SURVEY: TEXT-TO-SPEECH SYNTHESIS

|  |  |
| --- | --- |
| Abstract  Tổng hợp tiếng nói là task tạo ra tín hiệu tiếng nói tương ứng với văn bản đầu vào đã cho. Đầu ra được kỳ vọng nghe tự nhiên nhất có thể (ngữ điệu giống với người nói) và dễ hiểu (phát âm đúng).  Các kiến trúc SOTA cho TTS theo 3 giai đoạn xử lý: 1. chuyển ngữ của các biểu đồ đầu vào (input graphemes thành các âm vị sử dụng phonemizer,  2. chuyển đổi chuỗi âm vị thành melspectrogram biểu diễn taand số thời gian  3. Cuối cùng, tạo bản thô dạng sóng sử dụng melspectrogram.  Bài báo giới thiệu đa dạng kiến trúc cho text-to-mel và vocoder model.   1. INTRODUCTION    1. Problem Statement   Input: Text  Output: Speech signal (nghe tự nhiên và dễ hiểu)     * 1. Motivation   Tổng hợp tiếng nói, sản phẩm nhân tạo của tiếng nói con người là bài toán mà con người cố gắng giải quyết trong thời gian dài.  1779, nhà khoa học người Đan Mạch gốc Đức mô hình hóa giọng nói con người với 5 nguyên âm dài trong tiếng Anh.  Sau đó, nghiên cứu điều hòa tổng hợp giọng nói cho văn bản.  Ứng dụng:  + Hỗ trợ người khiếm thị /khuyết tật/ tổn thương thanh quản có thể đọc hiểu văn bản.  + Nhận dạng giọng nói tự động và dịch máy, loại bỏ rào cản giữa những người nói các ngôn ngữ khác nhau.   1. CLASSICAL APPROACHES TO SPEECH SYNTHESIS   Sử dụng CSDL của các âm đơn vị. Sóng tiếng nói thô được tạo bằng cách ghép nối các đơn vị tiếng nói theo thứ tự thích hợp.  = > Âm thanh tạo ra không tự nhiên   * 1. Diphone-based Speech Synthesis   Diphone được tạo từ 2 âm vị (đơn vị nhỏ nhất của âm) kết nối với nhau.  Có nhiều phương pháp dựa trên diphone sử dụng kỹ thuật xử lý tín hiệu như Pitch Synchronous Overlap-Add (PSOLA 1989), Frequency Domain PSOLA (FD-PSOLA), Time Domain PSOLA (TD-PSOLA). Những phương pháp này, tiếng nói đầu tiên được chia ra làm các segments nhỏ hơn (ở cấp độ diphones) sau đó kết hợp lại tạo ra output mong đợi (bổ sung chồng chéo overlapped addition).  Phân tích cấp độ ký tự của văn bản (đồ thị hoặc chuyển hóa thành các âm vị) có thể đóng vai trò như thông tin cho việc lựa chọn đơn vị tiếng nói nhỏ hơn.  + PSOLA: Tiếng nói được chia thành dạng sóng ngắn hạn đồng bộ *cao độ* (pitch), sau đó được thay đổi theo thời gian hoặc miền quang phổ để thu phiên bản tổng hợp khác nhau của một đơn vị tiếng nói. Vì phương pháp này hoạt động trực tiếp trên các dạng sóng thô của tín hiệu giọng nói nên không có hiện tượng mất mát thông tin.  + FD-PSOLA:  Đầu tiên, bao phủ của phổ (spectral envelop) được tính nhờ dùng phân tích dự báo tuyến tính.  Kế đến, cao độ của đoạn tiếng nói được modified thông qua một tuyến tính nội suy, bao gồm các đơn vị tổng hợp tiếng nói.  Do những thay đổi về độ lớn của domain tần số (không xét đến pha), nên có sự không liên tục không tự nhiên tại các ranh giới nối.  + TD-PSOLA: Âm điệu của dạng sóng giọng nói có thể được điều khiển bằng cách sử dụng phương pháp này. Điều này dẫn đến việc tạo ra các sửa đổi về cao độ và thang thời gian chất lượng cao. TD-PSOLA hiệu quả về mặt tính toán, nhưng yêu cầu bộ dữ liệu lớn.   * 1. Corpus-based Speech Synthesis   Chọn các âm vị (đơn vị âm thanh) từ CSDL lớn và kết nối chúng và tối tiểu hóa hàm chi phí. Văn bản có thể có chú thích bổ sung bao gồm thông tin prosodic và ngữ cảnh của âm vị.     1. TEXT TO MEL SPECTROGRAM    1. WaveNet    2. Tacotron2    3. FastSpeech    4. ForwardTacotron    5. Recent Advancements 2. VOCODERS    1. Waveglow    2. MelGAN    3. Parallel WaveGAN    4. HiFi-GAN    5. GAN Vocoder    6. CARGAN 3. VOICE CONVERSION    1. CycleGAN-VC Models    2. MaskCycleGAN-VC 4. Summary | Túm tắt:  Diphone: nối 2 âm vị có chồng lấp  Pitch?  Pitch-synchronous?  A **spectral envelope** is a curve which envelopes the magnitude STS, i.e. it wraps tightly around it, linking the peaks of the sinusoidal partials or passing close to the maxima of non-sinusoidal spectra.  Conv In Frequence |