Language barrier in spoken communication with locals is one of the major challenges that foreign tourists encounter when traveling to Vietnam. Vietnamese is a tonal language with 6 different tones that dictate the meaning of a word, making it difficult for non-native speakers to quickly absorb in a short amount of time (especially English speakers). We would like to propose a prototype of a cross-lingual voice translator application, which allows English-speaking travelers to order food, ask for directions, and communicate with locals by tapping their phones. Simultaneously, local vendors can also benefit from our system as they can provide services to foreign tourists without having to deal with unwanted language barriers. The proposed system consists of 3 core modules as illustrated below: (1) ASR to transcribe English audios to corresponding texts, (2) English-Vietnamese machine translation, and (3) voice cloning-based TTS which can generate Vietnamese-spoken audios using the original speaker's voice. Our main contribution is the voice cloning-based TTS model for Vietnamese, which incorporates the speaker embedding of English input audios to synthesize Vietnamese-spoken speeches with the voice of original speakers.