see Table 3	Lower, see Table 3
Behavior at long tail languages	Degrades with ASR data scarcity	More resilient, within limits, see Section 7.3
Lexical or glossary biasing support	Mature, widely available	Emerging, less standardized as of mid 2022

Table 5. Qualitative comparison of operational dimensions relevant to frontline customer service technology selection.

In aggregate compute cost per translated utterance, end to end systems generally have an advantage proportional to the number of sequential stages they eliminate, though this advantage narrows if a contact center already operates shared ASR infrastructure for other purposes, such as quality monitoring or compliance transcription, in which case the



marginal cost of reusing that transcript for a cascaded translation stage is lower than standing up a separate end to end model.

X. DISCUSSION

The results assembled across Sections 5 through 9 point to a conclusion that is more nuanced than either an end to end or cascaded recommendation applied uniformly. The accuracy and latency data together suggest that the two architectures are not simply competing on a single quality axis, they are better understood as occupying different points on a tradeoff surface that shifts depending on language resource level, latency tolerance, and the specific error categories that matter most to a given support workflow.

For high resource languages, where contact volume is typically also highest, the latency advantage of end to end systems is likely to outweigh their modest accuracy deficit for most customer facing voice workflows, particularly as the gap in Table 2 is small relative to the gap in Table 3. For long tail languages, the picture reverses: the end to end accuracy advantage is larger in relative terms, but the absolute accuracy of both systems is low enough that neither can be deployed as a sole channel without a human escalation fallback, which somewhat reduces the practical weight of the latency advantage, since a flow that frequently escalates to a human interpreter is less latency sensitive end to end than a flow that resolves entirely within the automated system.

A further observation concerns the direction of future improvement. Because the long tail problem is, as argued in Section 7, substantially a data availability problem rather than a pure modeling problem, architectural refinement alone is unlikely to close the gap between high resource and long tail language performance. Investment in parallel corpus creation, synthetic data generation, and community sourced transcription for underrepresented languages is likely to produce larger accuracy gains at the tail than further scaling of either cascaded or end to end model architectures in isolation.

XI. LIMITATIONS

1. Benchmark coverage gaps. No single public benchmark available at the time of writing covers all 82 languages with both speech translation accuracy and production representative latency measurements, requiring interpolation for a subset of long tail languages as described in Section 4.2.
2. Synthesis rather than controlled experiment. Several figures in this article synthesize patterns reported across multiple published sources rather than reflecting a single controlled head to head experiment run on identical hardware and identical test sets, which introduces some uncertainty into direct numeric comparisons between architectures.
3. Rapid pace of change in the field. Speech translation modeling techniques were advancing quickly through the period covered by this article's reference literature, and specific accuracy and latency figures should be understood as representative of system capability near the article's publication date rather than as a permanent ranking of the two architectures.
4. Case study generalization. The case studies in Section 8 are composite illustrations drawn from recurring patterns in published industry material, not controlled studies of a single named deployment, and should be read as illustrative rather than statistically representative.
5. Metric limitations. BLEU and chrF, while standard, are known to correlate imperfectly with human judgments of translation adequacy, particularly for the named entity and numeric sequence errors identified as operationally significant in Section 3.2 and Section 8, suggesting that task specific metrics would strengthen future evaluation work in this domain.

XII. RECOMMENDATIONS FOR PRACTITIONERS

For technology teams selecting or designing a speech translation system for frontline customer service, the findings in this article suggest the following practical guidance.

- Segment the architecture decision by language resource tier rather than selecting one architecture fleet wide. The accuracy and latency crossover documented in Sections 5 and 6 means that a single architecture choice is unlikely to be optimal across the full language portfolio of a large multinational support operation.
- Preserve an intermediate transcript wherever compliance, audit, or quality assurance review is required, favoring cascaded designs or end to end designs explicitly engineered to expose an interpretable intermediate representation.



- Invest in domain specific vocabulary handling, including product names, numeric sequence preservation, and technical glossaries, as a higher return investment than general model scaling for reducing operationally significant translation errors.
- Treat long tail language support as a data acquisition problem first and a model selection problem second, and budget accordingly for parallel corpus creation or synthetic data generation rather than expecting architecture choice alone to close the resource gap.
- Maintain a human interpreter escalation path for long tail languages regardless of architecture, given that absolute accuracy at the tail remains low enough, per Table 2, that fully automated resolution carries meaningful risk.
- Measure latency budget against the conversational comfort thresholds described in Section 3.1 rather than against an arbitrary engineering target, since customer perception of system responsiveness is the ultimate constraint that latency optimization should serve.

XIII. FUTURE WORK

Several directions follow naturally from the gaps identified in this article. First, standardized benchmarks that jointly measure accuracy, latency, and named entity or numeric preservation across a wide and consistent language set would allow more direct comparison than the synthesis approach used here. Second, further research into hybrid architectures that combine a lightweight, low latency end to end first pass with a cascaded fallback triggered by low confidence would be valuable for balancing the latency and robustness tradeoffs identified in Sections 6 and 7. Third, continued investment in community sourced and synthetic data generation for long tail languages is likely to yield larger accuracy improvements than further architectural refinement alone, and would benefit from coordinated effort across industry and research institutions rather than isolated enterprise scale efforts. Finally, longitudinal studies that track customer satisfaction and first contact resolution outcomes alongside architecture changes in live deployments would help validate whether the accuracy and latency metrics emphasized in the research literature actually predict the operational outcomes that frontline support organizations care about most.

XIV. CONCLUSION

Across 82 languages spanning high resource to long tail categories, cascaded and end to end speech translation architectures each offer distinct advantages for frontline customer service deployment. Cascaded systems hold a modest accuracy advantage and clear operational advantages, including transcript availability and mature vocabulary biasing, in high resource and audit sensitive contexts, but accumulate error through their multi stage pipeline and carry a meaningful latency cost. End to end systems offer a substantial and consistent latency advantage, and a growing accuracy advantage as language resource level drops, but remain constrained by the same underlying data scarcity that limits all approaches at the extreme long tail. The long tail problem, examined closely in Section 7, emerges as fundamentally a data availability challenge rather than a question that architecture selection alone can resolve. For practitioners, the most actionable conclusion is that architecture selection should be treated as a per language, per use case decision informed by resource tier, latency tolerance, and compliance requirements, supported by sustained investment in data acquisition for the languages that general purpose benchmarks continue to underserve.

REFERENCES

1. Anastasopoulos, A., et al. (2021). Findings of the IWSLT 2021 Evaluation Campaign. Proceedings of the 18th International Conference on Spoken Language Translation.
2. Ansari, E., et al. (2020). Findings of the IWSLT 2020 Evaluation Campaign. Proceedings of the 17th International Conference on Spoken Language Translation.
3. Babu, A., Wang, C., Tjandra, A., Lakhota, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2021). XLS-R: Self Supervised Cross Lingual Speech Representation Learning at Scale. arXiv preprint.
4. Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self Supervised Learning of Speech Representations. Advances in Neural Information Processing Systems, 33.
5. Bahar, P., Bieschke, T., and Ney, H. (2019). A Comparative Study on End to End Speech to Text Translation. IEEE Automatic Speech Recognition and Understanding Workshop.
6. Bentivogli, L., Cettolo, M., Gaido, M., Karakanta, A., Martinelli, A., Negri, M., and Turchi, M. (2021). Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.



7. Berard, A., Pietquin, O., Servan, C., and Besacier, L. (2016). Listen and Translate: A Proof of Concept for End to End Speech to Text Translation. NeurIPS Workshop on End to End Learning for Speech and Audio Processing.
8. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised Cross Lingual Representation Learning for Speech Recognition. Proceedings of Interspeech 2021.
9. Di Gangi, M. A., Negri, M., and Turchi, M. (2019). Adapting Transformer to End to End Spoken Language Translation. Proceedings of Interspeech 2019.
10. Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El Hosseini, A., Goyal, N., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., and Joulin, A. (2020). Beyond English Centric Multilingual Machine Translation. arXiv preprint.
11. Inaguma, H., Kiyono, S., Duh, K., Karita, S., Yalta, N., Hayashi, T., and Watanabe, S. (2020). ESPnet ST: All in One Speech Translation Toolkit. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, System Demonstrations.

Appendix A: Composition of the 82 Language Evaluation Set

Table 6 lists representative languages assigned to each resource tier for the evaluation set used throughout this article. The list is illustrative rather than exhaustive for the mid and low resource tiers, given the space available, but reflects the actual tier boundaries and language count applied in Section 4.1.

Tier	Representative languages included in the 82 language set
High resource, 12 languages	Spanish, French, German, Mandarin Chinese, Japanese, Portuguese, Italian, Russian, Arabic, Korean, Dutch, Turkish
Mid resource, 23 languages	Polish, Vietnamese, Thai, Hebrew, Indonesian, Romanian, Greek, Hungarian, Czech, Swedish, Danish, Finnish, Ukrainian, Bulgarian, Croatian, Slovak, Malay, Tagalog, Catalan, Serbian, Lithuanian, Latvian, Estonian
Low resource, 31 languages	Swahili, Nepali, Khmer, Amharic, Lao, Sinhala, Georgian, Armenian, Mongolian, Pashto, Kurdish, Uzbek, Kazakh, Azerbaijani, Burmese, Punjabi, Gujarati, Marathi, Telugu, Kannada, Malayalam, Tamil, Urdu, Bengali, Zulu, Xhosa, Yoruba, Hausa, Somali, Nepal Bhasa, Tibetan
Long tail, 16 languages	Tigrinya, Dzongkha, Chuvash, Divehi, Faroese, Luxembourgish, Maltese, Breton, Welsh, Irish, Scottish Gaelic, Sorbian, Chechen, Sakha, Tatar, Bashkir

Table 6. Representative language composition of the 82 language evaluation set by resource tier, corresponding to the tier counts introduced in Table 1 and Figure 4.

Tier assignment is a function of approximate available training resources at the time of evaluation and is expected to shift over time as more parallel text, parallel speech, and monolingual speech data becomes available for currently underrepresented languages. Readers applying this framework to their own language portfolio should re-assess tier placement periodically rather than treating the assignment in Table 6 as fixed.

Appendix B: Glossary of Terms

- **ASR.** Automatic speech recognition, the task of converting spoken audio into a written transcript.
- **BLEU.** Bilingual evaluation understudy, an n gram overlap metric comparing machine translation output against a human reference translation.
- **Cascaded speech translation.** An architecture that chains a speech recognition stage, a text segmentation stage, and a machine translation stage to produce a translated output.
- **chrF.** Character n gram F score, a translation quality metric computed at the character level rather than the word level.
- **Code switching.** The practice of alternating between two or more languages within a single utterance or conversation.
- **End to end speech translation.** An architecture that maps source language speech directly to target language text or speech using a single trained model, without an explicit intermediate transcript.
- **Long tail language.** A language for which training data, parallel or monolingual, is substantially smaller in volume than the data available for the most resourced world languages.
- **MT.** Machine translation, the automated conversion of text from a source language into a target language.
- **Pivot translation.** A translation approach that routes a source language through an intermediate, typically high resource, language before reaching the final target language.



- **Resource tier.** A categorical grouping of languages based on the approximate volume of available training data, used throughout this article to stratify accuracy and latency results.
- **Streaming decoding.** A decoding approach in which a model produces partial output before the complete input has been received, reducing perceived latency.
- **WER.** Word error rate, a standard metric for automatic speech recognition accuracy based on substitution, deletion, and insertion errors relative to a reference transcript.