SPEECH-TO-SPEECH ARTIFICIAL INTELLIGENCE AS A NEW STAGE IN AUTOMATED COMMUNICATION

Radoslav Dodnikov[1]
e-mail: radoslav.dodnikov@unwe.bg[1]

Abstract

The fast development of AI technologies has led to the emergence of different speech models. One of th ese models are the speech-to-speech models (S2S) that represent a new stage in automated communic ation. This paper analyzes its technological foundations, such as encoder-decoder architectures, multi lingual acoustic modeling, real-time speech synthesis and prosody transfer. It outlines key application s in education, customer service, healthcare, fintech, and digital accessibility. Big importance is given to the advantages of these models, like better naturalness, reduced latency, and preservation of emotio n and speaker characteristics. The paper also discusses current challenges related to model accuracy, language adaptation, data privacy, and ethical use. The analysis concludes that speech-to-speech AI r epresents a transformative step towards seamless human-machine interaction and opens new opportu nities for personalized and context-aware communication technologies.

Ключови думи: speech-to-speech models, ai communication, encoder-decoder architecture, real-time speech synthesis, multilingual acoustic modeling

JEL: C88, O33, L86

Understanding Speech-to-Speech AI

We have all seen sci-fi movies where characters meet other characters that speak a completely different language and their voice is instantly translated into the other language while keeping their emotion, style and personality. This is what we get with speech-to-speech AI models, an out-of-the-future technology that instantly transforms voice without first converting it to text.

Traditional approaches work with three points in their pipeline. First, we get the voice input translated into text (Speech-to-Text or STT). Second, it is processed, for example translated. Finally, it is converted back to speech (Text-to-Speech or TTS). However, each of these processes adds an additional delay and loses important information along the way.

S2S AI, by contrast, works like a direct pipeline. Your voice goes in one end, and transformed voice comes out the other, preserving all the subtle nuances that make human speech rich and expressive: the pauses, the emotional inflections, the rhythm, and even your unique vocal characteristics.

Why is this important

Moving from the traditional speech processing pipelines to S2S is really a big change in the way machines process and generate human speech. This technology enables more natural conversations with AI assistants and agents that sound human, real-time translation that also preserves your voice and emotion, makes communication accessible for people with speech impairments, and makes possible the personalization of voice assistants that can understand and respond in appropriate emotional context.

We have all seen how in recent years and even months, voice AI assistants and customer service bots increase their communication quality and emotional correctness. This is due to the S2S technologies that make our communication with computers more human-like. As we interact more and more with AI in our daily lives, the quality of these interactions becomes crucial. S2S AI represents the next step in

[1] Докторант към катедра ИТК, УНСС, email: radoslav.dodnikov@unwe.bg

making these interactions feel less like talking to a machine and more like conversing with another person.

How S2S AI Works

Understanding S2S AI means looking at four main technologies that work together to make direct voice transformation possible. Think of these as four essential parts of a machine that all need to work together perfectly.

Encoder-Decoder Architectures

At the heart of S2S AI is something called the encoder-decoder architecture. This is a type of neural network that works similar to how our brain processes and reproduces speech.

Imagine you are listening to a song and trying to hum it back. Your brain does not memorize every single sound wave. Instead, it captures the essence of the melody, rhythm, and emotion. This is the encoding process. Then, when you hum it back, your brain reconstructs the song from this compressed memory. This is the decoding process.

S2S AI works in a similar way. The encoder listens to the input speech and compresses it into what we call a context vector. This vector captures everything important about the speech: what is being said, how it sounds, and the emotional tone. Then the decoder takes this context vector and generates the output speech step by step, reconstructing the transformed voice while keeping all the characteristics that were captured.

One good example of this is Microsoft's SpeechT5 framework. It uses this encoder-decoder architecture and can handle multiple tasks like speech recognition and voice conversion using the same core system. Modern S2S systems use something called Transformer architectures with attention mechanisms. These allow the system to focus on the most relevant parts of the input when generating each part of the output, similar to how you might pay special attention to certain words when trying to understand someone's emotional state.

Multilingual Acoustic Modeling

Multilingual acoustic modeling is what allows S2S AI to work across different languages. It trains models on data from many languages at the same time, discovering patterns that are common across all languages while also preserving what makes each language unique.

Different languages have very different sounds and rhythms. Mandarin Chinese uses tones where the pitch changes the meaning of words. English relies more on stress patterns. Arabic has sounds that do not exist in European languages. Traditional systems would need separate models for each language, which is inefficient and hard to maintain.

Multilingual modeling solves this by creating shared representations that capture universal speech patterns. The model learns that certain features are common across languages, like the sound of laughter or emotional expressions, while others are specific to each language, like the rolled R in Spanish or the clicks in some African languages.

This approach has three main benefits. First, it enables cross-lingual speech synthesis, where the system can generate natural speech in languages it has not been extensively trained on by using knowledge from other languages. Second, it allows real-time translation where speech is translated from one language to another while preserving the speaker's voice and emotional tone. Third, it supports low-resource languages by transferring knowledge from languages with lots of training data to languages with limited data.

Imagine attending an international conference where everyone hears the speaker in their native language, but with the original speaker's voice and emotional delivery intact. This is what multilingual S2S AI makes possible.

Real-time Speech Synthesis

Real-time speech synthesis means generating natural-sounding speech fast enough for normal conversation. In human conversation, we expect responses within about 200 to 300 milliseconds. Delays longer than this make conversations feel unnatural and frustrating.

Traditional speech processing pipelines often introduce delays of 500 to 1000 milliseconds or more because the speech has to pass through multiple stages. First it converts speech to text, then processes the text, then converts it back to speech. Each step adds delay.

S2S AI achieves low latency by processing everything in one go. Instead of waiting to process the entire input before generating output, the system can begin producing speech as soon as it has enough information. This is like a simultaneous interpreter who starts translating while the speaker is still talking, rather than waiting for complete sentences.

Modern S2S systems also use optimized neural network designs that balance quality with speed. They process speech in small chunks, typically 10 to 50 milliseconds, which allows them to maintain low latency even for long utterances. This makes S2S AI suitable for live applications like simultaneous translation, interactive voice assistants, and real-time dubbing of live broadcasts.

Prosody Transfer

Prosody refers to the rhythm, stress, and intonation patterns of speech. It is essentially the music of language. Prosody conveys emotion, emphasis, and meaning beyond the words themselves. Prosody transfer is the ability to preserve and transfer these characteristics in synthesized speech.

Consider the sentence "I did not say he stole the money." Depending on which word you emphasize, the meaning changes completely. If you emphasize "I," it means someone else said it. If you emphasize "say," it means you implied it. If you emphasize "he," it means someone else stole it. Prosody also conveys emotion. The same words can sound happy, sad, angry, or sarcastic depending on how they are spoken.

Traditional text-based systems lose all this information because text does not capture pitch, rhythm, or emotional tone. S2S systems maintain prosodic information throughout the transformation process. They encode not just what is being said, but how it is being said: the pitch contours, speaking rate, voice quality, and emotional characteristics.

Advanced S2S models can even separate the linguistic content from the prosodic style. This allows the system to translate words to a different language while preserving the original speaker's emotional delivery. The system can also extract and preserve speaker-specific characteristics like voice timbre, speaking habits, and accent, creating a unique fingerprint for each speaker's voice.

This technology has important applications. It enables emotional speech synthesis for AI assistants that can express appropriate emotions. It allows voice cloning for people who have lost their ability to speak. It creates expressive audiobook narrators that sound engaging and natural with appropriate emotional variation.

Where S2S AI is Used

S2S AI is not just a theoretical technology. It has practical applications that are already improving people's lives in many different areas.

Education

In education, S2S AI is transforming how people learn languages and interact with educational content. Imagine a language learning app that does not just tell you if your pronunciation is correct, but actually demonstrates the correct pronunciation in your own voice, showing you exactly how to adjust. This is now possible with S2S AI.

Educational AI assistants can engage in natural conversations with students and adjust their speaking style based on the student's emotional state. They can speak more encouragingly when a student is frustrated, or more enthusiastically when they are making progress. Students with speech impairments can also participate in oral presentations using S2S systems that generate natural-sounding speech while preserving their intended emotional expression.

A student learning Mandarin Chinese can practice tonal pronunciation with an AI tutor that provides immediate, natural feedback in conversational form. The system can demonstrate correct tones using the student's own voice, making it easier to understand and remember the corrections.

Customer Service

Customer service is another area where S2S AI is making a big difference. Customer service bots powered by S2S AI can detect frustration in a caller's voice and respond with appropriate empathy. They can adjust their speaking style to match the customer's emotional state, being calm and reassuring for anxious customers, or efficient and direct for those who prefer quick resolution.

A single AI agent can also seamlessly switch between languages while maintaining consistent personality and service quality. A customer in Spain can speak Spanish, while a customer in Japan speaks Japanese, both interacting with what feels like the same helpful agent. Companies report that customers rate interactions with S2S-powered agents as significantly more satisfying than traditional text-to-speech bots.

Healthcare

In healthcare, S2S AI is improving both efficiency and accessibility. Doctors can dictate notes, order tests, and access patient information through natural conversation while maintaining sterile conditions or keeping their hands free during procedures. The system understands medical terminology and context, reducing documentation time.

Patients with speech disorders can communicate with healthcare providers using S2S systems that generate clear, natural speech while preserving their intended meaning and emotional context. This is particularly valuable for patients with conditions like ALS, stroke, or vocal cord damage. AI systems can also conduct check-in calls with patients, asking about symptoms and medication compliance in a natural, conversational manner.

A patient with Parkinson's disease, which affects speech clarity, can use an S2S system to communicate with their doctor. The system clarifies their speech while preserving their emotional expression, allowing the doctor to understand not just what the patient is saying, but how they are feeling about their condition.

Fintech

In the financial services industry, S2S AI enables secure voice-based authentication that is difficult to spoof. The system can verify identity through natural conversation, analyzing not just voice characteristics but also speaking patterns and prosody. AI financial advisors can provide guidance through natural conversation, adjusting their communication style based on the customer's financial literacy and emotional state.

S2S systems can also detect anomalies in voice patterns that might indicate stress or deception, helping identify potential fraud attempts during phone banking or transaction authorization. Visually impaired customers can manage their finances through natural voice interactions without needing to navigate visual interfaces.

A customer calling their bank to authorize a large transaction can be verified through voice biometrics during natural conversation. The system detects that they sound confident and calm, suggesting the transaction is legitimate, and processes the authorization without asking security questions or requiring PIN codes.

Digital Accessibility

Perhaps one of the most important applications of S2S AI is in digital accessibility. People with speech impairments can use S2S systems to generate natural-sounding speech that reflects their personality and emotional state. Unlike traditional text-to-speech systems that sound robotic, S2S can create personalized voices that feel authentic.

S2S AI can provide live audio descriptions of visual content for blind users, or generate captions with prosodic information indicating tone and emotion for deaf users. People with cognitive disabilities can interact with technology through simplified, natural voice interfaces that adapt to their communication style and pace.

A person with ALS who has lost the ability to speak can use an S2S system trained on recordings of their voice from before the disease progressed. They can communicate with family members using their own voice, maintaining their identity and emotional connection. This is transformative for people who are losing or have lost their ability to speak.

Global Communication

S2S AI is also breaking down language barriers in global communication. International conferences and meetings can use S2S AI for real-time interpretation that preserves speakers' voices and emotional delivery across languages. Business professionals can negotiate and collaborate across language barriers while maintaining the nuances of tone and emotion that are crucial for building trust.

Media content like podcasts, videos, and audiobooks can be translated to multiple languages while preserving the original speaker's voice and emotional delivery, creating more engaging localized content. Imagine a CEO giving a keynote speech at a global company meeting. Employees around the world hear the speech in their native language, but with the CEO's actual voice and emotional delivery intact. The passion and conviction in the CEO's voice comes through regardless of the listener's language.

Why S2S AI is Better

S2S AI offers several important advantages over traditional speech processing approaches. Understanding these benefits helps explain why this technology represents a real advancement.

Better Naturalness

When speech is converted to text and back to speech in traditional systems, crucial information is lost. Text does not capture the subtle variations in pitch, the natural pauses, the breathing patterns, or the small expressions in voice that make human speech rich and expressive. The result is speech that sounds robotic and unnatural.

S2S AI preserves all these acoustic nuances by processing speech directly without text conversion. It maintains breathing patterns, micro-pauses that indicate thinking or emphasis, pitch variations that convey emotion, voice quality changes that express feeling, and the way sounds blend together naturally in connected speech.

Studies show that listeners consistently rate S2S-generated speech as more natural and human-like compared to traditional TTS output. In some blind tests, S2S speech is indistinguishable from human speech, while traditional TTS is almost always identifiable as synthetic.

Compare these two scenarios. With traditional TTS, "I am so happy to help you today" is spoken in a monotone, robotic voice with even spacing between words. With S2S AI, the same sentence has natural enthusiasm, slight emphasis on "happy," and a warm, genuine tone. The words are identical, but the S2S version conveys genuine emotion and creates a more positive interaction.

Lower Latency

Traditional pipelines introduce significant delays because speech must pass through multiple stages. Speech to text takes 200 to 500 milliseconds, text processing takes 50 to 200 milliseconds, and text to speech takes another 200 to 500 milliseconds. The total delay is often 450 to 1200 milliseconds or more. These delays disrupt natural conversation flow and create frustrating user experiences.

S2S AI achieves much lower latency by eliminating intermediate stages. Direct transformation reduces total latency to 100 to 300 milliseconds, which is within the range of natural human conversation. This is achieved through single-pass processing where one model handles the entire transformation, streaming architecture where output generation begins before input is complete, and optimized neural networks that balance quality and speed.

Low latency enables applications that were previously impossible, like simultaneous interpretation for real-time translation during live conversations, interactive voice assistants with natural back-and-forth dialogue, and live dubbing for real-time translation of broadcasts. S2S AI brings machine conversation into the range of natural human interaction.

Keeping Emotion and Voice Identity

Traditional text-based systems face a fundamental problem. Text cannot encode emotional tone, speaker identity, or prosodic features. When speech is converted to text, all of this information disappears. When text is converted back to speech, the system must guess at appropriate emotion and prosody, often getting it wrong.

S2S systems maintain speaker characteristics throughout the transformation. Voice timbre, accent, speaking habits, and other unique characteristics that make each person's voice recognizable remain intact. The emotional state conveyed in the input speech is preserved in the output. Rhythm, intonation patterns, and emphasis are maintained. Individual speaking habits, like speaking quickly when excited or slowly when explaining something complex, are preserved.

This has important real-world applications. A person who is losing their voice due to illness can have their voice preserved and used to generate speech, maintaining their identity and emotional expressiveness. Actors' voices can be preserved when dubbing films into different languages, maintaining their emotional performance. Virtual assistants can maintain consistent personality and emotional appropriateness across interactions.

Imagine a motivational speaker giving a passionate speech. With S2S translation, the speaker's energy and passion come through in every language, their unique voice remains recognizable, emphasis on key points is preserved, and the emotional arc of the speech remains intact. With traditional translation, all of this would be lost, replaced by a generic voice reading translated text.

Challenges We Still Face

While S2S AI offers tremendous potential, it also faces significant challenges. Understanding these issues is important for the responsible development and use of the technology.

Technical Challenges

One major challenge is model accuracy. Human speech is incredibly diverse. People speak with different accents, dialects, speaking rates, voice qualities, and in various environments like quiet rooms, noisy streets, or over phone lines. Training S2S models that work well across all this diversity is extremely difficult.

A model trained primarily on American English might struggle with Scottish, Indian, or Australian English accents. This creates unfair performance where the technology works better for some groups than others. People also speak differently in different contexts, like formal presentations versus casual conversation. Models must handle all these variations. Background noise, poor microphone quality, and room acoustics all affect speech quality, and S2S systems must be robust to these real-world conditions.

Another challenge is language adaptation. While S2S AI can theoretically work across many languages, most training data exists for a small number of high-resource languages like English, Mandarin, and Spanish. The world's 7000 plus languages are vastly underrepresented. Many languages have little to no digital speech data available for training. Collecting such data is expensive and requires cooperation from native speaker communities.

Languages also differ dramatically in their sound systems, prosodic patterns, and grammatical structures. A model trained on English might struggle with tonal languages like Mandarin, click languages like Xhosa, or languages with complex consonant clusters like Georgian. Some languages lack standardized writing systems, making it difficult to create training datasets.

Data privacy is another serious concern. Training S2S models requires vast amounts of voice data. Voice is highly personal. It reveals not just what you say, but who you are, your emotional state, your health status, and potentially sensitive information about your identity and background. Voice recordings can be used to identify individuals and track their activities. Voice can reveal health conditions, emotional states, age, gender, ethnicity, and socioeconomic background.

Large databases of voice recordings are attractive targets for hackers and could be used for identity theft, fraud, or blackmail. Current protections like anonymization, encryption, and consent frameworks have limitations. As S2S AI becomes more common, we need clearer regulations specifically addressing voice data, technical innovations for privacy-preserving AI training, greater transparency about how voice data is used, and user-friendly controls for managing voice data.

Ethical Concerns

S2S AI can be used to create highly convincing fake audio, or deepfakes, that sound like someone saying things they never said. This technology can be weaponized for fraud, political manipulation, harassment, and misinformation.

Criminals can impersonate executives, family members, or authority figures to trick people into transferring money or revealing sensitive information. There have been documented cases of fraudsters using voice cloning to impersonate CEOs and authorize fraudulent wire transfers. Fake audio of politicians or public figures can be used to spread misinformation or manipulate elections. Individuals can be targeted with fake audio recordings that damage their reputation or relationships.

The existence of convincing fake audio also undermines the reliability of audio evidence in legal proceedings. "I did not say that, it is a deepfake" becomes a plausible defense even for genuine recordings. Current countermeasures like detection technologies, watermarking, authentication systems, and legal frameworks have limitations. We need industry-wide standards for responsible S2S development, built-in safeguards that prevent misuse, public education about synthetic audio, and international cooperation on legal frameworks.

Another ethical concern is bias in training data. S2S AI systems learn from training data, and if that data reflects societal biases, the systems will perpetuate and potentially amplify those biases. Models may perform better for majority groups than for underrepresented groups, creating unfair access to the technology. Certain accents or dialects may be treated as standard while others are marked as non-standard, reinforcing linguistic discrimination.

Prosodic norms and emotional expression vary across cultures. A model trained primarily on Western data might misinterpret or inappropriately modify emotional expression from other cultures. Training data often overrepresents educated, affluent speakers, potentially making systems less effective for working-class or economically disadvantaged users. People whose voices do not match the training data distribution may find S2S systems do not work well for them, effectively excluding them from services and opportunities.

Questions about consent and ownership are also important. S2S technology raises complex questions about voice ownership and control. Do you own your voice? Can you control how it is used? What rights do you have if someone creates a synthetic version of your voice? Should companies be allowed to use voice recordings to train S2S models without explicit consent? If an S2S system can generate speech in your voice, should it require your permission each time?

Voice rights are poorly defined in most jurisdictions. Voice actors whose livelihoods depend on their unique voices are concerned about being replaced by S2S technology. Some people might want to donate their voices for beneficial uses, like giving voice to people with speech impairments, but there are no clear frameworks for doing this ethically and legally. We need clear legal and ethical frameworks that define voice rights and ownership, establish consent requirements, provide compensation mechanisms, and balance innovation with protection of individual rights.

Finally, there are concerns about transparency and accountability. S2S AI systems are complex black boxes. Even their developers often cannot fully explain why they produce specific outputs. This lack of transparency creates accountability challenges. When an S2S system makes an error or produces problematic output, it is often difficult to understand why, making it hard to fix problems or assign responsibility.

People interacting with S2S systems may not realize they are talking to AI, especially as the technology becomes more convincing. This raises concerns about deception and informed consent. When S2S systems cause harm, it is unclear who is responsible: the developers, the deployers, the users, or the systems themselves. The complexity of S2S systems makes them difficult to audit for bias, errors, or malicious behavior.

We need clear standards for transparency in S2S development and deployment, mandatory disclosure when people are interacting with S2S systems, liability frameworks that assign responsibility for S2S-related harms, independent oversight and auditing mechanisms, and public registries of deployed S2S systems and their capabilities.

Looking to the Future

S2S AI represents a fundamental shift in how machines process and generate human speech. By eliminating the intermediate text conversion step, S2S systems achieve better naturalness, lower latency, and better preservation of emotional and speaker characteristics compared to traditional approaches. This is not just an incremental improvement. It is a qualitative change that makes human-machine interaction feel genuinely natural for the first time.

The technology's impact extends across many areas. Education becomes more personalized and accessible with AI tutors that can engage in natural, emotionally-aware conversations. Healthcare benefits from more efficient documentation, better patient communication, and accessible interfaces for people with disabilities. Customer service becomes more empathetic and satisfying with AI agents that can understand and respond to emotional context. Global communication becomes more seamless with real-time translation that preserves voice and emotion across language barriers. Accessibility improves dramatically, giving voice to people with speech impairments and enabling more inclusive technology.

However, realizing the full potential of S2S AI while avoiding its risks requires collaboration among many groups. Researchers must continue advancing the technology while prioritizing fairness, transparency, and safety. Developers must implement responsible AI practices, including bias testing, privacy protection, and misuse prevention. Policymakers must create legal frameworks that protect individual rights while enabling beneficial innovation. Civil society must advocate for fair access and hold developers accountable. Users must be educated about S2S capabilities and limitations and empowered to control their voice data.

Future Research Directions

The field of S2S AI is rapidly evolving with several promising research directions. Enhanced multilingual capabilities are needed to develop models that work seamlessly across all the world's languages, including low-resource and endangered languages. This requires new approaches to transfer learning and few-shot adaptation.

Improved prosody control will create systems with finer-grained control over emotional expression, speaking style, and prosodic features, enabling more nuanced and context-appropriate speech generation. Personalization and adaptation will build S2S systems that can adapt to individual users' preferences, communication styles, and needs, learning from user feedback and adjusting behavior over time.

Robustness and reliability improvements will enhance performance in challenging conditions like noisy environments, poor audio quality, and diverse accents, while reducing errors that could cause miscommunication or harm. Efficiency and sustainability work will develop more computationally efficient models that can run on mobile devices and edge hardware, reducing energy consumption and environmental impact.

Privacy-preserving techniques will create methods for training and deploying S2S systems that protect user privacy, such as federated learning, differential privacy, and secure multi-party computation. Ethical frameworks will establish industry standards, legal frameworks, and technical safeguards to ensure responsible development and deployment of S2S technology.

Final Thoughts

S2S AI is not just a technological advancement. It is a step toward more natural, empathetic, and inclusive human-machine interaction. As the technology matures, it has the potential to break down communication barriers, enhance accessibility, and create new forms of expression and connection.
However, this potential can only be realized if we address the significant technical and ethical challenges the technology presents. We must ensure that S2S AI works fairly for all people, regardless of their language, accent, or background. We must protect against misuse while enabling beneficial applications. We must respect individual rights to privacy and voice ownership while advancing the technology.

The future of S2S AI is not predetermined. It will be shaped by the choices we make today. By approaching this powerful technology with both enthusiasm and caution, technical innovation and ethical reflection, we can create a future where machines understand and speak with us in ways that feel genuinely human, opening new possibilities for communication, creativity, and connection.

References

1. Ao, J., Wang, R., Zhou, L., et al. (2022). "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing." arXiv:2110.07205. Available at: https://arxiv.org/abs/2110.07205
2. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). "Attention Is All You Need." Advances in Neural Information Processing Systems, 30.
3.M icrosoft SpeechT5 GitHub Repository. Available at: https://github.com/microsoft/SpeechT5
4.Partnership on AI. Guidelines for responsible development of synthetic media.
5.European Union. "EU AI Act: Regulatory framework for high-risk AI systems including speech technologies."