استاد : آقای دکتر اسلامی

دانشجو : سورنا لطفی ارجمند

شماره دانشجویی : 40114140111009

Abstract

Emotional voice conversion (EVC) aims to change the emotional state of an utterance while preserving the linguistic content and speaker identity. In this paper, we propose a novel 2-stage training strategy for sequence-to-sequence emotional voice conversion with a limited amount of emotional speech data. We note that the proposed EVC framework leverages textto-speech (TTS) as they share a common goal that is to generate high-quality expressive voice. In stage 1, we perform style initialization with a multi-speaker TTS corpus, to disentangle speaking style and linguistic content. In stage 2, we perform emotion training with a limited amount of emotional speech data, to learn how to disentangle emotional style and linguistic information from the speech. The proposed framework can perform both spectrum and prosody conversion and achieves significant improvement over the state-of-the-art baselines in both objective and subjective evaluation. Index Terms: Emotional voice conversion, sequence-tosequence, limited data

هدف تبدیل صدای عاطفی (EVC) تغییر حالت احساسی یک گفته در عین حفظ زبانی است.

محتوا و هویت گوینده در این مقاله، ما یک رمان را پیشنهاد می کنیم

استراتژی آموزش 2 مرحله ای برای عاطفی دنباله به سکانس

تبدیل صدا با مقدار محدودی از گفتار احساسی

داده ها. توجه داریم که چارچوب پیشنهادی EVC از متن به گفتار (TTS) استفاده می‌کند، زیرا آنها هدف مشترکی را ایجاد می‌کنند.

صدای رسا با کیفیت بالا در مرحله 1 سبک اجرا می کنیم

مقداردهی اولیه با یک پیکره TTS چند بلندگو، برای جدا کردن

سبک گفتاری و محتوای زبانی در مرحله 2 اجرا می کنیم

آموزش احساسات با مقدار محدود گفتار احساسی

داده ها، برای یادگیری نحوه تفکیک سبک عاطفی و زبانی

اطلاعات از سخنرانی چارچوب پیشنهادی می تواند

هم تبدیل طیف و هم تبدیل عروضی را انجام می دهد و به دست می آورد

بهبود قابل توجهی نسبت به خطوط پایه پیشرفته در

ارزیابی عینی و ذهنی

اصطلاحات شاخص: تبدیل صدای احساسی، توالی-توالی، داده های محدود

1. Introduction

Sequence-to-sequence (seq2seq) speech synthesis frameworks, such as Tacotron [1], can generate high-quality synthetic speech. However, such frameworks heavily rely on a large amount of training data. Furthermore, they generally lack emotional variance [2]. Emotional voice conversion aims to convert the emotional state of speech from one to another while preserving the linguistic content and speaker identity. This technique allows us to project a desired emotion into the generated speech, thus bears huge potential in real-world applications, such as expressive text-to-speech [3]. Emotion is inherently supra-segmental and complex with multiple signal attributes concerning both spectrum and prosody [4], thus it is insufficient to convert the emotion only with frame-wise spectral mapping. Prosodic features, such as pitch, energy, and duration, also need to be dealt with for emotional voice conversion. We believe that seq2seq training is a better solution for spectrum and duration conversion in EVC, which will be the focus of this paper. Emotional voice conversion is a special type of voice conversion [5]. Previous studies are focused on frame-based mapping of spectral features of source and target, including using statistical methods [6, 7] and deep learning methods, such as deep neural network [8], generative adversarial network (GAN) [9] and CycleGAN [10]. Inspired by the success in speaker voice conversion, these methods are adopted to model both spectral and prosodic parameters for emotional voice conversion. Successful attempts include GMM [11],

Sparse representation [12], deep bi-directional long-short-term memory (BLSTM) network [13], GAN-based [14–17] and autoencoder-based [18–21] methods. These frameworks model the mapping on a frame-by-frame basis. As emotional prosody is hierarchical in nature [4], frame-based methods are therefore not the best in handling prosody conversion [5].

.1مقدمه

چارچوب های سنتز گفتار دنباله به دنباله (seq2seq) مانند تاکوترون [1] ،

می تواند گفتار مصنوعی با کیفیت بالا تولید کند. با این حال، چنین چارچوب هایی به شدت به مقدار زیادی از داده های آموزشی متکی هستند.

علاوه بر این، آنها به طور کلی فاقد واریانس عاطفی [2] هستند.

تبدیل صدای عاطفی با هدف تبدیل حالت احساسی گفتار از یکی به دیگری با حفظ محتوای زبانی و هویت گوینده انجام می شود.

این تکنیک به ما امکان می دهد تا یک احساس دلخواه را در گفتار تولید شده فرافکنی کنیم، بنابراین پتانسیل عظیمی را در کاربردهای دنیای واقعی مانند تبدیل متن به گفتار بیانگر دارد [3].

احساسات ذاتاً فرابخشی و پیچیده با ویژگی های سیگنالی متعدد در رابطه با طیف و عروض است [4]، بنابراین برای تبدیل احساسات فقط با نگاشت طیفی چارچوبی کافی نیست. ویژگی‌های عروضی، مانند زیر و بم، انرژی و مدت زمان نیز باید برای تبدیل صدای احساسی مورد توجه قرار گیرند.

ما معتقدیم که آموزش seq2seq راه حل بهتری برای تبدیل طیف و مدت زمان در EVC است که تمرکز این مقاله خواهد بود.

تبدیل صدای احساسی نوع خاصی از تبدیل صدا است [5].

مطالعات قبلی بر روی نقشه برداری مبتنی بر فریم از ویژگی های طیفی منبع و هدف، از جمله استفاده از روش های آماری متمرکز شده اند [6، 7] و روش های یادگیری عمیق، مانند شبکه عصبی عمیق [8]، شبکه متخاصم مولد (9] (GAN] و 10] CycleGAN].

با الهام از موفقیت در تبدیل صدای بلندگو، این روش‌ها برای مدل‌سازی پارامترهای طیفی و عروضی برای تبدیل صدای احساسی اتخاذ می‌شوند.

تلاش های موفق شامل 11] GMM]، نمایش پراکنده [12]، شبکه حافظه کوتاه مدت دو جهته عمیق (13] (BLSTM]، مبتنی بر GAN 14-17]] و روش‌های مبتنی بر رمزگذار خودکار [18-21]. این چارچوب ها نقشه برداری را بر اساس فریم به فریم مدل می کنند.

از آنجایی که عروض عاطفی ماهیت سلسله مراتبی دارد [4]، بنابراین روش‌های مبتنی بر فریم در مدیریت تبدیل عروضی بهترین نیستند [5].

Recently, seq2seq models with attention mechanism have attracted much interests in speech synthesis [1, 22] and voice conversion such as SCENT [23], AttS2S-VC [24] and ConvS2S-VC [25]. Considering that VC and TTS share a similar motivation in a sense that they both aim to generate speech from internal representations [5], there are studies to leverage TTS systems to further improve seq2seq VC performance, such as adding text supervision [26] or leveraging TTS [27, 28]. Inspired by these studies, seq2seq frameworks have become popular in emotional voice conversion. For example, a seq2seq model is proposed in [29] to jointly model pitch and duration with parallel data. In [30], researchers propose a seq2seq model with multi-task learning for both emotional voice conversion and emotional text-to-speech. We note that these frameworks require tens of hours of emotional speech data to train, which is not practical for real-life scenarios

اخیراً مدل‌های seq2seq با مکانیسم توجه علاقه‌های زیادی را در سنتز گفتار به خود جلب کرده‌اند [1، 22] و تبدیل صدا مانند SCENT [23]،[24] AttS2S-VC و25] ConvS2S-VC.

با توجه به اینکه VC و TTS انگیزه مشابهی دارند به این معنا که هدف هر دو تولید گفتار از بازنمایی های داخلی است [5]، مطالعاتی برای استفاده از سیستم های TTS برای بهبود بیشتر عملکرد seq2seq VC وجود دارد، مانند اضافه کردن نظارت بر متن [26] یا استفاده از TTS 27و28 .

با الهام از این مطالعات، چارچوب‌های seq2seq در تبدیل صدای احساسی محبوب شده‌اند. به عنوان مثال، یک مدل seq2seq در پیشنهاد شده است29 برای مدل سازی مشترک گام و مدت زمان با داده های موازی.

در [30]، محققان یک مدل seq2seq با یادگیری چند وظیفه ای را برای تبدیل صدای احساسی و متن به گفتار احساسی پیشنهاد کردند. توجه داریم که این چارچوب‌ها به ده‌ها ساعت داده‌های گفتاری احساسی برای آموزش نیاز دارند، که برای سناریوهای واقعی عملی نیست.

In this paper, we propose a 2-stage training strategy for seq2seq emotional voice conversion. In stage 1, we perform style initialization, which aims to disentangle speaking style and the linguistic content with a multi-speaker TTS corpus. In stage 2, we perform emotion training, where all components of the network are trained with limited emotional speech data. By doing this, we obtain emotion encoder that learns to disentangle emotional style from the speech, and emotion classifier that further eliminates the emotion-related information in the linguistic space. The proposed framework achieves remarkable

performance by converting both spectrum and prosody with a limited amount of non-parallel emotional speech data.

در این مقاله، ما یک استراتژی آموزشی 2 مرحله‌ای برای تبدیل صدای عاطفی seq2seq پیشنهاد می‌کنیم. در مرحله 1، مقداردهی اولیه سبک را انجام می دهیم، که هدف آن تفکیک سبک گفتاری و محتوای زبانی با یک مجموعه TTS چند سخنران است. در مرحله 2، آموزش احساسات را انجام می دهیم، جایی که تمام اجزای شبکه با داده های گفتاری احساسی محدود آموزش می بینند.

با انجام این کار، رمزگذار احساسی به دست می‌آوریم که می‌آموزد سبک عاطفی را از گفتار جدا کند و طبقه‌بندی کننده احساسات که اطلاعات مربوط به احساسات را در فضای زبانی حذف می‌کند. چارچوب پیشنهادی با تبدیل هر دو طیف و عروض با مقدار محدودی از داده‌های گفتار عاطفی غیر موازی، به عملکرد قابل‌توجهی دست می‌یابد.

mited amount of non-parallel emotional speech data. The main contributions of this paper include: 1) we propose a seq2seq emotional voice conversion framework leveraging TTS without the need for parallel data, and flexible for manyto-many emotional voice conversion; 2) we propose a novel training strategy that requires a small amount of emotionlabelled data; 3) we significantly improve the performance by modelling the alignment between acoustic and linguistic embedding for emotion styles, which is a departure from framebased conversion paradigm; 4) we propose emotional finetuning for WaveRNN vocoder [31] training with the limited amount of emotional speech data to further improve the final performance. This paper is organized as follows: In Section 2, we motivate our study through the comparison with existing seq2seq EVC frameworks. In Section 3, we introduce our proposed framework and the proposed training strategy. In Section 4, we report the experiments. Section 5 concludes the study

مقدار کم داده های گفتاری عاطفی غیر موازی مشارکت های اصلی این مقاله عبارتند از: 1) ما یک چارچوب تبدیل صدای عاطفی seq2seq را پیشنهاد می‌کنیم که از TTS بدون نیاز به داده‌های موازی استفاده می‌کند و برای تبدیل صدای عاطفی بسیاری انعطاف‌پذیر است. 2) ما یک استراتژی آموزشی جدید پیشنهاد می کنیم که به مقدار کمی از داده های دارای برچسب احساسات نیاز دارد. 3) ما عملکرد را به طور قابل توجهی با مدل‌سازی هم‌ترازی بین جاسازی صوتی و زبانی برای سبک‌های احساسات بهبود می‌بخشیم، که انحراف از پارادایم تبدیل مبتنی بر چارچوب است. 4) ما تنظیم دقیق احساسی را برای Vocoder WaveRNN پیشنهاد می کنیم [31] آموزش با مقدار محدود داده های گفتاری احساسی برای بهبود بیشتر عملکرد نهایی.

این مقاله به شرح زیر تنظیم شده است: در بخش 2، ما انگیزه مطالعه خود را از طریق مقایسه با چارچوب‌های seq2seq EVC موجود ایجاد می‌کنیم. در بخش 3، چارچوب پیشنهادی خود و استراتژی آموزشی پیشنهادی را معرفی می‌کنیم. در بخش 4، آزمایش‌ها را گزارش می‌کنیم. بخش 5 مطالعه را به پایان می رساند.

## 2. Sequence-to-sequence

EVC The seq2seq model, which was first studied in machine translation [32], was found effective in speech synthesis [1, 22] and voice conversion [23, 25, 27, 28]. In voice conversion, the seq2seq model with attention mechanism has greatly improved the modelling ability by jointly learning the feature mapping and alignment. The seq2seq model marks a departure from the frame-wise modelling [5, 33]. First, the seq2seq model allows for the prediction of the speech duration at the run-time inference which is an essential factor of emotional prosody [11]. Second, emotion labels are usually annotated at the utterance level in speech corpus [33], while emotional prosody is suprasegmental and can be associated with only a few words. The attention mechanism makes it possible for the conversion to

focus on emotion-relevant regions, which will be our focus. There are only a few studies on emotional voice conversion with seq2seq modelling such as jointly modelling pitch and duration with parallel data [29], where the output pitch contour is conditioned on the syllable position and source signal; and multi-task learning where a single system is jointly trained for both emotional voice conversion and text-to-speech [30]. These frameworks perform well but rely on a large emotional speech corpus. In this paper, we would like to study a limited data solution. To the best of our knowledge, this is the first attempt with the seq2seq model that does not need a large amount of emotional speech training data for EVC.

2. دنباله به دنباله

EVC مدل seq2seq که برای اولین بار در ترجمه ماشینی مورد مطالعه قرار گرفت [32]، در سنتز گفتار مؤثر بود [1، 22] و تبدیل صدا [23، 25، 27، 28].

در تبدیل صدا، مدل seq2seq با مکانیسم توجه، توانایی مدل‌سازی را با یادگیری مشترک نقشه‌برداری و هم‌ترازی ویژگی‌ها بسیار بهبود بخشیده است. مدل seq2seq نشان دهنده انحراف از مدل سازی چارچوبی است [5، 33].

اول، مدل seq2seq امکان پیش‌بینی مدت زمان گفتار را در استنتاج زمان اجرا فراهم می‌کند که یک عامل اساسی عروض عاطفی است [11]. دوم، برچسب‌های احساسات معمولاً در سطح بیان در مجموعه گفتار حاشیه‌نویسی می‌شوند [33]، در حالی که عروض عاطفی فرابخشی است و تنها با چند کلمه می تواند مرتبط باشد. مکانیسم توجه این امکان را فراهم می کند که تبدیل بر روی مناطق مربوط به احساسات متمرکز شود، که تمرکز ما خواهد بود. تنها مطالعات کمی در مورد تبدیل صدای احساسی با مدل‌سازی seq2seq وجود دارد، مانند مدل‌سازی مشترک زیر و بم و مدت زمان با داده‌های موازی [29]، جایی که کانتور گام خروجی به موقعیت هجا و سیگنال منبع مشروط است، و یادگیری چند وظیفه ای که در آن یک سیستم واحد به طور مشترک هم برای تبدیل صدای احساسی و هم برای تبدیل متن به گفتار آموزش داده می شود [30].

این چارچوب ها به خوبی عمل می کنند اما بر یک مجموعه گفتاری عاطفی بزرگ متکی هستند. در این مقاله، ما می خواهیم یک راه حل داده محدود را مطالعه کنیم. تا جایی که ما می دانیم، این اولین تلاش با مدل seq2seq است که به حجم زیادی از داده های آموزش گفتار احساسی برای EVC نیاز ندارد.

3. Proposed seq2seq EVC model

We propose a seq2seq EVC framework that consists of 5 components, a text encoder, a seq2seq automatic speech recognition (ASR) encoder, a style encoder, a classifier, and a seq2seq decoder. We propose a 2-stage training strategy: 1) Stage I: Style initialization, which disentangles between the speaking style, i.e., speaker style, and the linguistic content with a multispeaker TTS corpus; 2) Stage II: Emotion training, where all components, initialized by stage I, are further trained with a limited amount of emotional speech data. Finally, during run-time conversion inference, the framework generates the utterance with reference emotion type by combining the source linguistic representation and the reference emotion representation. While the proposed model is trained to perform both EVC and emotional TTS, EVC will be the main focus of this paper.

3. مدل seq2seq EVC پیشنهادی

ما یک چارچوب seq2seq EVC را پیشنهاد می‌کنیم که از 5 جزء، یک رمزگذار متن، یک رمزگذار تشخیص خودکار گفتار (ASR seq2seq)، یک رمزگذار سبک، یک طبقه‌بندی‌کننده و یک رمزگشای seq2seq تشکیل شده است.

ما یک استراتژی آموزشی 2 مرحله ای را پیشنهاد می کنیم: 1) مرحله I: مقدار دهی اولیه سبک، که بین سبک گفتاری، یعنی سبک سخنران، و محتوای زبانی با یک مجموعه TTS چند سخنران جدا می شود. 2) مرحله دوم: آموزش عاطفی، که در آن همه مؤلفه‌ها، که توسط مرحله اول مقداردهی شده‌اند، با مقدار محدودی از داده‌های گفتار عاطفی بیشتر آموزش داده می‌شوند. در نهایت، در طول استنتاج تبدیل زمان اجرا، چارچوب با ترکیب بازنمایی زبانی مبدا و بازنمایی هیجان مرجع، گفتار را با نوع احساس مرجع تولید می‌کند. در حالی که مدل پیشنهادی برای انجام هر دو EVC و TTS احساسی آموزش داده شده است، EVC تمرکز اصلی این مقاله خواهد بود.

3.1. Training stage I: Style initialization

At stage I, we adopt a seq2seq VC framework [28], and pretrain it with a publicly available TTS corpus, as shown in Figure 2(a). The framework takes the acoustic features and one-hot phoneme sequences as the inputs. The text encoder and the seq2seq ASR encoder predict the linguistic embeddings from the audio input and the text input respectively. The style encoder embeds the acoustic features into the style embedding. Finally, the seq2seq decoder recovers the acoustic features with the style and linguistic embeddings either from audio or text inputs

در مرحله اول، ما یک چارچوب seq2seq VC را اتخاذ می کنیم [28]، و آن را با یک مجموعه TTS در دسترس عموم، همانطور که در شکل 2 (الف) نشان داده شده است، از قبل آموزش دهید.

چارچوب ویژگی‌های آکوستیک و دنباله‌های واجی تک داغ را به عنوان ورودی می‌گیرد. . رمزگذار متن و رمزگذار seq2seq ASR تعبیه‌های زبانی را به ترتیب از ورودی صدا و ورودی متن پیش‌بینی می‌کنند. . رمزگذار سبک ویژگی های صوتی را در جاسازی سبک تعبیه می کند. در نهایت، رمزگشای seq2seq ویژگی های صوتی را با سبک و جاسازی های زبانی از ورودی های صوتی یا متنی بازیابی می کند.

In stage I, the style encoder learns speaker-dependent information, i.e., speaker style, and excludes linguistic information from the acoustic features. To disentangle from speaker style, an adversarial training with a classifier is employed to further eliminate speaker information from the linguistic space. With the text inputs and adversarial training strategy, the framework learns to disentangle the linguistic and style information through a multi-speaker TTS corpus.

در مرحله اول، رمزگذار سبک اطلاعات وابسته به سخنران را می آموزد، به عنوان مثال، سبک سخنران، و اطلاعات زبانی را از ویژگی های صوتی حذف می کند. برای جدا شدن از سبک گوینده، یک آموزش خصمانه با یک طبقه‌بندی برای حذف بیشتر اطلاعات گوینده از فضای زبانی استفاده می‌شود. با ورودی‌های متن و استراتژی آموزش رقیب، چارچوب یاد می‌گیرد که اطلاعات زبانی و سبک را از طریق یک مجموعه TTS چند سخنران از هم جدا کند.

شکل بالای صفحه ی دوم سمت راست

Figure 1: Visualization of emotion embeddings derived from (a) style encoder and (b) emotion encoder. Each point represents the emotion embedding of a reference utterance

شکل 1: تجسم تعبیه‌های احساسات برگرفته از (الف) رمزگذار سبک و (ب) رمزگذار احساسات. هر نقطه نمایانگر تعبیه احساسات یک گفته مرجع است

However, since the style encoder learns the style information from an emotion-neutral TTS corpus, it does not learn to encode any specific speaking style during stage I, as shown in Figure 1(a). However, the style encoder has rich knowledge about the style and speaker information, we believe it has the

potential to learn the emotional style representation given a small amount of emotional speech data. Therefore, we consider stage I as the style initialization, and propose emotion training in stage II, where the style encoder acts as an emotion encoder to learn the emotional style representations.

با این حال، از آنجایی که رمزگذار سبک اطلاعات سبک را از یک پیکره TTS خنثی می‌آموزد، همانطور که در شکل 1(a) نشان داده شده است، در طول مرحله I یاد نمی‌گیرد که سبک صحبت خاصی را رمزگذاری کند. با این حال، رمزگذار سبک دانش غنی در مورد سبک و اطلاعات سخنران دارد، ما معتقدیم که با توجه به مقدار کمی از داده‌های گفتار احساسی، توانایی یادگیری بازنمایی سبک احساسی را دارد. بنابراین، ما مرحله I را به عنوان مقداردهی اولیه سبک در نظر می گیریم و آموزش احساسات را در مرحله دوم پیشنهاد می کنیم، جایی که رمزگذار سبک به عنوان رمزگذار احساسات برای یادگیری بازنمایی سبک احساسی عمل می کند.

3.2. Training stage II: Emotion training

We propose to retrain the framework with a limited amount of emotional speech data at stage II, as shown in Figure 2(b). We expect that the network has learnt the basic functions of VC and TTS with a styling mechanism during stage I. The style encoder is then ready to learn the emotional styles from additional emotion-labelled speech data. It acts as an emotion encoder to embed the acoustic features into an emotion vector $h^e$ . Furthermore, the classifier acts as an emotion classifier to eliminate the emotion information in the linguistic space. Both the emotion encoder and emotion classifier are trained in a supervised way with a one-hot emotion ID. 3.2.1. Training with limited emotion data The emotion encoder learns the emotional representations through the loss function Lc as below فرمول صفحه دوم سمت راست وسط صفحه :

3.2. مرحله دوم آموزش: آموزش احساسات

ما پیشنهاد می‌کنیم که چارچوب را با مقدار محدودی از داده‌های گفتار عاطفی در مرحله II بازآموزی کنیم، همانطور که در شکل 2 (b) نشان داده شده است. ما انتظار داریم که شبکه عملکردهای اساسی VC و TTS را با مکانیزم یک ظاهر طراحی در مرحله I آموخته باشد. سپس رمزگذار سبک آماده است تا سبک‌های احساسی را از داده‌های گفتاری برچسب‌گذاری شده با احساسات اضافی بیاموزد. این به عنوان یک رمزگذار احساسات عمل می کند تا ویژگی های صوتی را در یک بردار احساس $h^e$ جاسازی کند. علاوه بر این، طبقه‌بندی‌کننده به‌عنوان طبقه‌بندی‌کننده احساسات برای حذف اطلاعات احساسات در فضای زبانی عمل می‌کند. هم رمزگذار احساسات و هم طبقه‌بندی‌کننده احساسات به روشی تحت نظارت و با شناسه احساسی تکی.1.2.3 آموزش داده می‌شوند. آموزش با داده های احساسی محدود رمزگذار احساسات بازنمایی های احساسی را از طریق تابع ضرر Lc به شرح زیر می آموزد. فرمول

where CE(·) represents the cross entropy loss function, N represents the length of embedding sequence, ϕ and ϕˆn denote the one-hot emotion label and the predicted emotion probability respectively. As for the emotion classifier C s , the adversarial loss Ladv is modified as follows:فرمول دوم صفحه دوم

که در آن CE(·) تابع از دست دادن آنتروپی متقاطع را نشان می دهد، N

نشان دهنده طول دنباله جاسازی، ϕو ϕˆn نشان دهنده برچسب احساس یک داغ و احتمال احساسات پیش بینی شده به ترتیب. در مورد طبقه بندی احساساتCs ، متخاصم ضرر Ladv به شرح زیر اصلاح می شود: فرمول دوم صفحه دوم

where α = [1/R, ..., 1/R] T is an uniform distribution over the total number of emotion types R. And the emotion classification loss Lec is given as: فرمول سوم صفحه دوم

که در آن α = [1/R, ..., 1/R] T

Tیک توزیع یکنواخت بر روی تعداد کل انواع احساسات R است. و ضایعات طبقه بندی احساسات Lec به صورت زیر ارائه می شود: فرمول سوم صفحه

where V is the weight matrix of the emotion encoder. We note that all the components are initialized with the weights learnt at stage I, while the last projection layers of the emotion encoder and the emotion classifier are randomly initialized. At stage II, we update the entire network during training. The training allows the seq2seq ASR encoder and seq2seq decoder to learn a better alignment between acoustic frames and linguistic embedding sequence that particularly characterizes the emotional style of the utterance. Furthermore, we also adapt the speaker style encoder of stage I to an emotion

که در آن V ماتریس وزن رمزگذار احساسات است.

توجه می کنیم که تمام اجزاء با مقداردهی اولیه می شوند

وزن ها در مرحله I آموخته شد، در حالی که آخرین لایه های طرح ریزی از

رمزگذار احساسات و طبقه بندی کننده احساسات به صورت تصادفی هستند

اولیه شده است. در مرحله دوم، کل شبکه را در طول به روز می کنیم

آموزش. آموزش اجازه می دهد تا رمزگذار seq2seq ASR و

رمزگشا seq2seq برای یادگیری تراز بهتر بین آکوستیک

فریم ها و توالی تعبیه زبانی که به ویژه

سبک احساسی بیان را مشخص می کند. علاوه بر این،

ما همچنین رمزگذار سبک سخنران مرحله I را با یک احساس تطبیق می دهیم

صفحه سوم

نوشته ی زیر شکل:

Figure 2: The proposed 2-stage training strategy for seq2seq emotional voice conversion with limited emotional speech data.

شکل 2: استراتژی آموزش 2 مرحله ای پیشنهادی برای تبدیل صدای عاطفی seq2seq با داده های گفتار احساسی محدود.

encoder, the speaker classifier of stage I to an emotion classifier. Overall, the framework leverages the knowledge of disentanglement between linguistic and style information learnt at stage I, and effectively learns the emotional style disentanglement only with a limited amount of emotional speech data at stage II. The proposed 2-stage training further helps with obtaining better disentangled emotional representation without the support of large emotional speech data.

رمزگذار، طبقه‌بندی‌کننده بلندگو مرحله اول به طبقه‌بندی‌کننده احساسات.

به طور کلی، این چارچوب از دانش گسستگی بین اطلاعات زبانی و سبک آموخته شده استفاده می کند

در مرحله اول، و به طور موثری گره گشایی سبک عاطفی را تنها با مقدار محدودی از داده های گفتاری احساسی می آموزد

در مرحله دوم آموزش 2 مرحله ای پیشنهادی بیشتر کمک می کند

به دست آوردن بازنمایی عاطفی گسسته بهتر بدون پشتیبانی از داده های گفتاری عاطفی بزرگ

3.2.2. Style encoder vs. emotion encoder

During the emotion training, the style encoder acts as an emotion encoder and takes emotion ID as the input to effectively learn the emotion representation in the speech. To validate our idea, we use t-SNE [34] to visualize the emotion embedding of the reference utterances, which are derived by the style encoder from stage I and the emotion encoder from stage II respectively. To our delight, as shown in Figure 1, the emotion embeddings derived by the emotion encoder form separate groups for each emotion type, while those from the style encoder fail to provide a clear pattern. From Figure 1, we also observe a significant separation between the emotions with lower values of arousal and valence such as neutral and sad, and those with higher values such as angry, happy and surprise. These observations further validate our 2-stage training for EVC.

3.2.2. رمزگذار سبک در مقابل رمزگذار احساسات

در طول آموزش احساسات، رمزگذار سبک به عنوان رمزگذار احساسات عمل می کند و شناسه احساس را به عنوان ورودی برای یادگیری مؤثر بازنمایی احساسات در گفتار می گیرد. برای تایید ایده خود، از t-SNE استفاده می کنیم [34] برای تجسم تعبیه احساسات عبارات مرجع، که به ترتیب توسط رمزگذار سبک از مرحله I و رمزگذار احساسات از مرحله II مشتق شده اند. برای خوشحالی ما، همانطور که در شکل 1 نشان داده شده است، جاسازی های احساسات به دست آمده توسط رمزگذار احساسات گروه های جداگانه ای را برای هر نوع احساس تشکیل می دهند، در حالی که آنهایی که از رمزگذار سبک در ارائه یک الگوی واضح ناکام هستند. . از شکل 1، ما همچنین جدایی قابل توجهی را بین احساسات با ارزش‌های برانگیختگی و ظرفیت پایین‌تر مانند خنثی و غمگین، و احساسات با ارزش‌های بالاتر مانند عصبانیت، خوشحالی و تعجب مشاهده می‌کنیم. این مشاهدات آموزش 2 مرحله‌ای ما را برای EVC تأیید می‌کند.

3.3. Run-time inference

At run-time, we use the emotion encoder to generate the emotion embeddings from a set of reference utterances belonging to the same emotion category. We use the average emotion embedding to represent the emotion style. Given a source utterance and the intended emotion category, we use a seq2seq ASR encoder to derive the linguistic embedding of the source utterance, and apply the respective emotion embedding to the decoder. The converted acoustic features can be reconstructed by the seq2seq decoder

3.3.استنتاج زمان اجرا

در زمان اجرا، ما از رمزگذار احساسات برای ایجاد تعبیه‌های احساسات از مجموعه‌ای از گفته‌های مرجع استفاده می‌کنیم.

به همان دسته احساسات ما از احساسات متوسط استفاده می کنیم

تعبیه برای نشان دادن سبک احساسات با ذکر منبع بیان و مقوله احساس مورد نظر، از seq2seq استفاده می کنیم

رمزگذار ASR برای استخراج تعبیه زبانی منبع بیان، و تعبیه احساسات مربوطه را در آن اعمال کنید

رمزگشا ویژگی های صوتی تبدیل شده را می توان بازسازی کرد توسط رسیور seq2seq

3.4. Comparison with related work

The proposed seq2seq EVC framework shares a similar motivation with [28, 30] in terms of leveraging TTS but differs in many aspects. To start with, [28] only focuses on speaker disentanglement, and emotion has not been considered. The proposed 2-stage training strategy allows the network to learn emotion style in training stage II, thus, requires a smaller amount of emotional speech data. Compared with [30] which needs more than 30 hours of emotional speech data for training, our proposed framework only uses less than 50 minutes of emotional speech data. Besides, we further employ adversarial training with an emotion classifier to learn a better emotion disentanglement and use a seq2seq ASR with an explicit loss between the linguistic embedding of EVC and TTS to get a better alignment.

3.4. مقایسه با کار مرتبط

چارچوب پیشنهادی seq2seq EVC انگیزه مشابهی با [28، 30] از نظر استفاده از TTS دارد اما در بسیاری از جنبه‌ها متفاوت است. برای شروع، [28] فقط بر گسستگی گوینده تمرکز دارد و احساسات در نظر گرفته نشده است. استراتژی آموزش 2 مرحله‌ای پیشنهادی به شبکه اجازه می‌دهد تا سبک هیجانی را در مرحله دوم آموزش بیاموزد، بنابراین، به مقدار کمتری از داده‌های گفتاری احساسی نیاز دارد. در مقایسه با [30] که به بیش از 30 ساعت داده گفتار عاطفی برای آموزش نیاز دارد، چارچوب پیشنهادی ما تنها از کمتر از 50 دقیقه داده گفتاری احساسی استفاده می‌کند. علاوه بر این، ما از آموزش خصمانه با طبقه‌بندی‌کننده احساسات برای یادگیری بهتر تفکیک احساسات استفاده می‌کنیم و از seq2seq ASR با از دست دادن صریح بین تعبیه زبانی EVC و TTS استفاده می‌کنیم تا تراز بهتری داشته باشیم.

4. Experiments

We conduct emotion conversion from neutral to angry, sad, happy and surprise, denoted as Neu-Ang, Neu-Sad, Neu-Hap, and Neu-Sur respectively. We first use VCTK corpus [35] for stage I as shown in Figure 2(a), and then use the ESD database [36] for stage II as shown in Figure 2(b). For each emotion pair, we use 300 utterances for training, 30 utterances for reference, and 20 utterances for evaluation. The total duration of emotional speech data used in the stage II is around 50 minutes, which is small in the context of seq2seq training. The codes and implementation details of this work are publicly available at: https://github.com/KunZhou 9646/seq2seq-EVC. We implement two state-of-the-art methods, together with 4 seq2seq EVC systems:

4. آزمایشات

ما تبدیل احساسات را از حالت خنثی به عصبانی، غمگین، شاد و غافلگیرکننده انجام می دهیم که به ترتیب به عنوان Neu-Ang، Neu-Sad، Neu-Hap و Neu-Sur نشان داده می شوند. ما ابتدا از بدنه VCTK [35] برای مرحله I همانطور که در شکل 2(a) نشان داده شده است، و سپس از پایگاه داده ESD [36] برای مرحله II همانطور که در شکل 2(b) نشان داده شده است استفاده می

کنیم. برای هر جفت احساس، از 300 گفته برای آموزش، 30 گفته برای مرجع و 20 گفته برای ارزیابی استفاده می کنیم. کل مدت زمان داده های گفتار احساسی مورد استفاده در مرحله دوم حدود 50 دقیقه است که در زمینه آموزش seq2seq اندک است. کدها و جزئیات پیاده سازی این کار به صورت عمومی در دسترس هستند: https://github.com/KunZhou 9646/seq2seq-EVC.

ما دو روش پیشرفته را به همراه 4 سیستم EVC seq2seq پیاده سازی می کنیم:

• CycleGAN-EVC [15] (baseline): CycleGAN-based emotional voice conversion with WORLD vocoder.

• StarGAN-EVC [16] (baseline): StarGAN-based emotional voice conversion with WORLD vocoder.

• Baseline Seq2seq-EVC (baseline): Seq2seq-EVC trained directly with limited ESD data without any pretraining, and followed by a Griffin-Lim vocoder [37];

• Seq2seq-EVC-GL (proposed): Seq2seq-EVC followed by a Griffin-Lim vocoder;

• Seq2seq-EVC-WA1 (proposed): Seq2seq-EVC followed by a WaveRNN vocoder [31] that is pre-trained on VCTK corpus;

• Seq2seq-EVC-WA2 (proposed): Seq2seq-EVC followed by a WaveRNN vocoder that is pre-trained on VCTK corpus, and fine-tuned with limited ESD data.

) [15] CycleGAN-EVC •پایه): مبتنی برCycleGAN

تبدیل صدای احساسی با.Vocoder WORLD

) [16] StarGAN-EVC ·پایه): تبدیل صدای احساسی مبتنی بر StarGAN با.Vocoder WORLD

· Baseline Seq2seq-EVC (Baseline): Seq2seq-EVC

مستقیماً با داده‌های محدود ESD بدون هیچ گونه پیش‌آموزشی آموزش دیده و به دنبال آن یک.[37] Vocoder Griffin-Lim

) Seq2seq-EVC-GL ·پیشنهاد شده Seq2seq-EVC :(دنبال شد

توسط یک.Vocoder Griffin-Lim

) Seq2seq-EVC-WA1 ·پیشنهاد شده Seq2seq-EVC :(و پس از آن یک [31] Vocoder WaveRNN که از قبل آموزش داده شده است

در مجموعهVCTK ؛

) Seq2seq-EVC-WA2 ·پیشنهاد شده Seq2seq-EVC :(و پس از آن یک Vocoder WaveRNN که از قبل آموزش داده شده است

مجموعه VCTK، و با داده های محدود ESD تنظیم شده است.

We note that CycleGAN-EVC only can perform the oneto-one conversion, thus we train one CycleGAN-EVC for each emotion pair separately. Both StarGAN-EVC and our proposed Seq2seq-EVC use a unified model for all the emotion pairs.

توجه داریم که CycleGAN-EVC فقط می تواند تبدیل یک به یک را انجام دهد، بنابراین ما یک CycleGAN-EVC را برای هر یک آموزش می دهیم.

جفت احساس به صورت جداگانه هم StarGAN-EVC و هم پیشنهادی ماSeq2seq-EVC از یک مدل یکپارچه برای همه جفت های احساسات استفاده می کند.

4.1. Objective Evaluation

We calculate Mel-cepstral distortion (MCD) [5] and the average absolute differences of the utterance duration (DDUR) [28]

4.1.ارزیابی عینی

ما اعوجاج Mel-cepstral (MCD) [5] و میانگین را محاسبه می کنیم

تفاوت مطلق مدت زمان بیان (DDUR) [28]

صفحه چهارم

شکل بالای صفحه4

Figure 3: Acoustic-linguistic alignment visualization of utterances that are converted from neutral to (a) happy and (b) sad. for the voiced parts to measure the spectral distortion and duration difference respectively. In Seq2seq-EVC models, Melspectrograms are adopted as acoustic features, and the Melcepstral coefficients (MCEPs) are extracted directly from the waveform to calculate MCD values.

شکل 3: تجسم تراز صوتی-زبانی جملاتی که از حالت خنثی به (الف) شاد و (ب) غمگین تبدیل می شوند.

برای قطعات صدا برای اندازه گیری اعوجاج طیفی و

به ترتیب اختلاف مدت در مدل‌هایSeq2seq-EVC ، Melspectrogramها به عنوان ویژگی‌های صوتی پذیرفته می‌شوند و ضرایب Melcepstral (MCEPs) مستقیماً از

شکل موج برای محاسبه مقادیر MCD.

To motivate our proposed 2-stage training, we first conduct experiments with the Baseline Seq2seq-EVC model that is trained directly with limited ESD data without any pre-training procedures. We note that in all experiments, it consistently achieves the worst results in terms of MCD and DDUR values. This observation shows that seq2seq-EVC does not work well with limited training data, which further shows the necessity of our proposed 2-stage training strategy.

برای ایجاد انگیزه برای آموزش 2 مرحله ای پیشنهادی ما، ابتدا انجام می دهیم

آزمایشات با مدل پایه Seq2seq-EVC که است

به طور مستقیم با داده های محدود ESD بدون هیچ گونه پیش آموزشی آموزش دیده است

رویه ها توجه داشته باشیم که در تمام آزمایشات، به طور مداوم

بدترین نتایج را از نظر مقادیر MCD و DDUR به دست می آورد.

این مشاهدات نشان می دهد که seq2seq-EVC به خوبی کار نمی کند

با داده های آموزشی محدود، که بیشتر ضرورت را نشان می دهد

استراتژی آموزشی 2 مرحله ای پیشنهادی ما

As shown in Table 1, our proposed Seq2seq-EVC models with Griffin-Lim, WaveRNN and fine-tuned WaveRNN always outperform baseline CycleGAN-EVC and StarGAN-EVC. We also note that the proposed Seq2seq-EVC-WA1 and Seq2seqEVC-WA2 consistently achieve the best results of MCD values for all the emotion pairs, which shows the effectiveness of our proposed 2-stage training strategy for limited data EVC

همانطور که در جدول 1 نشان داده شده است، مدل های Seq2seq-EVC پیشنهادی ما

همیشه باGriffin-Lim ، WaveRNNو WaveRNN دقیق تنظیم شده است

بهتر از CycleGAN-EVC و StarGAN-EVC پایه. ما

همچنین توجه داشته باشید که Seq2seq-EVC-WA1 و Seq2seqEVC-WA2 پیشنهادی به طور مداوم بهترین نتایج مقادیر MCD را به دست می‌آورند.

برای همه جفت های احساسات، که اثربخشی ما را نشان می دهد

استراتژی آموزشی 2 مرحله ای برای EVC داده های محدود پیشنهاد شده است

We note that the attention mechanism allows us to vary the phonetic duration from source to target during the decoding, which is crucial for EVC. Figure 3 shows an example of the acoustic-linguistic alignment for (a) converted happy and (b) converted sad utterances. We note that the source utterance is the same for both conversion mappings, while the converted utterances have different duration. These results further show that our proposed framework is capable of duration manipulation

توجه می کنیم که مکانیسم توجه به ما اجازه می دهد تا آن را تغییر دهیم

مدت زمان آوایی از منبع به مقصد در طول رمزگشایی،

که برای EVC بسیار مهم است. شکل 3 نمونه ای از آن را نشان می دهد

تراز صوتی-زبانی برای (الف) تبدیل شاد و (ب(

تبدیل جملات غم انگیز توجه می کنیم که قول منبع است

برای هر دو نگاشت تبدیل یکسان است، در حالی که گفته های تبدیل شده مدت زمان متفاوتی دارند. این نتایج بیشتر نشان می دهد که

چارچوب پیشنهادی ما قادر به دستکاری مدت زمان است

We further report DDUR results to evaluate duration conversion performance in Table 2. We observe that the proposed Seq2seq-EVC-WA1 (with WaveRNN) and Seq2seq-EVC-WA2 (with fine-tuned

WaveRNN) consistently achieves the best DDUR results for all emotion pairs. We noted that baseline frameworks CycleGAN-EVC and StarGAN-EVC cannot modify speech duration, hence they are not reported in the table. These results show the effectiveness of our proposed Seq2seqEVC framework in terms of duration conversion.

ما بیشتر نتایج DDUR را برای ارزیابی عملکرد تبدیل مدت زمان در جدول 2 گزارش می‌کنیم. مشاهده می‌کنیم که پیشنهاد شده است

) Seq2seq-EVC-WA1با (WaveRNN وSeq2seq-EVC-WA2

)با WaveRNN دقیق تنظیم شده) به طور مداوم بهترین ها را به دست می آورد

نتایج DDUR برای همه جفت های احساسات. ما به آن خط مبنا اشاره کردیم چارچوب‌های CycleGAN-EVC و StarGAN-EVC نمی‌توانند مدت زمان گفتار را تغییر دهند، بنابراین در جدول گزارش نشده‌اند.

این نتایج اثربخشی چارچوب Seq2seqEVC پیشنهادی ما را از نظر تبدیل مدت زمان نشان می‌دهد.

4.2. Subjective Evaluation We conduct listening tests to assess the emotion similarity and speech quality. 15 subjects participated in all the experiments and each listened to 128 converted utterances in total. We first report the emotion similarity results as shown in Figure 4. We use the baseline frameworks CycleGAN-EVC and StarGAN-EVC; and the proposed framework Seq2seq-EVC with Griffin-Lim (Seq2seq-EVC-GL), WaveRNN (Seq2seqEVC-WA1), and fine-tuned WaveRNN (Seq2seq-EVC-WA2).

4.2.ارزیابی ذهنی

ما تست های شنیداری را برای ارزیابی شباهت احساسات وکیفیت گفتار 15 نفر در تمام آزمایش ها شرکت کردند و هر کدام در مجموع به 128 گفته تبدیل شده گوش دادند.

ما ابتدا نتایج شباهت عاطفی را همانطور که در نشان داده شده است گزارش می کنیم

شکل 4. ما از چارچوب های پایه CycleGAN-EVC و استفاده می کنیم

StarGAN-EVC;وچارچوب پیشنهادیSeq2seq-EVC

با (Seq2seq-EVC-GL) Griffin-Lim، (WaveRNN (Seq2seqEVC-WA1)، و WaveRNN دقیق تنظیم شده (Seq2seq-EVC-WA2).

شکل بالای سمت راست صفحه4

Figure 4: Emotional similarity results with 95% confidence interval to evaluate emotion similarity with target speech in a scale of -2 to 2 (-2: absolutely different; -1: different; 0: cannot tell; +1: similar; +2: absolutely similar)

شکل 4: نتایج شباهت عاطفی با 95% اطمینان

فاصله برای ارزیابی شباهت عاطفی با گفتار هدف در الف

مقیاس -2 تا 2 (-2: کاملاً متفاوت؛ -1: متفاوت؛ 0: نمی تواند بگوید +1: مشابه +2: کاملا مشابه)

Table 3: Best Worst Scaling (BWS) listening experiments to evaluate the overall speech quality

جدول 3: آزمایش‌های شنیداری بهترین بدترین مقیاس‌بندی.(BWS) کیفیت کلی گفتار را ارزیابی کنید

All participants are asked to listen to the reference target speech first, and then score the speech samples in terms of the emotion similarity to the reference target speech. It is encouraging to see that the proposed Seq2seq-EVC framework with WaveRNN (Seq2seq-EVC-WA1) and fine-tuned WaveRNN (Seq2seq-EVC-WA2) significantly outperform the baselines for all the emotion pairs, especially for Neu-Sur

از همه شرکت‌کنندگان خواسته می‌شود ابتدا به گفتار هدف مرجع گوش دهند و سپس نمونه‌های گفتار را از نظر شباهت احساسی به گفتار هدف مرجع امتیاز دهند.

مایه دلگرمی است که ببینیم چارچوب پیشنهادی Seq2seq-EVC با WaveRNN (Seq2seq-EVC-WA1) و WaveRNN دقیق تنظیم شده (Seq2seq-EVC-WA2) به طور قابل توجهی از خطوط پایه برای همه جفت های احساسات، به ویژه برای Neu-Sur بهتر است.

We further conduct the best-worst scaling (BWS) [38] test in terms of speech quality of our proposed Seq2seq-EVC framework with 1) Griffin-Lim (Seq2seq-EVC-GL), 2) WaveRNN (Seq2seq-EVC-WA1), and 3) fine-tuned WaveRNN (Seq2seqEVC-WA2). All participants are asked to choose the best one and the worst one in terms of the overall quality. From Table 3, Seq2seq-EVC-WA2 outperforms the baseline consistently, which proves the effectiveness of our emotional fine-tuning strategy on WaveRNN vocoder

ما همچنین آزمون بهترین-بدترین مقیاس بندی [38] (BWS) را انجام می دهیم

از نظر کیفیت گفتار چارچوب Seq2seq-EVC پیشنهادی ما با 1(Seq2seq-EVC-GL) Griffin-Lim (، WaveRNN (2

(Seq2seq-EVC-WA1)، و 3 WaveRNN (با تنظیم دقیق .(Seq2seqEVC-WA2) از همه شرکت کنندگان خواسته می شود که بهترین را انتخاب کنند و از نظر کیفیت کلی بدترین. از جدول

3، Seq2seq-EVC-WA2به طور مداوم از خط پایه بهتر عمل می کند،که اثربخشی تنظیم دقیق عاطفی ما را ثابت می کند

استراتژی در Vocoder WaveRNN

5. Conclusion In this paper, we propose a novel training strategy for seq2seq emotional voice conversion leveraging text-to-speech without the need for parallel data. To our best knowledge, this is the first work of seq2seq emotional voice conversion that only needs a limited amount of emotional speech data to train. Moreover, the proposed framework can do many-to-many emotional voice conversion, and conduct spectral and duration mapping at the same time. We also investigate the training strategy of emotion fine-tuning for WaveRNN vocoder training. Experimental results show a significant improvement of the conversion performance over the baselines.

.5نتیجه گیری

در این مقاله، ما یک استراتژی آموزشی جدید برای seq2seq پیشنهاد می‌کنیم

تبدیل صدای احساسی با استفاده از متن به گفتار بدون نیاز به داده های موازی تا جایی که ما می دانیم، این اولین مورد است

کار تبدیل صدای احساسی seq2seq که فقط نیاز به یک مقدار محدودی از داده های گفتار عاطفی برای آموزش. علاوه بر این،

چارچوب پیشنهادی می تواند صدای عاطفی بسیار به چند نفر را انجام دهد تبدیل، و انجام نقشه برداری طیفی و مدت زمان درهمان زمان. ما همچنین استراتژی آموزشی تنظیم دقیق احساسات را برای آموزش رمزگذار صوتی WaveRNN بررسی می‌کنیم. تجربی نتایج نشان دهنده بهبود قابل توجهی در عملکرد تبدیل نسبت به خطوط پایه است.

6. Acknowledgment The research is funded by SUTD Start-up Grant Artificial Intelligence for Human Voice Conversion (SRG ISTD 2020 158) and SUTD AI Grant - Thrust 2 Discovery by AI (SGPAIRS1821), the National Research Foundation, Singapore under its AI Singapore Programme (Award No: AISG-GC2019-002) and (Award No: AISG-100E-2018-006), and its National Robotics Programme (Grant No. 192 25 00054), and by RIE2020 Advanced Manufacturing and Engineering Programmatic Grants A1687b0033, and A18A2b0046.

.6تصدیق

بودجه این تحقیق توسط SUTD Start-up Grant Artificial تامین می شود

هوش برای تبدیل صدای انسانSRG ISTD 2020)

(158و SUTD AI Grant - Thrust 2 Discovery توسطAI (SGPAIRS1821) ، بنیاد تحقیقات ملی، سنگاپور

تحت برنامه هوش مصنوعی سنگاپور (شماره جایزه AISG-GC2019-002) :و (شماره جایزه AISG-100E-2018-006) :و آن

برنامه ملی رباتیک (گرنت شماره 192 25 00054)،

و توسطRIE2020 Advanced Manufacturing and Engineering

کمک های مالی برنامه ای A1687b0033 و.A18A2b0046