1 — Конвейер из четырёх шагов

```
Услышать -> ASR (речь в текст) -> MT (перевод) -> TTS (текст в речь) -> доставка.
Один конвейер — и для наушников, и для видеозвонка.
```

2 — «Наушники с переводом» = аудио-точка перед конвейером

```
На устройстве (Apple Live Translation, iOS 26): приватно, офлайн, лимит железа.
Телефон + облако (Pixel / Galaxy): уровень Gemini, 70+ языков, нужна сеть.
```

3 — Каскад vs прямая модель (как собрать мозг)

```
Каскад (ASR->MT->TTS): текст -> субтитры, глоссарий, отладка, аудит. Больше задержка.
Прямая S2ST (Seamless / GPT Realtime / Gemini Live): меньше задержка, голос сохранён.
Регулируемое + субтитры -> каскад. Критична задержка + голос -> прямая. Часто оба.
```

4 — Почему переписывает себя + бюджет задержки

```
Языки меняют порядок слов -> ранний перевод догадка. ЗАМЕНЯЙТЕ строку до финала.
Держитесь ~2-3 с позади (ear-voice span / wait-k). Речь ждёт дольше субтитров.
Бюджет: ~1 с машина (TTS макс) + ~2-3 с ожидание. <800 мс живо; потолок AIIC 3-5 с.
```

5 — В WebRTC-звонке: перевести один раз на SFU, раздать по языкам

```
Снять звук на SFU, VAD-гейт, перевести один раз на язык слушателя.
Субтитры -> data channel (RFC 8831). Звук -> Opus-дорожка (RFC 6716), оригинал тише.
Стоимость ~ языки x минуты: 3 яз x 60 мин x $0,12 = $21,60/час. Остальное гейтит VAD.
```

6 — Чек-лист «строить или покупать»

- [ ] Групповой звонок с записью / аудитом? -> каскад на SFU (текст для субтитров + запись).
- [ ] Двое, нужен живой голос? -> прямая модель S2ST, выразительный вариант.
- [ ] Self-host звонка? -> mediasoup / Janus / LiveKit (свой съём + вставка дорожки).
- [ ] Выберите ASR + MT + TTS (каскад) или одну S2ST; меняйте по задержке + точности.
- [ ] <60 одновременных потоков -> облачный API; сильно выше -> self-host (решает комплаенс).
- [ ] Всегда: VAD-гейт, глоссарий имён, план согласия на клон голоса + запись.