بسم اللّه الرحمن الرحیم

دانشگاه صنعتی شریف
آزمایشگاه پردازش گفتار

طراحی یک سامانه‌ی جامع تبدیل متن به گفتار فارسی

دفاع کارشناسی ارشد

حامد جمشیدیان
استاد راهنما: دکتر حسین صامتی
استاد داور: دکتر مهدیه سلیمانی

شهریور ۱۴۰۰

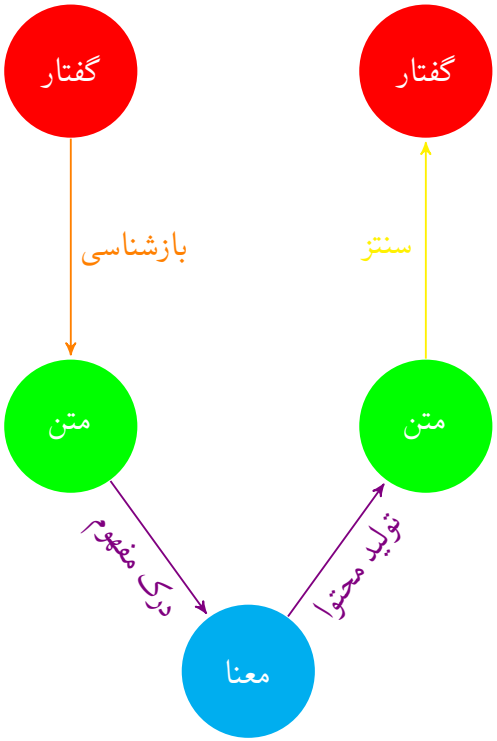
سیستم‌های قادر به تکلم

- گفتار به عنوان یکی از روش‌های ارتباطی انسان‌ها
- استفاده از گفتار در بهبود تعاملات بین انسان و کامپیوتر
- طراحی سیستمی که قادر به تکلم بوده و انسان قادر به تشخیص مصنوعی بودن این سیستم نباشد از مسائل مطرح در هوش‌مصنوعی است.

تقسیم‌بندی وظایف سیستم قادر به مکالمه

برای ایجاد یک سامانه‌ی قادر به صحبت با انسان باید مسئله را تقسیم کنیم

- گفتار به متن
- تبدیل متن به معنا
- تبدیل معنا به متن
- متن به گفتار



شکل: تقسیم‌بندی وظایف یک سیستم کامپیوتری با توانایی درک و تولید گفتار برای ارتباط با کاربر انسانی

اهداف سامانه‌های متن به گفتار

- تبدیل دنباله‌ای از کلمات معنادار در زبانی خاص به صوت حاوی خوانش آن دنباله
- تا حد امکان طبیعی و رسا
- تک گوینده، چند گوینده یا با گوینده‌ی تطبیق‌پذیر
- تولید صدا با سرعت بالا
- استفاده‌ی بهینه از سخت افزار

کاربردهای متن به گفتار

- پیاده‌سازی سیستم هوشمند با توانایی ایجاد ارتباط گفتاری
- کمک به نابینایان یا افرادی که توانایی خواندن ندارند
- کمک به افراد با مشکل تکلم
- استفاده در سیستم‌های نوبت‌دهی یا اعلام هشدار
- صداگذاری و دوبله‌ی ویدیوهای مختلف

روش‌های پیش از یادگیری ژرف

نسل‌های اولیه

سنتز بر پایه‌ی فرمنت
سنتز بر پایه‌ی اجزای تولید صدا
سنتز الحاقی

نسل‌های بعدی

روش‌های مبتنی بر انتخاب واحد
مدل مخفی مارکوف

مشکلات روش‌های قبلی

- سنتز فرمنت: مدت زمان ادای واحدها و تعیین مقادیر فیلترها
- سنتز مبتنی بر اجزای تولید صدا: اختلاف خروجی با صدای طبیعی انسان
- سنتز الحاقی: مشکل جمع آوری واحدهای الحاقی
- سنتز مبتنی بر انتخاب واحد: تشدید مشکل سنتز الحاقی
- مدل مخفی مارکوف: پایین بودن کیفیت صدا و عدم توانایی مدل در تعمیم‌پذیری

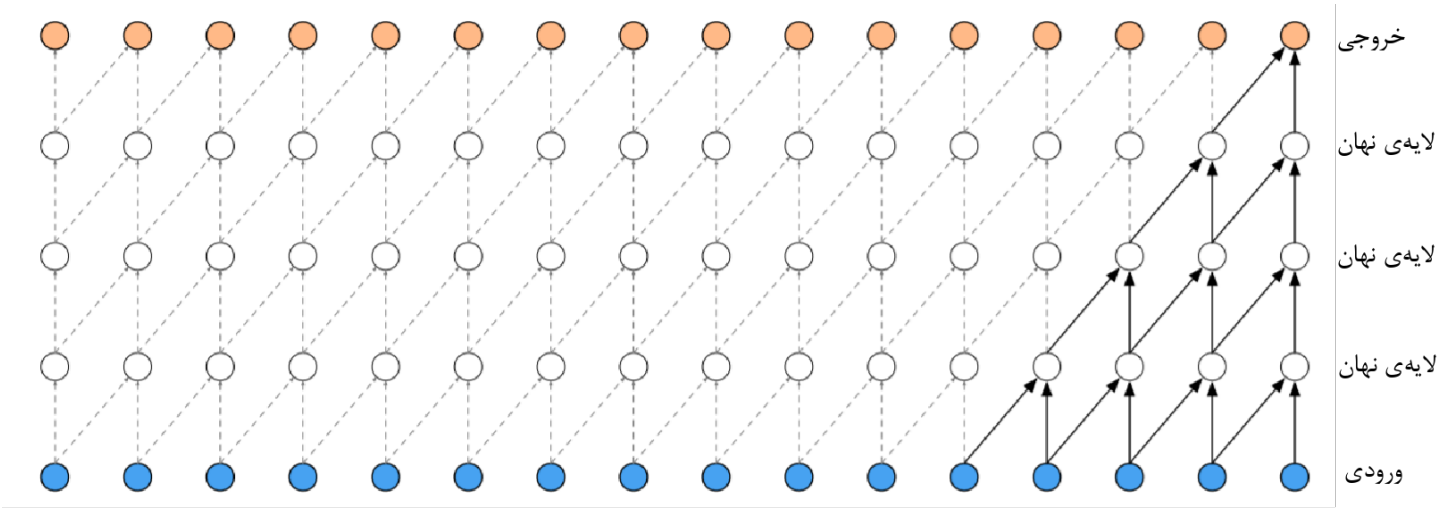
به علت مشکلات موجود در روش‌های قبلی و رواج استفاده از روش‌های مبتنی بر یادگیری ژرف در این مسئله نیز استفاده از این گونه راه حل‌ها مطرح شد

مدل WaveNet

- یک شبکه‌ی خودبازگشتی مولد
احتمال توأم توالی $\mathbf{x} = \{x_1, ..., x_T\}$ با

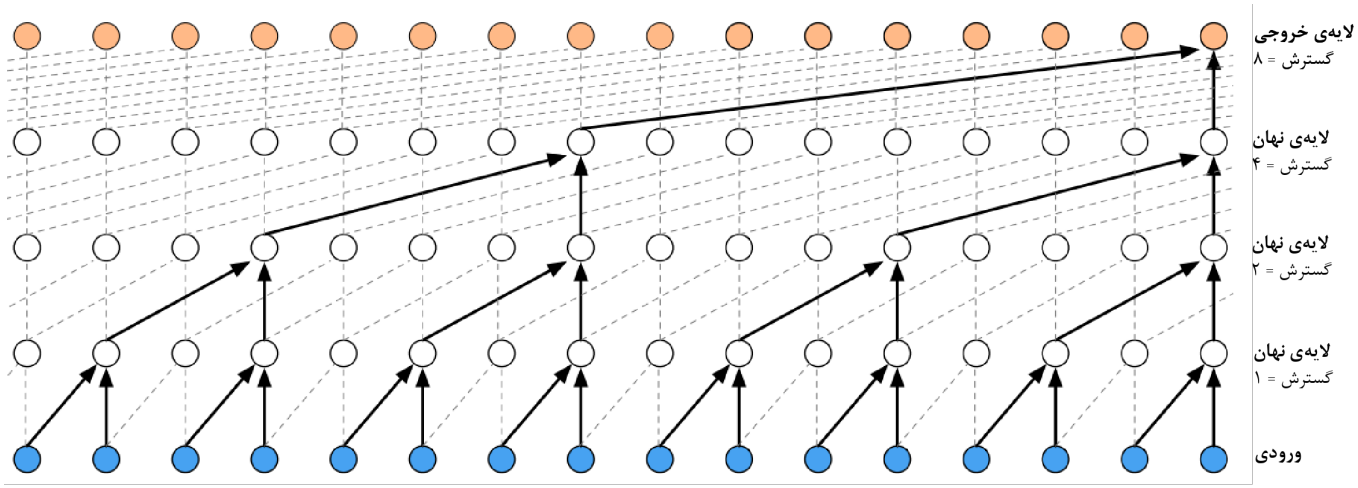
$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_1, ..., x_{t-1})$$

مدل شده است
- احتمال‌های مشروط در خروجی بر اساس پشته‌ای از لایه‌های پیچشی مدل می‌شود
- خروجی در گام t توزیع دسته‌ای برای مقدار x_t است
- استفاده از شبکه‌ی عِلّی برای حفظ محدودیت سیگنال صوتی



شکل: نمونه‌ای از پشته‌ی لایه‌های پیچشی علّی

مشکلی که این حالت از معماری شبکه داشت کوچک بودن سایز ناحیه‌ی ادراکی در هر گام زمانی بود



شکل: نمونه‌ای از پشته‌ی لایه‌های پیچشی علّی گسترش‌یافته

در هر لایه با پرش از تعدادی از ورودی‌ها ناحیه‌ی ادراکی وسیعتر می‌شود این نوع از لایه‌های شبکه به پیچشی عِلّی گسترش‌یافته موسوم است

- ورودی شبکه ویژگی‌های زبانی استخراج شده از متن و خروجی توزیع دسته‌ای روی نمونه‌های سیگنال صوت است
- آموزش مدل با دو دادگان تک گوینده در انگلیسی و چینی با طول ۲۴٫۶ و ۳۴٫۸ ساعت
- ارزیابی با معیار MOS

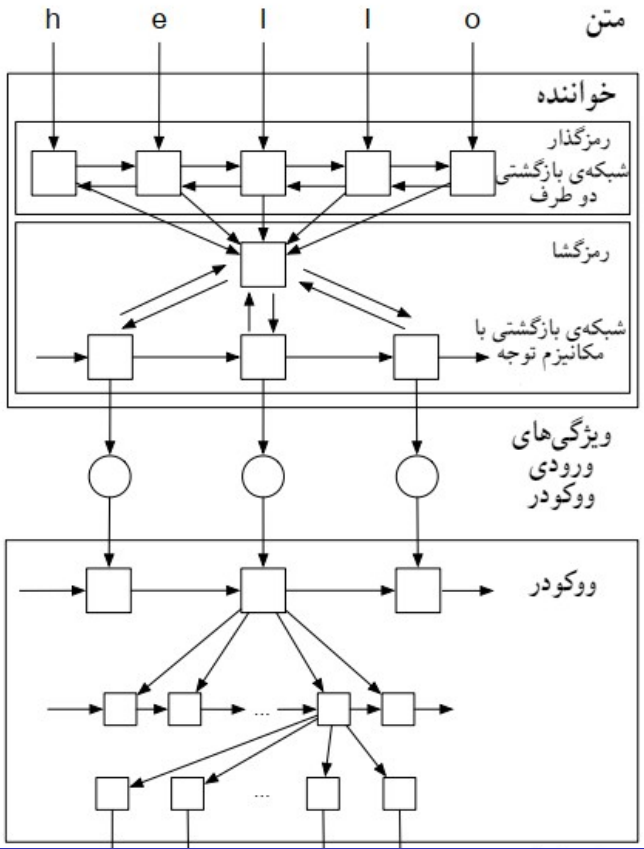
مدل	ارزیابی شاخص میانگین با مقیاس ۵	
	انگلیسی	چینی
LSTM-RNN پارامتری	۳٫۶۷	۳٫۷۹
مبتنی بر مدل مخفی مارکوف و انتخاب واحد	۳٫۸۶	۳٫۴۷
WaveNet	۴٫۲۱	۴٫۰۸

جدول: ارزیابی مدل WaveNet ارائه شده در مقاله‌ی ارائه‌دهنده‌ی مدل

مدل Char2Wav

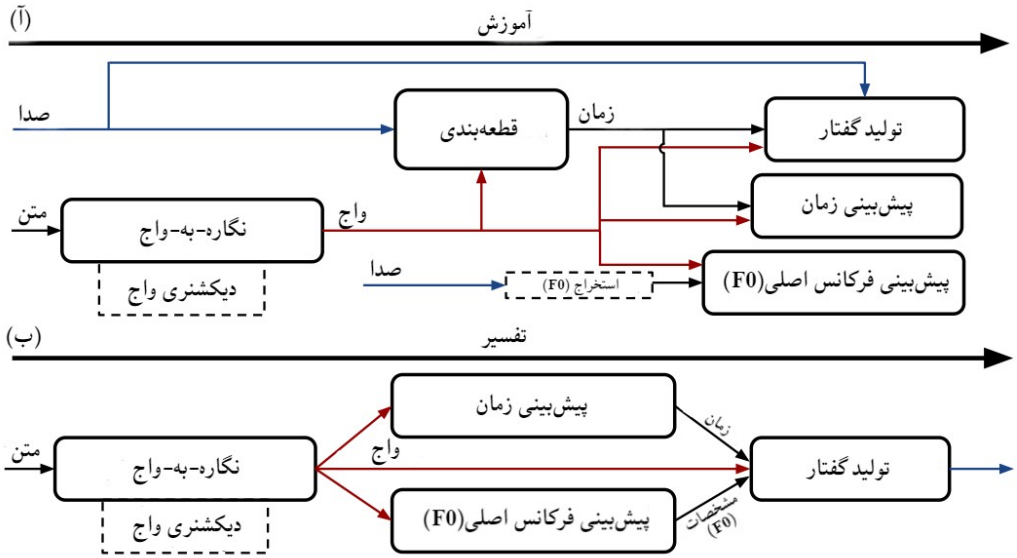
- مدل انتها-به-انتها مبتنی بر معماری رمزگذار-رمزگشا و مکانیزم توجه
- با فرض داشتن دنباله‌ی جاسازی حروف X دنباله‌ی $h = (h_1, ..., h_L)$ توسط رمزگذار تولید می‌شود و سپس دنباله‌ی ویژگی‌های ووکودر $Y = (y_1, ..., y_t)$ توسط رمزگشا و با توجه به مکانیزم توجه به صورت زیر به دست می‌آید:

$$\alpha_i = \text{Attend}(s_{i-1}, \alpha_{i-1}, h)$$
$$g_i = \sum_{j=1}^{L} \alpha_{i,j} h_j$$
$$y_i \sim \text{Generate}(s_{i-1}, g_i)$$
$$s_i = \text{RNN}(s_{i-1}, g_i, y_i)$$



مدل DeepVoice

- مدل انتها-به-انتها برای تولید نمونه‌های صوت
- تقسیم مسأله به زیر مسئله‌های زیر
  - نگاره-به-واج
  - قطعه‌بندی
  - پیش‌بینی زمان
  - پیش‌بینی فرکانس اصلی
  - تولید گفتار



- مدل با دادگانی به طول ۲۰ ساعت و متشکل از زوج‌های متن-صوت به تعداد ۱۳۰۷۹ آموزش داده شده است

طرح مسئله

- در مدل‌های قبل از شبکه‌های عمیق برای زبان فارسی حالت رباتی بودن صدا احساس می‌شد
- مدل‌های مبتنی بر یادگیری ژرف نیازمند داده‌ی آموزشی با حجم بالا هستند
- این دادگان باید ویژگی‌های خاصی را برآورده کند
  - تک گوینده
  - بدون نویز
  - عاری از کلمات شکسته و محاوره
  - پوشش کلمات متنوع از زبان
  - رعایت نوا و استرس در تلفظ عبارات
- در زبان فارسی داده‌ی با حجم بالا به صورت آزاد موجود نیست
- پیچیدگی‌های زبان فارسی مانند حروفی که در بعضی مواقع خوانده نمی‌شوند: «خواهر»، «موسی»

کلیت روش

- استفاده از دو قسمت انتهای پیشین و انتهای جلودار
- انتهای پسین: آماده‌سازی دنباله‌ی کاراکتر ورودی برای تبدیل به شکل موج صوت.
  شامل واحدهای:
  - هنجارساز
  - نگاره به واج
  - اضافه‌کننده‌ی استرس

- انتهای جلودار: تبدیل دنباله‌ی کاراکتر ورودی به شکل موج صوت. شامل واحدهای:
  - تکاترون
  - مِل‌گن

هنجارساز

تبدیل ناهنجاری‌های موجود در متون با نگارش‌های متفاوت به نگارشی مطابق با انتهای جلودار

- جایگذاری حروف تعریف نشده در دادگان با کاراکتر مناسب: «ي» ← «ی»، «ۀ» ← «ه» و ...
- حذف فاصله‌های اضافی بین کلمات و علائم نگارشی: «به کار بردن این قانون الزامی می‌باشد.» ← «به کار بردن این قانون الزامی می‌باشد.»
- تصحیح اشتباهات ناشی از استفاده یا عدم استفاده‌ی نیم‌فاصله: «میباشد» ← «می‌باشد»،
- تبدیل تاریخ و زمان به معادل نوشتاری با کاراکترهای تعریف شده‌ی فارسی: «۱۳۵۷/۱۱/۲۲» ← «بیست و دو بهمن هزار و سیصد و پنجاه و هفت»، «۱۳ : ۵۵» ← «سیزده ساعت و پنجاه و پنج دقیقه»

نگاره به واج

تبدیل دنباله‌ی کاراکترهای نوشته شده در زبانی خاص به دنباله‌ی واجی متناظر

«پرداخت این وام از طرف صندوق وام عمرانی آمریکا اعلام شد»

↓

p/rdaxte @in vam | @/z t/r/fe s/nduqe vame @omraniye @amrika | @e@lam $od

- شبکه‌ی تشخیص کسره‌ی اضافه
- شبکه‌ی تشخیص کلمات هم‌نگاره
- شبکه‌ی پیش‌بینی تلفظ کلمات خارج از لغت‌نامه

اضافه‌کننده‌ی استرس

- در زبان فارسی در حالت معمول برای تلفظ هر کلمه روی یکی از مصوت‌های آن کلمه تاکید می‌شود
- خطا در تلفظ استرس در خروجی مدل سنتز گفتار
- کمک به مدل با نشان‌دار کردن استرس کلمات:

«p/rdaxte @in vam | @/z t/r/fe s/nduqe vame @omraniye @amrika | @e@lam $od»

↓

«p/r/daxtE @In vAm | @%z t/r/fE s/ndoqE vamE @omraniniyE @amrikA | @e@lAm $Od»

- در زبان فارسی می‌توان با استفاده از قوانین موجود استرس کلمات را مشخص کرد

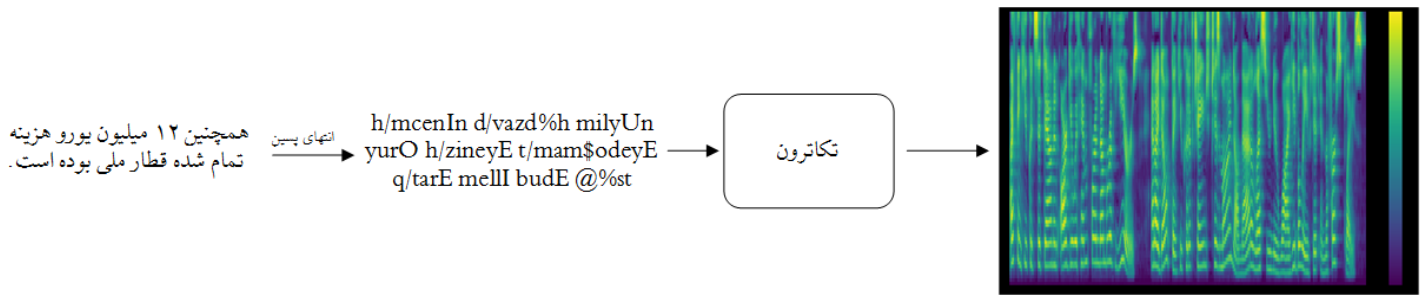
قوانین اضافه کردن استرس

- استرس روی تنها مصوت در کلمات تک‌هجایی: «سم» ← «s/m» ← «s%m»
- استرس روی هجای قبل از کسره‌ی اضافه: «توپ والیبال» ← «tupe valibal» ← «tUpe valibAl»
- فعل در جملات بدون استرس تلفظ می‌شود: «می‌رود» ← «mir/v/d»
- در کلمات با ضمایر ملکی متصل استرس روی هجای قبل از ضمیر ملکی می‌آید: «حواسشان» ← «h/vase$an» ← «h/vasE$an»
- در لیست محدودی از کلمات استرس روی هجای اول کلمه قرار می‌گیرد: «بلکه» ← «b/lke» ← «b%lke»
- در صورتی که کلمه‌ای جزو قوانین قبلی نبود استرس روی هجای آخر کلمه قرار می‌گیرد

مدل سنتز گفتار

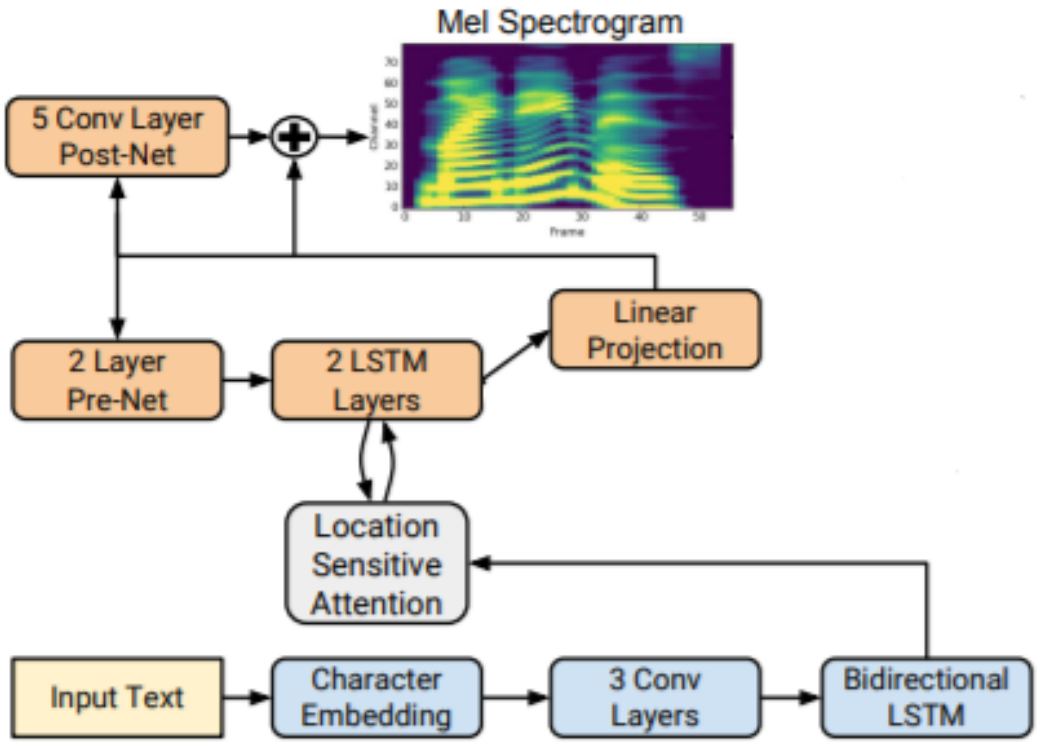
بخش اصلی سامانه‌ی تبدیل متن به گفتار که در ویژگی‌های استخراج شده از متن را به نوعی بازنمایی از صوت تبدیل می‌کند

- استفاده از مدل تکاترون ۲
- ورودی دنباله‌ی واجی آماده شده در انتهای پیشین
- خروجی طیف‌نگاشت مل



شکل: نمایی از شکل ورودی و خروجی واحد تکاترون در روش پیشنهادی

مدل تکاترون ۲



شکل: معماری شبکه‌ی تکاترون ۲

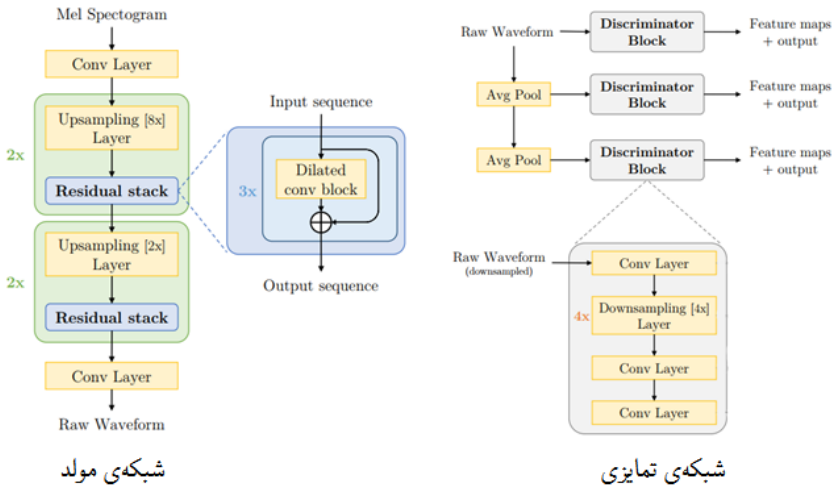
- از مدل‌هایی است که اخیراً ارائه شده
- در حال حاضر حجم زیادی از تحقیقات در حوزه‌ی مسائل متن به گفتار با تغییر روی این مدل انجام می‌شود
- معماری رمزگذار-رمزگشا
- مکانیزم توجه حساس به مکان
- ورودی جاسازی حروف و خروجی طیف‌نگاشت مل

وُکودر

- استفاده از طیف‌نگاشت به علت وجود نمونه‌های خروجی کمتر
- کیفیت بهتر صدای خروجی بعد از استفاده از طیف‌نگاشت برای تبدیل کاراکتر
- حذف اطلاعات فاز سیگنال در طیف‌نگاشت
- نیاز به الگوریتم یا مدل برای تخمین نمونه‌های شکل موج صوت
  - گریفین.لیم
  - ویونت
  - **ملگن**

مل‌گن

- مدل بر پایه‌ی شبکه‌ی مولد تخاصمی
- استفاده از سه واحد تمایزی روی مقیاس‌های متفاوت از نمونه‌های صوت
- شبکه‌ی مولد نسبتاً ساده، تماماً بر پایه‌ی لایه‌های پیچشی معماری رمزگذار–رمزگشا
- مکانیزم توجه حساس به مکان



شبکه‌ی مولد　　شبکه‌ی تمایزی

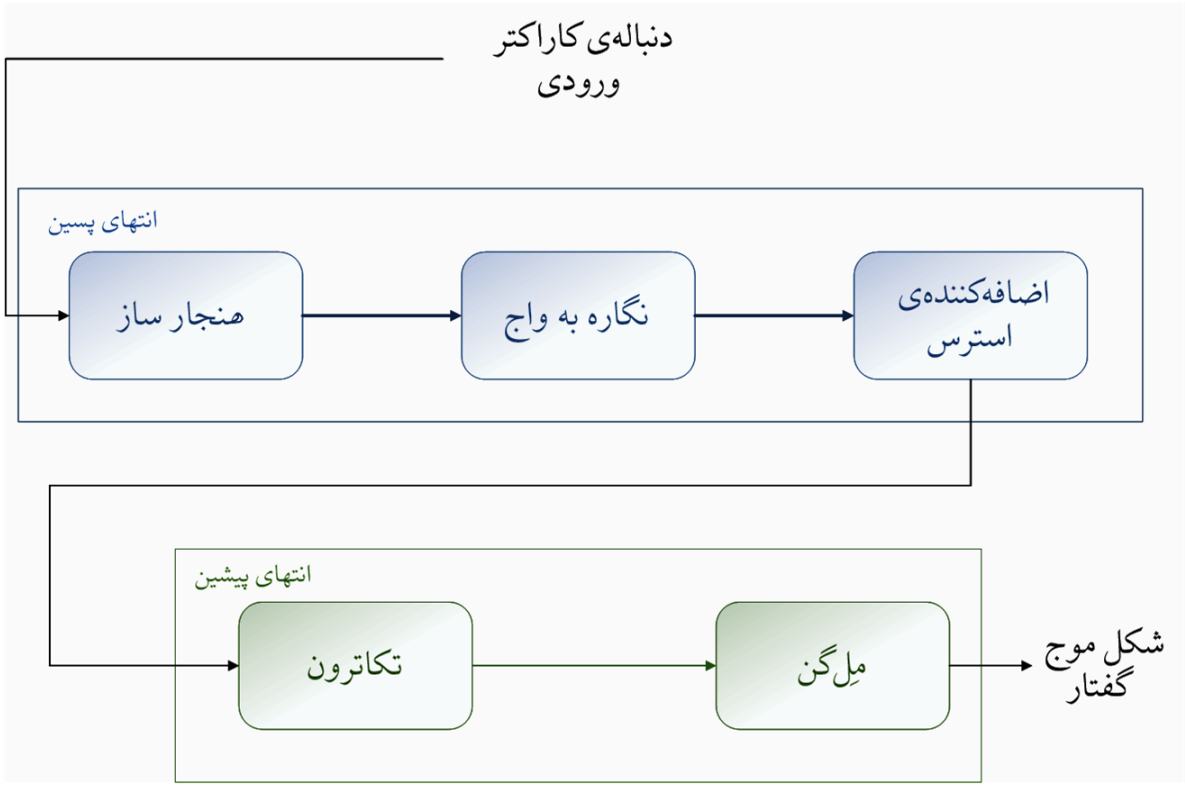
- آموزش پارامترهای شبکه‌ها بر اساس تابع هدف:

شکل: معماری شبکه‌ی مل‌گن

$$\min_{D_k} \mathbb{E}_x\left[\min(0, 1 + D_k(x))\right] + \mathbb{E}_z\left[\min(0, 1 - D_k(G(s, z)))\right], \forall k = 1, 2, 3$$

$$\min_{G} \mathbb{E}_z\left[-\sum_{k=1,2,3} D_k(G(z))\right]$$

شمای کلی سیستم پیشنهادی



شکل: شمایی از اتصال واحدهای استفاده شده در روش پیشنهادی

آموزش نگاره به واج

- استفاده از پیکره‌ی بیجن‌خان برای آموزش هر سه شبکه‌ی واحد نگاره به واج
- استفاده از مجموعه کاراکتر استفاده شده در داده‌ی آموزشی مدل سنتز گفتار برای برچسب گذاری دادگان این سه شبکه
- استفاده از تعدادی از کلمات این پیکره همراه با تلفظ آن‌ها برای آموزش شبکه‌ی پیش‌بینی تلفظ کلمات خارج از لغت‌نامه
- استفاده از لیست جمع‌آوری شده برای کلمات هم‌نگاره در زبان فارسی و جداسازی این کلمات از پیکره‌ی بیجن‌خان برای آموزش شبکه‌ی تشخیص کلمات هم‌نگاره
- جداسازی جملات حاوی کسره اضافه و نشان‌دار کردن کلمات حاوی کسره اضافه برای آموزش مدل تشخیص کسره‌ی اضافه

آموزش تکاترون و ملگن

- استفاده از دادگان سنتز گفتار شرکت عصر گویش پرداز
- شامل متون خبری خوانده شده توسط ۳ گوینده‌ی حرفه‌ای با نام‌های شفیعی(مرد)، کریمی(مرد) و صارمی(زن) ضبط شده در شرایط استودیویی
- مدت زمان فایل‌های خوانده شده توسط سه گوینده:
  شفیعی: ۵ ساعت و ۵ دقیقه
  کریمی: ۱ ساعت و ۵۰ دقیقه
  صارمی: ۵ ساعت و ۲۹ دقیقه
- آموزش سه مدل از شبکه‌ی تکاترون برای هر دادگان هر کدام از گویندگان
- آموزش ملگن روی مجموع دادگان سه گوینده

ارزیابی شاخص میانگین

جدول: ارزیابی شاخص نظرخواهی میانگین سه مدل آموزش داده شده با ۴ امتیاز دهنده و مدل‌های شاخص دیگر برگرفته از مقاله‌ی تکاترون ۲

	شاخص نظرخواهی میانگین
تکاترون	۴٫۰۰۱ ± ۰٫۰۸۷
مدل پارامتری مبتنی بر حافظه کوتاه-مدت ماندگار	۳٫۶۹ ± ۰٫۱۰۹
مدل مبتنی بر سنتز الحاقی	۴٫۱۶۶ ± ۰٫۰۹۱
ویونت	۴٫۳۴۱ ± ۰٫۰۵۱
تکاترون ۲	۴٫۵۲۶ ± ۰٫۰۶۶
مدل شفیعی	۴٫۲۲ ± ۰٫۰۷۳
مدل صارمی	۴٫۱۷ ± ۰٫۰۲۱
مدل کریمی	۳٫۹ ± ۰٫۰۴۵

معیار عامل بلادرنگ

- نشان‌دهنده‌ی سرعت تولید گفتار
- مدت زمان سپری شده به ثانیه برای تولید ۱ ثانیه از صوت خروجی
- با پیکربندی Intel(R) Core(TM) i5-3570K CPU @ 3.40GHz و Nvidia GeForce GTX 1080 Ti به صورت زیر می‌باشد:

معیار عامل بلادرنگ		
شفیعی	صارمی	کریمی
۰٫۱۷۸	۰٫۱۶۶	۰٫۱۷۶

کارهای آتی

- رویکرد جمع‌آوری داده‌ی آموزشی بیشتر
- تمرکز بیشتر روی نوای صحیح گفتار خروجی
- گوینده‌ی تطبیق‌پذیر یا مدل‌های آموزش داده شده با تعداد بیشتر گوینده
- تمرکز بر روی مقاوم‌سازی مدل نسبت به خطاهای مکانیزم توجه در جملات خاص یا بلند

با تشکر از شما