DeepL Voice: Real-Time Speech-to-Speech Translation

Johannes Ernesti[1] , **Peter Kaiser**[1] , **Jonas Heinze**[1] , **Elnaz Shafaei-Bajestan**[1] , **Kristina Geißler**[1] , **Weiyue Wang**[1] , **Johannes Beck**[1] , **Sascha Brinker**[1] , **Thorben Finke**[1]

[1]DeepL Research

{johannes, peter, jonas.heinze, elnaz.shafaei, kristina.geissler, weiyue.wang, johannes.beck, sascha.brinker, thorben.finke}@deepl.com

Abstract

DeepL Voice is a real-time speech-to-speech translation system for global business communication, following a pragmatic incremental approach: developing a production-grade cascaded speech-to-speech-translation (S2ST) system, while exploring end-to-end solutions in parallel. The production system (launched November 2024) achieves competitive transcription quality through proprietary real-time ASR models and eliminates translation "flickering" via stable text streaming while maintaining low latency. Supporting 18 input languages and 30+ target languages, it offers *DeepL Voice for Meetings* (Microsoft Teams/Zoom integration), *DeepL Voice for Conversations* (mobile apps), as well as the *DeepL API for Voice*. Key features include customizable formality and glossary support for business-appropriate communication, with voice cloning TTS under development.

1 Introduction

Real-time speech-to-speech translation (RT-S2ST) has evolved from futuristic vision to practical reality. Growing customer demand and commercial interest have driven rapid development. Commercial systems from Google, Microsoft, and DeepL demonstrate the importance of real-time communication for private and professional use-cases across language barriers.

DeepL Voice employs a cascaded architecture (ASR → MT → TTS), prioritizing production quality and reliability. While end-to-end approaches [Jia *et al.*, 2019; Jia *et al.*, 2022; Meta AI, 2023; Labiausse *et al.*, 2025] offer potential advantages in latency and prosody preservation, cascaded systems provide modularity, interpretability, and immediate production readiness by leveraging proven ASR (e.g. [Radford *et al.*, 2023]) and MT components. We explore both cascaded optimization and end-to-end research in parallel to deliver continuous customer value while advancing state-of-the-art capabilities.

2 Challenges of Real-Time S2ST

Developing DeepL Voice for production deployment required addressing fundamental technical challenges in real-time speech-to-speech translation. This section outlines the key challenges we encountered.

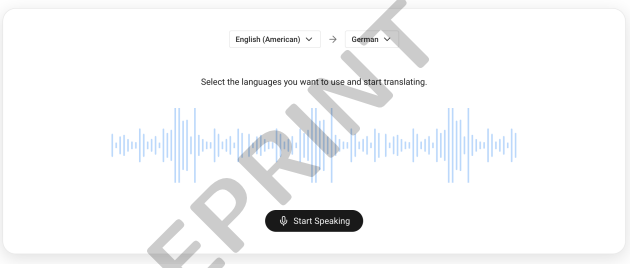

Figure 1: DeepL Voice web interface (https://www.deepl.com/products/voice) showing language selection (English to German) and Start Speaking button.

Quality-Latency Trade-offs. The ambition of real-time S2ST is to deliver high-quality translations with low latency which creates an inherent tension. Optimizing for end-to-end latency often requires trading off translation quality as translation decisions may be taken prematurely. At the same time, delivering high-quality (e.g. sentence-based) translations may imply unacceptable latency depending on the customer use-case. Cascaded systems accumulate latency across pipeline stages and may behave inefficiently due to local and heuristic decision-making between components. End-to-end systems on the other hand provide the chance to do smart and situation-dependent trade-offs between quality and latency.

Error Propagation and Robustness. Cascaded systems suffer from error propagation through ASR, MT, and TTS stages. E.g., ASR misrecognitions to acoustically similar but semantically divergent words lead to syntactically correct but semantically wrong translations. Additional challenges include noisy input, disfluencies (hesitations, false starts), and missing punctuation. Furthermore, post-hoc error correction is more difficult for TTS than for text displays, since new audio takes time to articulate whereas text can be swapped instantly.

Data Scarcity and Coverage. While text translation benefits from abundant corpora, spontaneous utterance translation faces severe data scarcity. Paired speech datasets remain limited, with expensive collection particularly acute for low-resource pairs and specialized domains (e.g. meetings,

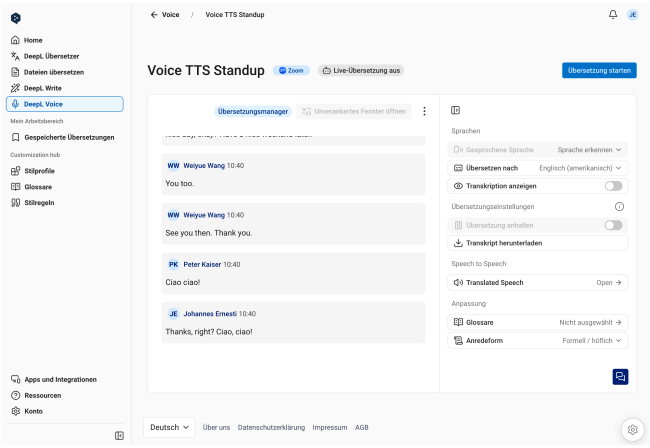

Figure 2: DeepL Voice for Meetings transcription interface showing real-time speech-to-text translation. Settings panel allows language pair selection and translation mode configuration.

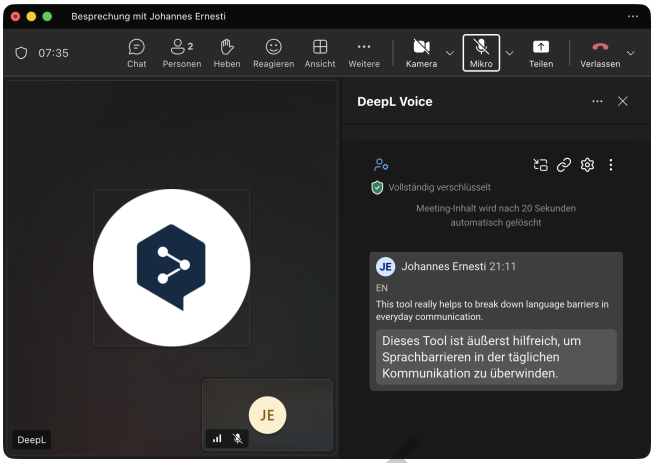

Figure 3: DeepL Voice seamlessly integrated into Microsoft Teams via an app displaying live translated captions during video calls

medical, legal). Architecturally, cascaded systems leverage separate datasets for ASR, MT, and TTS, while end-to-end approaches require costly paired speech data or sophisticated training techniques. Recent work explores reducing alignment requirements through reinforcement learning approaches [Labiausse *et al.*, 2026].

Prosody and Voice Preservation. Beyond semantic content, effective S2ST must convey paralinguistic information: emotion, emphasis, style, and voice. The text bottleneck of cascaded S2ST systems discards prosody, yielding generic voice synthesis results. End-to-end systems can preserve prosody, but they are more difficult to control as they offer only limited control over individual inference steps. Current solutions include voice cloning and style transfer, though natural, emotionally appropriate output remains a challenging problem in real-time applications.

Having established these fundamental challenges, we now describe how DeepL Voice addresses them through systematic engineering and thoughtful feature design.

3 DeepL Voice

DeepL Voice demonstrates a pragmatic incremental approach to production S2ST deployment. The production system (launched in November 2024) implements real-time speech-to-text translation combining state-of-the-art proprietary real-time ASR models with DeepL's industry-leading machine translation solutions [DeepL, 2024]. It supports 18 input languages and more than 30 target languages, leveraging DeepL's leading MT quality. DeepL is actively developing proprietary TTS models with prosody-aware synthesis, already available to selected customers through an early access program. These models will enable a full S2ST experience in upcoming iterations. In parallel, we experiment with end-to-end approaches to inform our research roadmap and deliver customer-facing improvements as fast as possible. Figures 2 and 3 illustrate DeepL Voice in production environments.

3.1 Use-Case

DeepL Voice addresses global business communication challenges across two primary variants. *DeepL Voice for Meetings* provides enterprise live captions seamlessly integrated into Microsoft Teams and Zoom for multilingual video conferencing. *DeepL Voice for Conversations* delivers mobile real-time bilateral translation for face-to-face interactions. In addition, the *DeepL API for Voice* enables custom solutions. Key application scenarios include international team collaboration, client meetings across language barriers, customer support in multiple languages, and multilingual conferences.

The system can be explored interactively on the web-based landing page at https://www.deepl.com/products/voice (Figure 1).

In the following, we discuss some particular challenges that we were facing when bringing DeepL Voice from the research prototype stage into production.

3.2 Stable Text Streaming and the Quality-Latency Tradeoff

Real-time speech translation systems face a fundamental challenge: producing high-quality translations with minimal latency from continuous audio input without clear sentence boundaries. Naive approaches that immediately translate each recognized word fragment create a "flickering" effect—translations constantly update and change as more context becomes available, creating jarring user experiences and making output unsuitable for text-to-speech synthesis.

DeepL Voice addresses this through stable text streaming, an approach to balance three competing objectives: translation quality, latency, and output stability. For DeepL Voice, we explore intelligent policies that determine optimal points to finalize translation decisions, ensuring that once text is displayed, it remains stable while maintaining acceptably low latency.

This approach eliminates flickering while preserving translation quality by leveraging DeepL's industry-leading neural machine translation. The stable text stream also provides a solid foundation for voice synthesis integration, as consistent

text enables natural-sounding prosody in text-to-speech output.

3.3 Language-specific Challenges

Different language pairs present varying degrees of difficulty for real-time translation due to fundamental differences in grammar and word order. Some translation directions, such as English to German, are relatively straightforward as both languages can follow similar subject-verb-object patterns, allowing incremental translation with minimal restructuring.

However, other language pairs pose significant challenges. German to English translation must handle verb-final word order, where critical verbs appear only at the end of subordinate clauses. This grammatical structure forces the system to either wait for complete clauses (increasing latency) or produce preliminary translations that may require substantial revision (reducing quality and stability).

Japanese presents even greater complexity, with fundamentally different word order (subject-object-verb) and grammatical particles that determine semantic roles. English to Japanese translation requires comprehensive sentence restructuring, making high-quality real-time translation particularly challenging. For such language pairs, the quality-latency tradeoff must be carefully reconsidered, potentially accepting higher latency to maintain translation accuracy and user trust.

3.4 Customization: Tone and Terminology

Professional and business communication requires not only accurate translation but also appropriate tone and consistent terminology. DeepL Voice integrates DeepL's proven customization features to address these requirements for real-time spoken communication.

Formality Control. The formality feature allows users to select between formal and informal registers in the target language, determining the pronouns, verb forms, and stylistic choices used in translation. This capability is essential for business contexts where cultural norms dictate appropriate levels of formality—for instance, using formal address in German business meetings ("Sie") versus informal conversation ("du"). The feature supports German, French, Italian, Spanish, Dutch, Polish, Portuguese, and Japanese.

Translation Glossaries. DeepL's glossary feature enables users to define custom translations for domain-specific terminology, brand names, product names, and internal jargon. Unlike simple find-and-replace approaches, glossaries intelligently adapt terms to grammatical context, automatically handling inflections, plural forms, and verb conjugations while maintaining semantic consistency. Organizations can create glossaries through three methods: uploading existing translation memory files, manual entry of term pairs, or customizing industry-specific sample glossaries. Shared glossaries ensure consistent terminology across teams and projects, critical for maintaining brand voice and technical accuracy in multilingual collaborations.

Spoken Terms for DeepL Voice. Complementing translation glossaries, we developed Spoken Terms[1] specifically for DeepL Voice to address the ASR component. This feature allows users to define internal terminology, acronyms, product names, and technical vocabulary that the speech recognition system should accurately transcribe. By ensuring correct transcription of company-specific terms before translation, Spoken Terms prevents cascading errors and improves end-to-end quality for specialized business conversations.

Together, these customization features enable DeepL Voice to deliver not just accurate translations, but appropriate, consistent, and context-aware communication that respects cultural norms and organizational terminology standards.

[1] Available to selected customers via an early access program.

3.5 Stability and Technical Complexity

A tool that provides reliable real-time translations at a quality-level that enables multilingual meetings without interpreters imposes very high standards on reliability. The latency requirements impose specific challenges in the deployment setup. For achieving easy accessibility during meetings, the product needs to be integrated into established meeting platforms like Microsoft Teams or Zoom, as well as be available as apps on common mobile operating systems.

4 Conclusion and Future Directions

We presented DeepL Voice, a pragmatic, quality-first approach to production S2ST deployment, balancing innovation with delivery through incremental development. Key contributions include stable text streaming to eliminate translation "flickering," customization features for business communication (formality control, glossaries, spoken terms), and seamless enterprise platform integration.

Acknowledgements

The content presented in this paper is based on the tireless work of the entire DeepL Voice team. We especially thank Faried Abu Zaid, Deniz Adalar, Pablo Aubele, Eileen Bailey, Raíssa Bastos-Schehl, Amiran Chyb, Nicolò Dalla Rosa, Leonardo Doin, Christoph Ebert, Anat Elhalal, Lennart Erpenbeck, Laura Escolar, Leonardo Ferreira, Markus Grasselt, Tassilo Holtzwart, Hadi Inja, Edward John, Fabian Joswig, Joachim Klein, Jakub Kowalisko, Christian König, Hoye Lam, Luca Lovisa, Mateo Luzi, Aliasgar Makda, Yoshia Makino, Maurizio Martino, Jonas Mehlhaus, Christian Meißner, Lisa Piotrowski, Shiv Prakash, Santhos Ramalingam, Pauline Ruiz Seiler, Patrick Schlarmann, Ivo Slegers, Martin Ueding, Jannis Veerkamp, and Lianna Wehinger.

We thank DeepL for providing such a productive research environment that made this work possible.

References

[DeepL, 2024] DeepL. DeepL Voice: Real-time speech translation. https://www.deepl.com/en/ai-labs/voice-to-voice, 2024. Launched November 2024.

[Jia *et al.*, 2019] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. Direct speech-to-speech translation with a sequence-to-sequence model. *arXiv preprint arXiv:1904.06037*, 2019.

[Jia *et al.*, 2022] Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *Proceedings of ICML*, 2022.

[Labiausse *et al.*, 2025] Tom Labiausse, Laurent Mazaré, Edouard Grave, Patrick Pérez, Alexandre Défossez, and Neil Zeghidour. High-fidelity simultaneous speech-to-speech translation. *arXiv preprint arXiv:2502.03382*, 2025.

[Labiausse *et al.*, 2026] Tom Labiausse, Romain Fabre, Yannick Estève, Alexandre Défossez, and Neil Zeghidour. Simultaneous speech-to-speech translation without aligned data. *arXiv preprint arXiv:2602.11072*, 2026.

[Meta AI, 2023] Meta AI. SeamlessM4T: Massively multilingual and multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, 2023.