



درس:

پردازش سیگنال‌های دیجیتال

موضوع:

**ترجمه صوتی از سن به سن به روش زبانی برای
بهبود عملکرد تشخیص گفتار در محیط‌های واقعی**

استاد:

جناب آقای دکتر مهدی اسلامی

دانشجو:

حمیدرضا پورمحمد

شماره دانشجویی:

۴۰۰۱۴۱۴۰۱۱۱۰۳۳

چکیده	۳
اصطلاحات شاخص	۳
مقدمه	۳
کارهای مرتبط	۵
ترجمه سیگنال صوتی جفت زبانی	۶
خوشه‌بندی واج‌شناسی بدون نظارت	۷
ترجمه صوتی سن به سن	۱۰
راه‌اندازی آزمایشی	۱۲
مجموعه داده	۱۲
جزئیات پیاده‌سازی	۱۳
استفاده از سیستم باز ASR	۱۴
نتایج	۱۴
اجرای ASR در گفتار اصلی	۱۵
اجرای ASR در گفتار تبدیل شده	۱۵
اندازه‌گیری کیفیت صدای تبدیل شده	۱۷
تحلیل و بررسی	۱۹
مقایسه عملکرد با توجه به تعداد خوشه‌های زبانی	۲۰
نتیجه‌گیری و کارهای آینده	۲۱
منابع	۲۲

چکیده:

به یک مشکل کم کارایی سالمندان در تشخیص خودکار گفتار (ASR) از طریق سازگاری ویژگی‌های آگنوستیک به ASR رسیدگی شده است. بیشتر مجموعه داده‌های مدل‌های تشخیص گفتار از مجموعه داده‌های جمع‌آوری شده از سخنرانان بزرگسال تشکیل شده‌اند. در نتیجه، اکثر سیستم‌های تشخیص گفتار تجاری معمولاً روی سخنرانان بزرگسال عملکرد خوبی دارند. به عبارت دیگر، تنوع محدود سخنرانان در مجموعه داده‌های آموزشی، عملکرد غیرقابل اعتمادی را برای سخنرانان اقلیت (به عنوان مثال، افراد مسن) به دلیل دستیابی غیرممکن از داده‌های آموزشی ایجاد می‌کند. در پاسخ، این مقاله یک چارچوب تبدیل صدا مبتنی بر شبکه عصبی را برای تقویت تشخیص گفتار اقلیت پیشنهاد می‌کند. برای این منظور، یک مدل ترجمه صوتی شامل یک خوشه‌بندی واج‌شناسی بدون نظارت برای استخراج اطلاعات زبانی برای گفتار اقلیت در چارچوب مدل آگوستیک فعلی پیشنهاد شده است. یک روش انطباق ویژگی طیفی است که می‌تواند در مقابل هر سیستم ASR تجاری یا باز قرار گیرد و از تغییر مستقیم تشخیص دهنده گفتار اجتناب شود. نتایج تجربی و تجزیه و تحلیل اثربخشی روش پیشنهادی از طریق بهبود دقت تشخیص گفتار سالمندان نشان می‌دهد.

اصطلاحات شاخص: تشخیص گفتار، ترجمه صوتی، تبدیل ویژگی طیفی، گفتار بر حسب سن

مقدمه:

فناوری تشخیص خودکار گفتار (ASR) به درک یک گفتار ورودی متوالی به عنوان یک کلمه یا کاراکتر مربوطه و تبدیل آن‌ها به یک جمله متنی کامل اشاره دارد. برای ساختن متن از ویژگی‌های آگوستیک در مرحله تشخیص، مدل‌های آگوستیک و زبان پیچیده مورد نیاز است. در میان چندین آزمایش برای افزایش عملکرد ASR، سیستم ASR مبتنی بر شبکه عصبی عمیق - مدل پنهان مارکوف (DNN-HMM) به پیشرفت‌های قابل توجهی در سیستم‌های تشخیص موجود منجر شده است [۱] [۵]. دو ماژول در این مکانیسم ترکیبی به ترتیب نقش‌های متفاوتی دارند. DNN احتمال مشاهدات را برای همه حالت‌های تلفن سه‌گانه محاسبه می‌کند و HMM ویژگی‌های متوالی اطلاعات واج به دست آمده از مدل DNN را محاسبه می‌کند. این ویژگی DNN به دلیل تخمین توزیع مستقیم هنگام مدل‌سازی احتمال خلفی ویژگی‌های صوتی گفتار امکان‌پذیر است.

بر خلاف سیستم تشخیص گفتار موجود مبتنی بر HMM، که در آن زمان جریان اصلی بود [۶]، روش‌های ASR انتهایی منتشر شدند و به تدریج به‌روزرسانی شدند [۷] [۱۶] که تشخیص گفتار با

کارایی بالا را حتی با یک سیستم بزرگ فعال کرد. مقیاس مجموعه داده گفتار واژگان علاوه بر این، در تنظیمات عمومی بدون نظارت، مدل‌های جدید ASR عملکرد بهتری نسبت به یادگیری تحت نظارت در وظایف پایین دستی نشان داده‌اند [۱۷] [۲۶]. به این ترتیب، سیستم ASR به دلیل کاربرد عملی آن توجه زیادی را از حوزه‌های مختلف به خود جلب کرده است و به سرعت در حال توسعه است. در این مقاله، بر روی حل سوگیری داده‌ها با استفاده از تشخیص دهنده‌های گفتار تجاری واقعی که با واژگان در مقیاس بزرگ آموزش داده شده‌اند، تمرکز شده است.

در حالی که اکثر سیستم‌های تشخیص گفتار معمولاً روی بلندگوهای بزرگسال عملکرد خوبی دارند، ادعا می‌شود که سیستم‌های ASR فعلی به دلیل تفاوت در تنوع همبستگی صوتی و کلامی، مستعد ارائه عملکرد غیرقابل اعتماد برای اقلیت (مانند افراد مسن) هستند [۲۷] [۳۱]. بدتر از آن، بیشتر مجموعه داده‌هایی که برای آموزش سیستم‌های ASR استفاده می‌شوند، عمدتاً از گفتار جمع‌آوری شده از بزرگسالان عادی تشکیل شده‌اند. بنابراین، بخش کوچکی از اکثر داده‌ها در برابر یک مشکل عدم تعادل معمولی آسیب‌پذیر هستند [۳۲]، که مانع از درک دقیق گفته‌های سالمندان توسط مدل‌های ASR می‌شود.

اگرچه تنظیم مجدد یک مدل ASR از پیش آموزش‌دیده به بهبود عملکرد تشخیص گفتار سیستم‌های ASR کمک می‌کند، مجموعه‌های داده‌ای که به ندرت حاوی صدای افراد مسن هستند همچنان دامنه پیشرفت‌ها را محدود می‌کنند. حتی اگر بتوان مدل را با صدای سالمندان کوک کرد، حل مشکلات ریشه‌ای که بیشتر به ویژگی‌های مختلف در گفته‌های سالمندان نسبت داده می‌شود، سخت است. علاوه بر این، ساخت یک مجموعه داده بزرگ گفتار سالمندان در مقایسه با بزرگسالان نسبتاً پر زحمت و پرهزینه است. علاوه بر این، ساخت مدل ASR از ابتدا یک کار دشوار است که به دانش، قدرت محاسباتی و منابع داده قابل توجهی نیاز دارد.

برای مقابله با این مشکل، یک پارادایم جدید برای بهبود عملکرد مدل‌های عمومی ASR پیشنهاد شده است که معمولاً تشخیص ضعیف گفته‌های سالمندان را نشان می‌دهند. برای این منظور، یک چارچوب تبدیل صدای سن به سن را پیشنهاد کرده‌اند، که یک روش خوشه‌بندی واج‌شناسی بدون نظارت را برای پل زدن ویژگی‌های واجی مربوطه بزرگسالان و سالمندان معرفی می‌کند. در عمل، این روش می‌تواند به آرامی در مقابل هر مدل تجاری ASR به عنوان یک رویکرد انطباق ویژگی طیفی ملحق شود. رویکرد پیشنهادی عملکرد تشخیص گفتار سالمندان را افزایش می‌دهد و مستقیماً بر عملکرد سیستم ASR اصلی تأثیر نمی‌گذارد. از این نظر، هدف ارائه عملکرد ASR بالا برای سالمندان با تبدیل صدای آنها به صدای بزرگسالان است.

برای پیاده‌سازی مدل پیشنهادی، از یک مجموعه داده سالمندان باز استفاده شده است و از مجموعه داده گفتار بزرگسالان دیگر (۲۰۰ ساعت) برای پل زدن ویژگی‌های واجی دو گروه سنی در خوشه‌بندی واج‌شناسی استفاده شده است. برای آموزش مدل تبدیل صدا، داده‌هایی را جمع‌آوری کردند که حاوی جملات کوتاه دستوری است که معمولاً برای کارکردن دستگاه‌های خانه هوشمند با صدای بزرگسالان در کاربرد عملی استفاده می‌شوند. صداهای دو گروه سالمندان و بزرگسالان به ترتیب جفت می‌شوند.

برای تأیید صحت کامل چارچوب پیشنهادی، از یک سیستم ASR باز استفاده شده است. نتایج تجربی کارآیی مدل ترجمه صوتی پیشنهادی از طریق بهبود دقت تشخیص گفتار تأیید می‌کند.

کارهای مرتبط:

تبدیل صدا (VC) به فرآیند تبدیل سبک صدای ورودی به هدف حفظ اطلاعات زبانی ورودی اشاره دارد. VoiceGAN [۳۳] یک چارچوب انتقال سبک برای VC بر اساس شبکه‌های متخاصم مولد پیشنهاد کرد. این مدل یاد گرفته است که سبک گفتاری گفته‌های گوینده ورودی را بدون جاسازی اطلاعات زبانی اضافی به هدف تبدیل کند. Parrotron [۳۴] با موفقیت گفتار را از گویندگان ناتوان به گفتار معمولی تبدیل کرد و عملکرد تشخیص گفتار را برای افرادی که از نظر جسمی در بیان خود محدود هستند بالا برد. برای بهبود عملکرد تبدیل سیگنال به سیگنال، یک شبکه تشخیص گفتار کمکی نیز به رمزگذار متصل شد که نشان داد تشخیص آموزش چند وظیفه‌ای برای استحکام مدل موثر است. در کار قبلی [۳۵]، یک VC سرتاسر در سطح طیف‌گرای گفتار ورودی بدون خوشه‌بندی زبانی انجام شد.

برخی از تحقیقات قبلی برای به دست آوردن نمایش‌های قدرتمند از طریق شبکه Transformer [۳۶] که قادر به محاسبه متن به گفتار (TTS) و VC است، تلاش کرده است. هوانگ و همکاران [۳۷] VC را از طریق یک مدل TTS از پیش آموزش دیده انجام داد، که شبکه مبتنی بر ترانسفورماتور است که با مجموعه‌ای در مقیاس بزرگ برای انتقال دانش برای فرآیند تبدیل آموزش داده شده است. از آنجایی که این رویکرد وزن‌های از پیش آموزش دیده را از مدل TTS به ارث برده است، گفتار تولید شده از رمزگشا از نظر تنوع محدود است. لیو و همکاران [۳۸] بر مکانیسم حفاظت از زمینه مبتنی بر شبکه ترانسفورماتور و یک مدل TTS تک بلندگوی از پیش آموزش دیده در منظر انطباق مدل برای VC یک به یک متمرکز شد. در حالی که یادگیری دوگانه متشکل از TTS و ASR

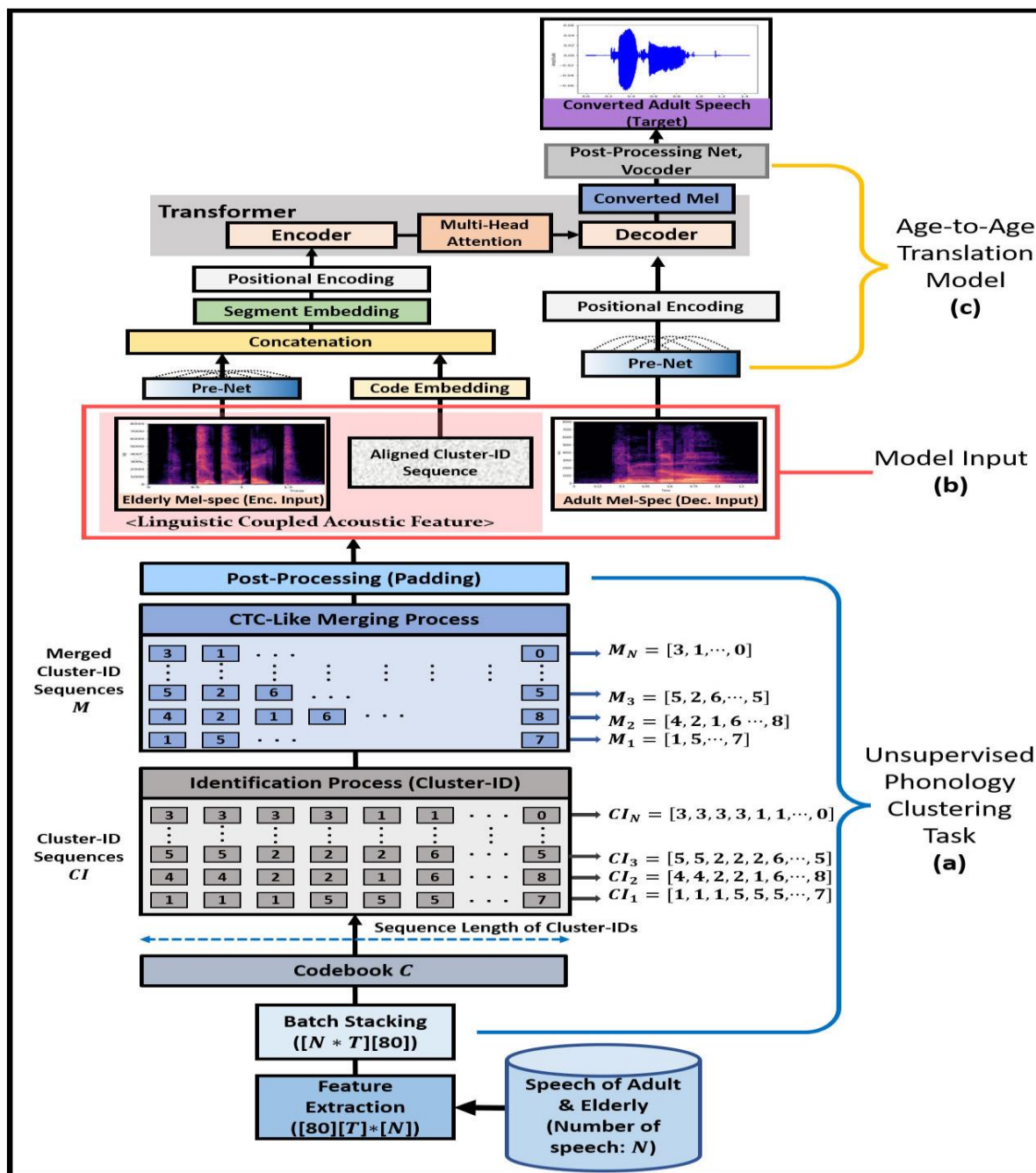
تنها بر روی گرفتن بازنمایی‌های نهفته برای متن و گفتار متمرکز است [۳۹]، [۴۰]، این رویکرد سعی می‌کند ارزش‌های نهفته زبانی را در گفتار ناهمگون بزرگسالان و سالمندان پل کند.

هدف بهبود عملکرد کم ASR در گفتار سالمندان همانطور که در Parrotron [۳۴] توضیح داده شده است. با این حال، به جای وظایف کمکی، اطلاعات زبانی بیان شده از واج‌های مشابه در گروه‌های سنی مربوطه را با روش خوشه‌بندی واج‌شناسی بدون نظارت، جفت می‌کنند.

ترجمه سیگنال صوتی جفت زبانی:

در این بخش، ترجمه سیگنال صوتی همراه با اطلاعات زبانی را شرح داده شده است. این برای استخراج ویژگی‌های ذاتی در گفته‌های سالمندان که در مجموعه داده‌های معیار پراکنده هستند، ابداع شده است. همچنین روشی را ارائه شده است که ترجمه صوتی موفق را بین سخنرانانی که سبک‌های گفتار و تلفظ متفاوتی دارند به دلیل اندام‌های صوتی پیرشان ممکن می‌سازد. به عنوان یک رویکرد جدید، یک روش ساده بدون نظارت را برای اتصال یک رابطه واج‌شناسی بین ویژگی‌های گفتاری دو گروه که در هر قاب Mel-spectrogram وجود دارد، اعمال می‌شود. به عبارت دیگر، فریم‌های Mel-Spectrogram به دست آمده از واج‌های یکسان به داشتن ویژگی‌های مشابه در گروه‌های مربوطه مرتبط هستند. اطلاعات زبانی مبتنی بر گفتار را با استفاده از خوشه‌بندی K-means استخراج می‌شود تا اطلاعات زبانی را برای فریم‌های ویژگی مربوطه از داده‌های گفتار ترکیب شود. شکل ۱ روش کلی پیشنهادی را نشان می‌دهد. مشارکت‌های اصلی به شرح زیر است:

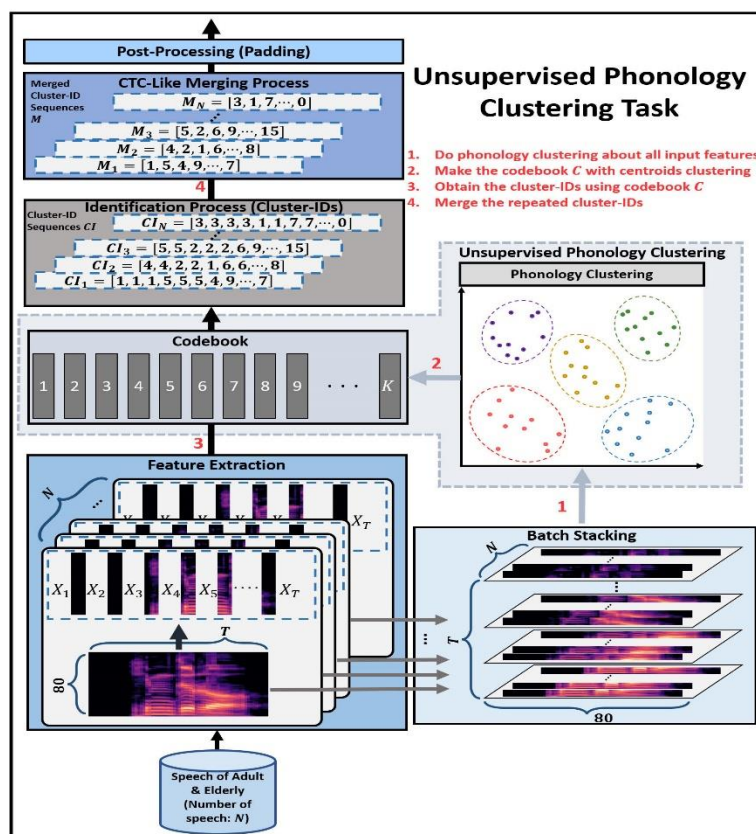
- یک روش خوشه‌بندی واج‌شناسی بدون نظارت را پیشنهاد کردند تا ارتباط بین واج‌های مشابهی را که هم در گفته‌های بزرگسالان و هم در بزرگسالان یافت می‌شود و برای پل زدن ویژگی‌های همگن در سیگنال‌های گفتاری گروه مربوطه، ایجاد شود.
- سپس می‌توان با استفاده از کتاب کدی که از خوشه‌بندی K-means به دست می‌آید، از چارچوب ویژگی کوانتیزه شده به دست آورد. در این فرآیند، به دلیل سرعت متفاوت گفته‌ها، شناسه خوشه بین دو سخنران که یک جمله را بیان می‌کنند ممکن است با فریم به فریم مطابقت نداشته باشد. برای این منظور، فرآیند ادغام برای ایجاد یک توالی Cluster-ID منحصر به فرد انجام می‌شود.
- در نهایت، نحوه استفاده از اطلاعات زبانی را به عنوان ورودی مدل ترجمه صوتی پیشنهادی و آموزش آن با روش جاسازی کد نشان داده شده است.



شکل ۱. شماتیک کامل مدل ترجمه صوتی از سن به سن همراه با زبان

خوشه بندی واج شناسی بدون نظارت

دشواری های جمع آوری صداهای سالمندان و کودکان باعث شده است که بیشتر مجموعه داده باز از گفته های بزرگسالان تشکیل شود. از این نظر، مشکل داده های گفتاری نسبتاً کمیاب سالمندان را با استفاده از داده های بزرگسالان حل کردند. برای این منظور، ابتدا ویژگی های همگن از گفته های هر گروه سنی را پل کردند. شکل ۲ فرآیند گام به گام روش خوشه بندی واج شناسی پیشنهادی را نشان می دهد:



شکل ۲. روش خوشه‌بندی واج‌شناسی بدون نظارت برای استخراج اطلاعات زبانی

در اینجا، هر گفتار را هم در بزرگسالان و هم در افراد مسن سازماندهی شده است و مجموعه داده مانند بزرگسالان $A = a_1, a_2, \dots, a_n$ و افراد مسن $E = e_1, e_2, \dots, e_n$. بخش ساکت در هر گفته‌ای است از طریق یک الگوریتم تشخیص فعالیت صوتی پردازش می‌شود. همه ویژگی‌های گفتار از طریق یک لاگ ۸۰ بعدی استخراج می‌شوند Mel-Iterbank و شکل هر ویژگی می‌شود $(80, T)$.

سپس طیف‌نگارهای Mel مربوطه را برای نوشتن پشته می‌کنیم. دسته‌ها در نتیجه، کل پایگاه داده را می‌توان بیان کرد مانند $(80, N \times T)$ ، که در آن N تعداد داده‌ها و T طول توالی هر طیف Mel، است و $N \times T$ هم $\sum_{i=1}^N T$ است.

پس از رویه انباشتگی دسته‌ای، دریافت می‌کنیم $X = \{x\}_{i=1, t=1}^{N, T}$ که شامل تمام بردارهای ویژگی Mel-spectrogram جدا شده برای گام تمام زمان از A و E («۱» را در شکل ۲ ببینید). سپس کمیت می‌کنیم X با پارتیشن‌بندی R^{80} به مناطق K که r_1, \dots, r_k . این محاسبات را می‌توان با ساخت یک کتاب کد پردازش کرد $C = c_1, \dots, c_k$ ، جایی که c_k یک کلمه رمز را نشان می‌دهد که مجموعه‌ای از مرکزها با استفاده از K -نزدیکترین همسایه است. بنابراین، K شناسه‌های کتاب کد برای نمایش X از طریق واج‌شناسی استفاده می‌شود (خوشه‌بندی «۲» را در شکل ۲ ببینید).

توجه داشته باشید که کتاب کد از روش ما با رویکرد *Gumbel-Softmax* آموزش داده نشده است شرح داده شده در [۴۱]، [۴۲].

ما کتاب کد C را مقداردهی اولیه می‌کنیم و بهینه را کشف می‌کنیم پارتیشن‌های R_k از:

$$R_k = \{x: d(x, c_k) \leq d(x, c_j), 1 \leq k \neq j \leq K\} \quad (1)$$

که در آن $d\{(a, b)\}$ فاصله اقلیدسی بین a و b است، c شاخص کتاب کد است، j یک مرکز متفاوت و K است تعداد کلمات رمز که در آن $k = 1, \dots, K$. سپس اطلاعات واج شناسی با استفاده از کلمه رمز k -امین c_k استخراج می‌کنیم از:

$$c_k \leftarrow \underset{c_k}{\operatorname{argmin}} E[(d(x, c_k) | x \in R_k)], \quad k = 1, \dots, K \quad (2)$$

بر اساس این خوشه‌بندی، تمام بردارهای ویژگی X با نزدیک‌ترین کلمه رمز به‌عنوان شناسه خوشه نشان داده می‌شوند («۳» را در شکل ۲ ببینید). از طریق این روش، می‌توان به طور مؤثری بر واج چارچوبی که شباهت زبانی را در مجموعه داده‌های سالمندان و بزرگسالان نشان می‌دهد، پل زد. در عمل، A و E به دنباله‌های CI مانند شکل ۲ تبدیل می‌شوند.

با این حال، از آنجایی که سرعت گفتار برای سخنرانان مسن نسبتاً کندتر از بزرگسالان است، CI از A و E ممکن است فریم به فریم مطابقت نداشته باشد. به طور دقیق، ما نیاز داریم نه تعداد هر خوشه- ID ، بلکه یک دنباله منحصر به فرد یک $Cluster-ID$ برای نشان دادن گفته تشکیل شده است. به منظور رفع مشکلات فوق، تکنیک برگرفته از کاتیون رده‌بندی زمانی ارتباط‌گرا (CTC) پیشنهاد می‌شود [۴۳].

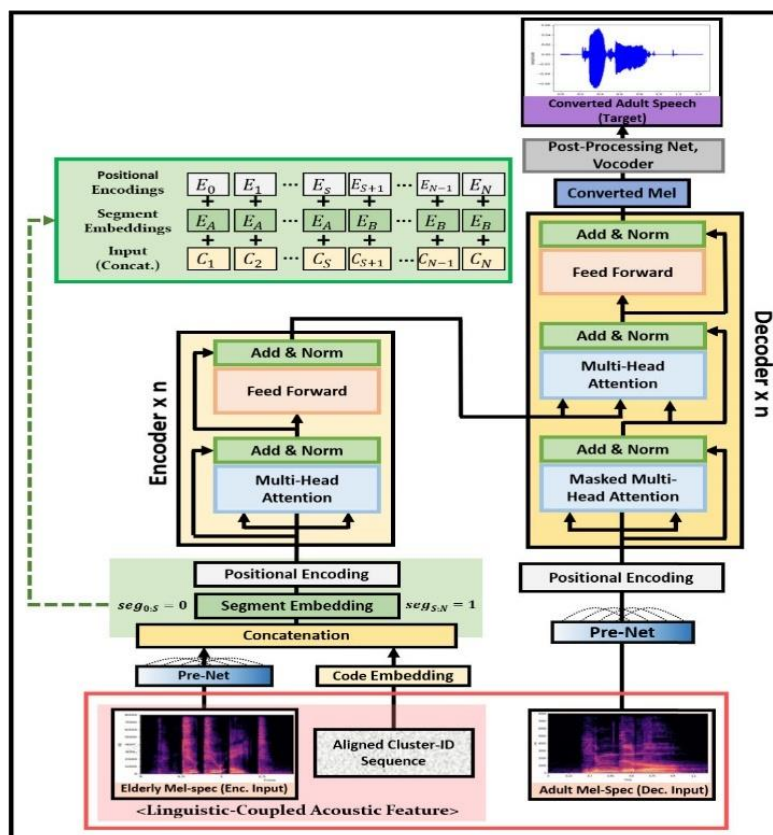
اگر زمان فعلی، حداکثر طول زمان و مقادیر i -th به ترتیب با i نشان داده می‌شوند، T و $value_i$ ، ما شناسه‌های خوشه‌ای مکرر را از i به $t+i$ ادغام می‌کنیم تا اینکه $value_i \neq value_{t+i}$ برای به دست آوردن توالی $Cluster-ID$ غیر تکراری. در فرآیند ادغام، در مواجهه با عدم تکرار- $cluster-ID$ ، گام زمانی به مرحله بعدی افزایش می‌یابد. i و این روند بازه تا زمانی ادامه می‌یابد که T در این مرحله به این ترتیب، به حداقل رساندن خطا بین آنها امکان پذیر می‌شود (CI ‌های دو گروه سنی به "۴" در شکل ۲ مراجعه کنید).

انتظار این است که اعمال رویه فوق‌الذکر قادر به بهبود عملکرد ترجمه صوتی باشد. علاوه بر این، بین دو نسل مختلف سنی ویژگی‌های زبانی مبتنی بر گفتار استخراج شده از خوشه‌بندی و ویژگی‌های $\log Mel-spectrogram$ برای تشکیل ویژگی آکوستیک همراه زبانی می‌توان از آن برای استخراج نمایش معنادار استفاده کرد.

ترجمه صوتی سن به سن (A2AVT):

شبکه مبتنی بر ترانسفورماتور [۳۶] که قادر به محاسبات موازی سریع است جزء اصلی مدل پیشنهادی است. مطالعات اخیر تأیید کرده‌اند که شبکه‌های مبتنی بر ترانسفورماتور برای تبدیل صدا مناسب هستند [۳۵]، [۳۷]، [۳۸]. برخلاف استفاده معمولی از مدل‌های ترانسفورماتور، مدل تبدیل صوتی پیشنهادی به وزنه‌های از پیش آموزش دیده نیازی ندارد. همچنین، یک ساختار رمزگذار متفاوت از رمزگشا را با جفت کردن اطلاعات زبانی با ویژگی‌های گفتاری با استفاده از شناسه‌های خوشه‌ای به دست آمده پیشنهاد شده است.

شکل ۳ مدل ترجمه سن به سن ما را نشان می‌دهد. در ماژول رمزگذار، طیف نگارهای log Mel گفتار سالمندان و توالی Cluster-ID مربوطه به عنوان ویژگی صوتی همراه با زبانی ورودی تغذیه می‌شوند. سپس ویژگی‌ها از Pre-Net [۴۴] عبور می‌کنند و Cluster-ID از طریق کد تعبیه شده با شماره کلاس $C + 1$ ادامه می‌یابد که در آن مقدار padding برای مطابقت با حداکثر طول اضافه می‌شود. استفاده از Pre-Net به تبدیل صدای واضح‌تر کمک می‌کند که هر دو ویژگی Cluster-ID و گفتار را به هم متصل می‌کند.



شکل ۳. معماری سرتاسر ترجمه صوتی متشکل از رمزگذار به رمزگشا

ویژگی های گفتار برای به هم پیوستن هر دو ویژگی Mel-spectrogram و cluster-ID، مفهوم تعبیه بخش و BERT [۴۵] را از آن قرض گرفته و آن را به ویژگی ورودی ما تغییر می دهد. در حالی که نشانه "[SEP]" بین جفت جمله قرار می گیرد در BERT، فقط ویژگی گفتار و دنباله را جفت می کند. شناسه های خوشه ای بدون رمز، موقعیت بخش مختلف شاخص ها را برای تشخیص ویژگی گفتار و Cluster-ID است. مقدار "۰" به موقعیت ویژگی گفتار ($seg_{0:s} = 0$) و "۱" به مکان اختصاص داده شده است از خوشه-ID ($seg_{s:N} = 1$). در نتیجه، واژگان اندازه جاسازی بخش "۲" است. موقعیت سپس رمزگذاری برای کل ورودی الحاق شده به آن اضافه می شود مقادیر موقعیت مطلق را بدست می آید. فقط در مازول مرحله آموزش رمزگشا طیفنگار Mel بزرگسالان هدف به عنوان ورودی تغذیه می شود.

Pre-Net از دو لایه کاملاً متصل تشکیل شده است که همانطور که در [۴۴] توضیح داده شد، ۰.۵ احتمال خروج دارند. اندازه پنهان Pre-Net مورد استفاده در معماری ما ۲۵۶ است. هر لایه رمزگذار منفرد از دو لایه فرعی تشکیل شده است: خود توجهی چند سر و لایه فوروارد. دو جزء فرعی از طریق یک شبکه باقیمانده به هم متصل می شوند و نرمال سازی لایه [۴۶] برای هر یک از آنها اعمال می شود.

ماژول رمزگشا تقریباً مشابه رمزگذار است به جز اعمال پوشش نگاه به جلو در توجه چند سر. در مرحله زمانی t ماژول رمزگشا، پوشش نگاه به جلو از ویژگی های گفتاری آینده جلوگیری می کند. $m_{t+1}, m_{t+2}, \dots, m_T$ که فریم های Mel-spectrogram هدفی هستند که در معرض مدل قرار می گیرند. هنگامی که t در حال افزایش است، به دلیل مکانیسم خود رگرسیون، پوشش نگاه به آینده به تدریج باریک می شود. علاوه بر این، قسمت صفر تا حداکثر طول در محاسبه مرحله توجه خود نادیده گرفته می شود.

برای ارزیابی کیفیت صدای تبدیل شده، به یک شبکه پس پردازش [۴۴] برای تبدیل طیفنگار Mel به طیفنگار خطی، و یک سینتسایزر [۴۷] برای بازسازی طیفنگار خطی به شکل موج نیاز است. این فرآیند بازسازی در شرایط مشابه [۴۴] اعمال می شود، به جز ضریب کاهش τ .

تابع هدف ما L_{sum} شامل دو ضرر L_{mel} و L_{spec} است به عنوان:

$$L_{sum} = L_{mel} + L_{spec} \quad (3)$$

جایی که از دست دادن L_{mel} بین بزرگسالان هدف مشتق می شود Mel-spectrogram و نتیجه تبدیل شده که از دست دادن L_{mel} از آخرین بلوک رمزگشا خروجی است. به عنوان:

$$L_{mel} = L_1\lambda_1 + L_2\lambda_2 = \sum_{i=1}^N \sum_{k=1}^T |m_{ik} - \hat{m}_{ik}| \lambda_1 + \sum_{i=1}^N \sum_{k=1}^T (m_{ik} - \hat{m}_{ik})^2 \lambda_2 \quad (4)$$

که در آن N تعداد داده‌ها، T طول زمانی هر مجموعه داده، m قاب Mel-طیف‌گرام هدف، و \hat{m} به ترتیب قاب Mel-spectrogram پیش‌بینی شده است. L_{mel} به عقب انتشار می‌یابد و بر وزن‌ها در مدل تبدیل صدای پیشنهادی تأثیر می‌گذارد که تبدیل بین سخنرانی‌های دو نسل را مدیریت می‌کند.

از سوی دیگر، از دست دادن L_{spec} در معادله (۳) توضیح داده شده است. بین طیف‌نگار خطی بزرگسالان هدف به دست می‌آید و طیف نگار خطی پیش‌بینی شده تبدیل شده-Mel spectrogram (به دست آمده از آخرین بلوک رمزگشا). را بیان ریاضی از دست دادن L_{spec} به صورت V است:

$$L_{spec} = L_1\lambda_1 + L_2\lambda_2 = \sum_{i=1}^N \sum_{k=1}^T |s_{ik} - \hat{s}_{ik}| \lambda_1 + \sum_{i=1}^N \sum_{k=1}^T (s_{ik} - \hat{s}_{ik})^2 \lambda_2 \quad (5)$$

که در آن s یک قاب طیف‌گرام خطی هدف و \hat{s} یک قاب پیش‌بینی شده است قاب طیف‌گرام خطی \hat{m} در معادله (۴) است. در دیگر کلمات، L_{spec} در شبکه پس از پردازش که است آموزش دیده برای حدس زدن تمایل قدر طیفی خطی با تخمین ویژگی‌های طیفی نمونه برداری شده در مقیاس فرکانس مل با مقیاس فرکانس خطی [۴۴] است. بر اساس این روش، می‌شود طیف‌نگار خطی را بازسازی کرد به عنوان شکل موج با استفاده از سینت سائزر استفاده می‌شود. الگوریتم Griffin-Lim [۴۷] به عنوان سینت سائزر برای بازسازی طیف نگار خطی تبدیل شده به عنوان شکل موج است. ۱ و ۲ مورد استفاده در هر دو معادله (۴) و معادله (۵) پارامترهای فوق هستند برای تنظیم نسبت هر لاجیت و فرمول ریاضی نسبت به صورت زیر توضیح داده شده است:

$$\lambda_1 + \lambda_2 = 1.0 \quad (6)$$

راه اندازی آزمایشی:

مجموعه داده:

از مجموعه داده گفتار VOTE400، مجموعه داده گفتار AIHub و مجموعه داده مرجع بزرگسالان (ARD) استفاده شده که از دو سخنران جمع آوری شده است. مجموعه داده VOTE400 به طور کلی شامل ۴۰۰ ساعت صحبت ثبت شده سالمندانی است که میانگین سنی آنها ۷۹.۴۷ سالمند

است. نسبت جنسیت در این مجموعه داده ۵.۲۹:۱ (زنان در مقابل مردان) است. بر روی عباراتی تمرکز شده که کلمات دستوری کوتاهی دارند و عمدتاً برای کار با دستگاه‌های خانه هوشمند گفته می‌شوند. بر اساس این شرط، ۱۳۳۰ لغت آموزشی ضبط شده توسط ۱۰۳ سخنران و ۱۶۱ آزمون ثبت شده توسط ۶۲ سخنران (نسبت ۹:۱) نمونه برداری شد.

مجموعه داده AIHub از حدود ۱۰۰۰ ساعت مکالمه تشکیل شده است که توسط ۲۰۰۰ بزرگسال بیان شده است. این مجموعه داده در یک محیط گفتگوی روزانه که در آن نویز، خنده و تنفس گنجانده شده است، ثبت شده است. از میان کل مجموعه داده، فقط ۲۰۰ ساعت از مجموعه داده AIHub را انتخاب کردند تا فقط برای عملکرد مدل‌سازی خوشه واج شناسی، نه برای ترجمه سن به سن انجام شود. به عبارت دیگر، این ۲۰۰ ساعت داده را با مجموعه داده سالمندان VOTE400 ترکیب شده تا شناسه‌های خوشه‌ای تراز شده را برای آماده‌سازی آموزش تبدیل صدا به دست آوردند.

برای ایجاد یک داده هدف که با داده‌های ورودی سالمندان مطابقت دارد، مجبور شدند صدای بزرگسالان را ضبط کنند. ۲ سخنران، یک مرد سی ساله و یک زن بیست و هفت ساله را جمع‌آوری کردند. ضبط در یک اتاق کوچک CE انجام شد و مرد صدای خود را با تلفن همراه خود که دارای نرخ نمونه برداری ۴۸ کیلوهرتز است، ضبط کرد. همچنین این زن صدای خود را با تلفن همراه خود که دارای نرخ نمونه برداری ۴۴.۱ کیلوهرتز است ضبط کرده است. در نتیجه، از ARD و گفته‌های مربوط به سالمندان با رونویسی یکسان برای آموزش مدل تبدیل صدای پیشنهادی استفاده کردند. توجه داشته باشید که سه مجموعه داده فوق در شرایط مختلف ثبت شدند. این از این جهت قابل توجه است که روش پیشنهادی می‌تواند در محیط‌های واقعی کار کند.

جزئیات پیاده‌سازی:

به طور جزئی، بعد پنهان مدل در شکل ۳ ۲۵۶ است، تعداد توجه چند سر ۸ است، نرخ انصراف ۰.۱ تنظیم شده است و اندازه بعد پنهان پیشخور ۱۰۲۴ است. مدل دارای به ترتیب ۶ لایه رمزگذار و ۴ لایه رمزگشا است. در ماژول ورودی، ماسک‌سازی اعمال می‌شود تا توجه به تمرکز بر بخش‌های بدون لایه ویژگی‌های صوتی همراه با زبان از جمله شناسه‌های خوشه‌ای و طیف‌نگاری Mel را متوقف کند. زمانی که خودتوجهی در بلوک توجه چند سر انجام می‌شود، موقعیت‌های پوشاندن اعمال می‌شود.

سایز دسته را ۱۶ قرار دادند و مدل را آموزش دادند تا ۸۵۰۰۰ قدم ≈ 1060 دوران با یک NVIDIA RTX، که ۲۴ گیگابایت حافظه دارد. از بهینه‌ساز Adam استفاده شده است [۴۸] با $\beta_1 = 0.9$ و

$\beta_2 = 0.98$ و عضوی از $1e-9$ و 4000 مرحله است. هر کدام از هر دو λ_1 و λ_2 برابر 0.5 است (۴)(۵).

از آنجایی که نرخ‌های نمونه‌گیری مختلفی در مجموعه داده‌ها وجود دارد، همه نرخ‌های نمونه‌برداری را 16 کیلوهرتز در نظر گرفتند و یک نمونه‌گیری مجدد روی همه آن‌ها انجام دادند. از طیف‌نگار مل 80 بُعدی با اندازه پنجره 32 میلی‌ثانیه و اندازه همپوشانی 16 میلی‌ثانیه و تبدیل فوریه 512 نقطه‌ای استفاده کردند و همه آن‌ها نرمال می‌شوند.

برای ساخت خوشه‌های زبانی بدون نظارت، اندازه کتاب کد 32 است. الفبای کره‌ای شامل 19 صامت است. 10 تک صدایی، و 11 دوفتونگ. اگر صداها انتقالی در نظر گرفته نشود، مجموعه واج را میتوان به 32 تا 36 واحد تقسیم کرد.

استفاده از سیستم باز ASR:

برای مقایسه عینی عملکرد تشخیص صداها، اصلی سالمندان و نتایج تبدیل شده، از یک بازشناس باز تجاری استفاده کردند که قادر به مدیریت واژگان در مقیاس بزرگ است. از آنجایی که سیستم **ASR** باز فقط به شکل موج‌ها به عنوان ورودی سیستم اجازه می‌دهد، شبکه پس پردازش و سینت سائزر **Grifn-Lim** را برای بازیابی طیف‌نگارهای **Mel** تبدیل شده به شکل موج اضافه کردند تا مدل را ارزیابی کنند.

نتایج:

نتیجه عملکرد **ASR** در گفتار اصلی سالمندان و **ARD** در این بخش نشان داده شده است. سپس عملکرد بهبود یافته **ASR** از صدای تبدیل شده حاصل از مدل پیشنهادی را نشان می‌دهند. علاوه بر این، توزیع شناسه‌های خوشه‌ای از گفته‌ها، که در رونویسی‌های مشابه توسط افراد مسن و بزرگسالان بیان می‌شوند، مقایسه می‌شود. علاوه بر این، نتایج ارزیابی میانگین امتیاز نظر (**MOS**) بر روی صدا ارائه شده است.

جدول ۱. نتایج تشخیص گفتار برای گفته‌های سخنرانان بزرگسال و مسن

Group	Gender	CER (%)	Average CER (%)
Adult	Male	11.08	12.80
	Female	14.51	
Elderly	Male	28.82	27.13
	Female	25.44	

اجرای ASR در گفتار اصلی:

نرخ خطای کاراکتر (CER) را به عنوان معیاری برای ارزیابی دقت تشخیص گفتار تنظیم کردند. ابتدا، عملکرد هر دو مجموعه آزمون سالمندان و مجموعه داده‌های بزرگسالان هدف را ارزیابی می‌کنند. همانطور که در جدول ۱ نشان داده شده است، CER میانگین مجموعه تست سالمندان ۲۷.۱۳٪ است در حالی که CER بزرگسالان مرد ۱۱.۰۸٪ و بزرگسالان زن ۱۴.۵۱٪ و میانگین CER بزرگسالان ۱۲.۸۰٪ بود.

با توجه به نتایج، تشخیص‌دهنده گفتار تجاری به سخنرانان بزرگسال عادی بیشتر از گویندگان مسن عادت دارد.

قبل از اعمال رویکرد مبتنی بر یادگیری، علاوه بر این، ما از روش معمولی سازی طول دستگاه صوتی (VTLN) برای ارزیابی میزان بهبود عملکرد استفاده می‌کنیم. داده‌های گفتاری مردان و زنان مسن به ترتیب بر اساس مجاری صوتی مردانه و زنانه نرمال می‌شوند. جدول ۲ عملکرد روش VTLN را نشان می‌دهد. VTLN تا حدودی برای گفتار مردان مسن موثر است، اما پیشرفت کمی برای گفتار زنان دارد. این نشان می‌دهد که روش عادی سازی محدودیت‌هایی در بهبود عملکرد تشخیص گفتار سالمندان دارد.

جدول ۲. نتایج تشخیص گفتار پس از VTLN داده های سالمندان

VTLN		CER (%)	ERR (%)
Source	Normalization Target		
Elderly Male	Adult Male	20.71	28.14
Elderly Female	Adult Female	25.11	1.30

اجرای ASR در گفتار تبدیل شده:

هدف اصلی بهبود تشخیص گفتار سالمندان فراتر از نسل گفتار سالمندان از طریق ترجمه صوتی سن به سن است. از این منظر، عملکرد تشخیص گفتار سالمندان را با استفاده از تکنیک تولید صدای مرسوم مقایسه می‌کنیم.

برای این منظور، ارزیابی را با استفاده از رمزگذار خودکار حذف نویز (DAE) انجام دادند. DAE تفاوت‌های گفتار بزرگسالان را که در گفتار سالمندان به دلیل پیری اندام‌های صوتی رخ می‌دهد،

به عنوان «نویز» تعریف می‌کند. صداهای تبدیل شده افراد مسن را می‌توان از مدل DAE تحت شرایط آزمایشی مشابهی که مدل A2AVT ما آموزش داده شد به دست آورد.

نتایج چهار الگوریتم (A2AVT، DAE، A2AVT+Linguistic-ID، و A2AVT+ merged_Linguistic-ID) را در معیارهای CER گزارش کردند. نرخ کاهش خطا (ERR) در جدول ۳ نشان دهنده نسبت بهبود نسبی بین CER از نتیجه تبدیل شده و CER از سخنرانی اصلی سالمندان در جدول ۱ است.

جدول ۳. مقایسه عملکرد بین روش های DAE و A2AVT.

Method	Conversion Target	CER (%)	ERR (%)
DAE	Adult Male	20.53	28.76
	Adult Female	18.76	26.26
A2AVT	Adult Male	22.96	20.33
	Adult Female	20.97	17.57
A2AVT + Linguistic-ID	Adult Male	24.50	14.99
	Adult Female	22.51	11.52
A2AVT + merged_Linguistic-ID	Adult Male	19.21	33.34
	Adult Female	14.35	43.59

۲۸.۷۶٪ از میانگین ERR را از طریق روش DAE معمولی دریافت شده است. ۱۸.۹۵٪ از میانگین ERR را از طریق مدل A2AVT بدست آورده شده است. از سوی دیگر، نتایج ارزیابی A2AVT + Linguistic-ID 13.25 درصد از میانگین ERR را نشان می‌دهد که کمی بیشتر از گفتار اصلی سالمندان است. بر اساس نتایج شناسایی شده، انتظار بر این است که با توجه به سرعت گفتار متفاوت سخنرانان بزرگسال و مسن، تراز بخشی از دنباله زبانی-ID مورد نیاز است.

بنابراین، بهترین پیشرفت را در روش A2AVT + merged_Linguistic-ID برای ادغام شناسه‌های مکرر از همان بخش های زبانی در یک شناسه واحد به دست آورد. در حالی که میانگین CER سخنرانی اصلی سالمندان ۲۷.۱۳٪ است، مدل پیشنهادی A2AVT + merged_Linguistic-ID 16.78٪ از میانگین CER و ۳۸.۴۷٪ از میانگین ERR را نشان می‌دهد. این نشان می‌دهد که تکنیک ادغام پیشنهادی در این مقاله می‌تواند ویژگی‌های همگن را در مد گفتار ناهمگن در خوشه‌بندی واج‌شناسی به هم متصل کند.

هنگامی که صداهای سالمندان به گفتار مردان بزرگسال ترجمه می‌شوند، ERR نسبتاً کمتر از مواردی بود که به گفتار زنان بالغ ترجمه می‌شوند. این ویژگی این گرایش به دلیل کمبود سخنان مردان مسن است. این نشان می‌دهد که عملکرد ترجمه صوتی بین یک جنس مؤثرتر است و انتظار می‌رود اگر مدل پیشنهادی برای هر جنسیت آموزش داده شود، بهترین نتایج حاصل می‌شود.

علاوه بر این، عملکرد تشخیص گفتار را با اعمال تبدیل صدا به واژگانی که در یادگیری A2AVT استفاده نمی‌شود، ارزیابی کرده‌اند. روش پیشنهادی را برای ۳۰ گفته سالمند تصادفی به کار بردند و نتایج در جدول ۴ ارائه شده است.

جدول ۴. عملکرد مدل A2AVT برای واژگان دلخواه.

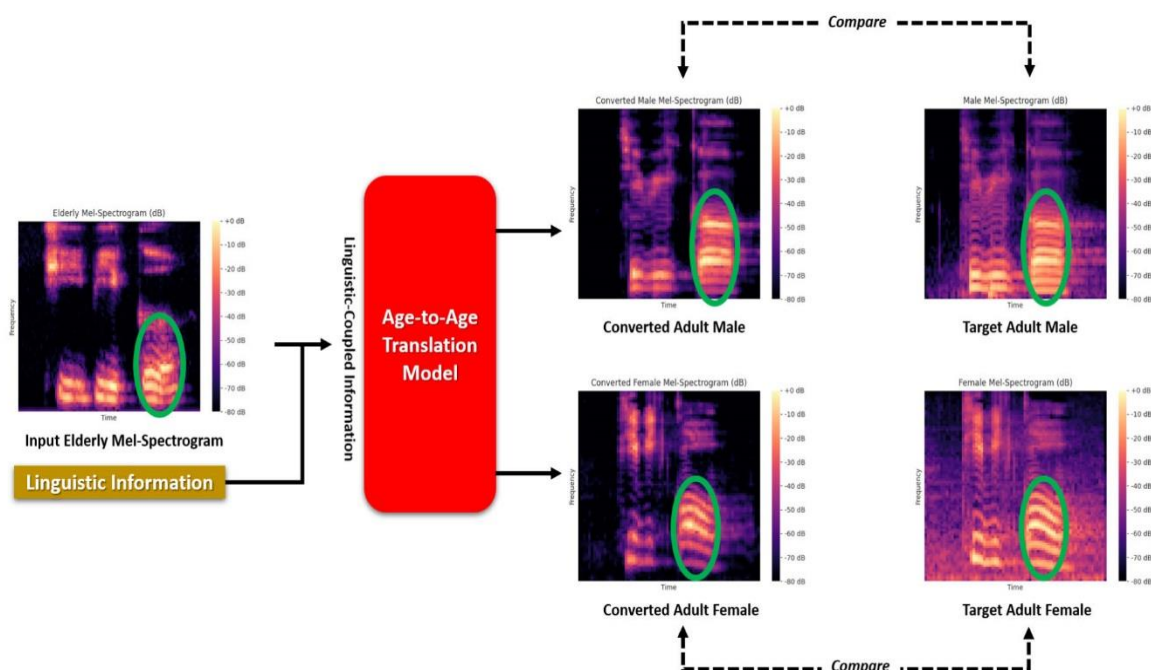
Method	Conversion Target	CER (%)	ERR (%)
No Conversion		30	
A2AVT + merged_Linguistic-ID	Adult Male	25	16.67
	Adult Female	12.5	58.33

جدول ۴ نشان می‌دهد که روش پیشنهادی برای واژگان دلخواه که به عنوان جفت تبدیل در یادگیری A2AVT استفاده نمی‌شوند، کار می‌کند.

در نتیجه، بهبود عملکرد تشخیص، اثربخشی A2AVT و روش اطلاعاتی همراه با زبان را تأیید می‌کند و نشان می‌دهد که روش پیشنهادی را می‌توان بدون هیچ گونه تغییری در سیستم ASR تجاری اتخاذ کرد.

اندازه گیری کیفیت صدای تبدیل شده:

برای ارزیابی کیفیت صدای تبدیل شده خود، از میانگین امتیاز نظر (MOS) و اندازه گیری اعوجاج Mel-cepstral (MCD) استفاده کنید.



شکل ۴. مقایسه طیف‌نگار Mel ورودی سالمندان، نتایج تبدیل شده و طیف‌نگارهای Mel بزرگسالان هدف

در تنظیمات ارزیابی MOS، ۶۰ آزمودنی بومی کره‌ای را که سن آنها بین ۲۰ تا ۳۹ سال است، انتخاب شده است. ارزیابی در یک محیط آرام انجام شد و به شرکت‌کنندگان پیشنهاد شد که از هدست استفاده کنند تا روی قضاوت عینی تمرکز کنند. شرکت‌کنندگان عمدتاً دو مرحله کیفیت صدای تبدیل شده را به دست آوردند. جنبه اول طبیعی بودن است. معیار امتیاز برای درجه طبیعی بودن بستگی به میزان طبیعی بودن صداهای تبدیل شده در مقایسه با متن داده شده دارد. جنبه دوم شباهت است. ارزیابی میزان شباهت به این بستگی دارد که صداهای سالمندان تغییر شکل یافته با گفتار مردان و زنان بالغ چقدر شبیه است. شرکت‌کنندگان در مقیاس ۱-۵ امتیازی برای دو جنبه امتیاز گرفتند و نمره بالاتر بهترین است.

جدول ۵. ارزیابی MOS برای صدای تبدیل شده.

Task	Target		
	Male	Female	Average
Naturalness	4.19±0.17	4.63±0.11	4.41±0.15
Similarity	4.14±0.18	4.69±0.10	4.41±0.16

جدول ۵ نتایج ارزیابی MOS را شرح می‌دهد. نمره طبیعی بودن گفتار سالمندان تبدیل شده به بزرگسالان مرد 4.19 ± 0.17 و نمره شباهت 4.14 ± 0.18 است. نمره طبیعی بودن گفتار تبدیل شده از سالمند به زن 4.63 ± 0.11 و نمره شباهت 4.69 ± 0.10 است. مانند نتایج ASR، صداهای تبدیل شده دوم نمرات بالاتری نسبت به همتای خود می‌گیرند. میانگین کلی نمره طبیعی بودن 4.41 ± 0.15 و میانگین نمره شباهت 4.41 ± 0.16 است.

علاوه بر این، نتایج اندازه‌گیری MCD را با صدای بزرگسالان با استفاده از صدای اصلی سالمند و صدای تبدیل شده سالمند ارائه داده شده است. برای اندازه‌گیری MCD، به جای ضرایب مغزی فرکانس Mel، از طیف نگار Mel 80 بُعدی استفاده شده است. اگر دو صدا طول متفاوتی داشته باشند، بالشتک‌های صفر اضافه می‌شوند تا با طیف‌نگار Mel طولانی‌تر مطابقت داشته باشند. جدول ۶ نشان می‌دهد که صدای تبدیل شده سالمندان در مقایسه با صدای اصلی سالمندان به صدای بزرگسالان نزدیک‌تر است.

جدول ۶. نتایج اندازه‌گیری MCD با صدای بزرگسالان

MCD Pair	Distortion Value (dB)
Adult - Original Elderly	13.77
Adult - Converted Elderly	6.78

تحلیل و بررسی:

صداهاى سالمندان تبدیل شده به بزرگسالان را از طریق مدل پیشنهادی تجزیه و تحلیل شد. علاوه بر این، تجزیه و تحلیل شباهت کسینوس بین توالی زبانی-ID و توزیع توالی خوشه-ID زبانی همان گفته از دو گروه ارائه شده است.

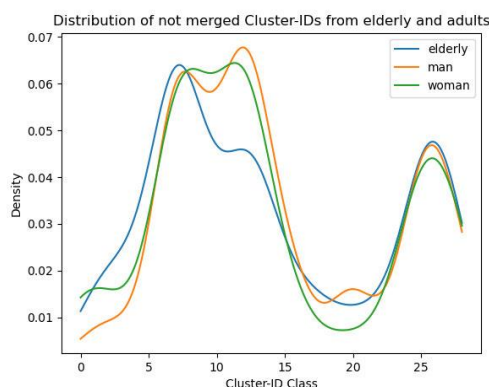
شکل ۴ نتایج ترجمه سن به سن را با استفاده از روش پیشنهادی نشان می‌دهد. طیف‌نگار مل سالمندان ورودی، نتایج تبدیل شده آن، و طیف‌نگار مل بزرگسالان هدف از راست به چپ نمایش داده می‌شوند. در میان دو طیف‌نگار Mel بالا، سمت چپ به ترتیب نشان‌دهنده نتیجه تبدیل شده و سمت راست صدای مرد بزرگسال هدف را نشان می‌دهد. در پایین، طیف‌نگارهای Mel نتیجه تبدیل شده و صدای زن بزرگسال هدف آن ارائه شده است.

همانطور که در شکل ۴ نشان داده شده است، شکل‌های هر طیف‌نگار Mel به رنگ سبز هستند. در مقایسه، در حالی که شکل‌های موجود در نتیجه تبدیل شده و طیف‌نگار Mel هدف دارای اشکال مشابهی هستند، طیف‌نگار Mel سالمندان ورودی با نتایج آن متفاوت است. به طور خاص، هر دو طیف‌نگار Mel هدف، شکل‌دهنده‌های افقی را نشان می‌دهند. طیف‌نگارهای Mel بزرگسالان ماده افزایش جزئی و کاهش تدریجی را نشان می‌دهند، در حالی که طیف‌نگاری مل اصلی شکل‌دهنده‌های موجی را نشان می‌دهد.

جدول ۷. شباهت کسینوس شناسه‌های خوشه‌ای از جنسیت‌ها و سنین مختلف

Group A	Group B	Cosine Similarity (A, B)	
		Not Merged	Merged
Adult Male	Adult Female	0.9549	0.9596
Elderly	Adult	0.8522	0.8833

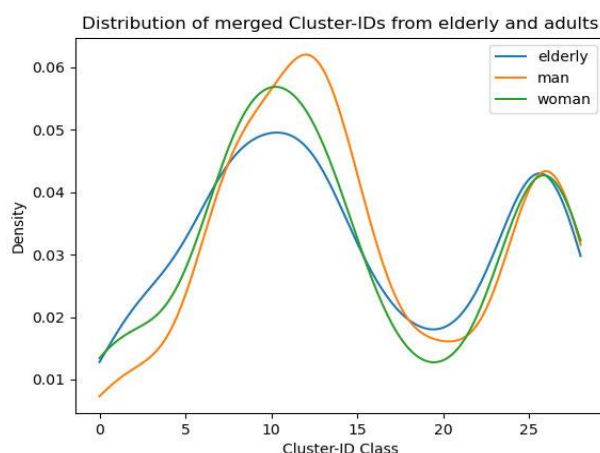
در اینجا، نتایج شباهت کسینوس بین دو دنباله زبانی-ID را گزارش شده است. همانطور که در جدول ۷ نشان داده شده است، به شباهت یکسانی برای دو جنس در گروه بزرگسالان دست یافته‌اند. حدس زده می‌شود که شباهت‌های کسینوس قابل مقایسه دو گروه بزرگسال به سرعت گفتاری مشابه آن‌ها نسبت داده می‌شود. از سوی دیگر، استفاده از تکنیک ادغام نتایج متفاوتی را در مقایسه با صداهاى سالمندان و بزرگسالان نشان می‌دهد. شناسه‌های ادغام شده شباهت بیشتری نشان داده‌اند، بنابراین، روش ادغام می‌تواند ویژگی‌های همگنی را در ویژگی‌های گفتاری از دو گروه سنی مختلف ایجاد کند.



شکل ۵. توزیع شناسه های خوشه ای سالمندان و بزرگسالان بدون فرآیند ادغام به دست آمده است.

علاوه بر این، دو تابع توزیع به دست آمده با و بدون استفاده از فرآیند ادغام را تجزیه و تحلیل کردند. همانطور که در شکل ۵ توضیح داده شد، شناسه های زبانی همه گویندگان یک شکل پاکت مشابه را ترسیم می کنند. توابع توزیع نشان می دهد که روش خوشه بندی واج شناسی پیشنهادی به خوبی واج ها را از دو گروه سنی ترسیم می کند. در حالی که بیشتر کلاس ها پاکت های مشابهی را نشان می دهند، شکل ۹ تا ۱۴ کلاس سالمندان با بزرگسالان متفاوت است. دلیل آن تفاوت در سرعت صحبت افراد مسن و بزرگسال است.

شکل ۶ توابع توزیع شناسه های خوشه ادغام شده را نشان می دهد. به لطف تکنیک ادغام، می توان توزیع های مشابه بیشتری را نسبت به شکل ۵ مشاهده کرد. بنابراین، تکنیک ادغام می تواند با استفاده از روش خوشه بندی واج شناسی پیشنهادی با تکنیک ادغام، پل سازی زبانی دو گروه را فراهم کند.



شکل ۶. توزیع شناسه های خوشه ای سالمندان و بزرگسالان به دست آمده با فرآیند ادغام

مقایسه عملکرد با توجه به تعداد خوشه های زبانی:

تأثیر تعداد خوشه های زبانی، K را برای مدل $A2AVT + merged_Linguistic-ID$ نشان داده شده است. وقتی K به ترتیب ۸، ۱۶، ۳۲ و ۶۴ باشد، CER صداهای تبدیل شده با هم مقایسه می شوند.

جدول ۸. مقایسه عملکرد با توجه به تعداد خوشه های زبانی

K	Conversion Target	CER (%)	ERR (%)
8	Adult Male	20.75	28.00
	Adult Female	24.50	3.69
16	Adult Male	20.31	29.53
	Adult Female	15.45	39.27
32	Adult Male	19.21	33.34
	Adult Female	14.35	43.59
64	Adult Male	34.22	-18.74
	Adult Female	24.06	5.42

در جدول ۸ بهترین عملکرد زمانی به دست می آید که 32 K باشد. این نشان می دهد که خوشه بندی بر اساس مجموعه های واج کره ای به عنوان اطلاعات زبانی برای روش A2AVT موثر است.

نتیجه گیری و کارهای آینده:

در این مقاله، بر روی بهبود عملکرد سیستم تجاری ASR متمرکز شده که در تشخیص صداها دورتر مانند افراد مسن ضعیف است. برای این منظور، اطلاعات مرتبط با زبان را از طریق روش خوشه بندی واج شناسی بدون نظارت معرفی شده است و ترجمه صوتی سن به سن را با استفاده از اطلاعات همراه با زبان برای بهبود عملکرد تشخیص گفتار برای سالمندان پیشنهاد شده است. از این نظر، روش پیشنهادی روش انطباق جایگزین است که می تواند در مقابل هر سیستم ASR تجاری یا باز قرار گیرد. اثربخشی روش پیشنهادی A2AVT را نشان داده شد و اطلاعات مرتبط با زبان را از طریق بهبود دقت تشخیص گفتار از سیستم تجاری ASR ادغام شد.

به عنوان کارهای آینده، از جمله تشخیص گفتار سالمندان، همچنین روشی برای تشخیص گفتار کودکان، افراد ناتوان با لکنت زبان، و همچنین لهجه ها و گویش های مختلف به کار برده می شود. علاوه بر این، روش را طوری طراحی کردند که بتوان در شرایط غیر زوجی با استفاده از عبارات جملات مختلف عمل کرد.

- [1] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop Deep Learn. Speech Recognit. Rel. Appl.*, Vancouver, BC, Canada, 2009, vol. 1, no. 9, p. 39.
- [2] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pretrained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30_42, Jan. 2011.
- [3] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2012, pp. 131_136.
- [4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82_97, Nov. 2012.
- [5] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8614_8618.
- [6] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580_4584.
- [7] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*. [Online]. Available: <http://arxiv.org/abs/1412.5567>
- [8] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1764_1772.
- [9] Y. Miao, M. Gowayyed, and F. Metze, "EESSEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 167_174.
- [10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 577_585.
- [11] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960_4964.
- [12] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, and J. Chen, "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 173_182.
- [13] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4835_4839.
- [14] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5884_5888.
- [15] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4774_4778.
- [16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 5036_5040.
- [17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*. [Online]. Available: <http://arxiv.org/abs/1807.03748>
- [18] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 3465_3469.
- [19] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Proc. Interspeech*, Sep. 2019, pp. 146_150.
- [20] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2041_2053, Dec. 2019.
- [21] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Prasad, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end ASR: From supervised to semi-supervised learning with modern architectures," 2019, *arXiv:1911.08460*. [Online]. Available: <http://arxiv.org/abs/1911.08460>
- [22] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *Proc. Int. Conf. Learn. Represent.*, 2019.
- [23] A. T. Liu, S.-W. Yang, P.-H. Chi, P.-C. Hsu, and H.-Y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6419_6423.
- [24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020.
- [25] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020, *arXiv:2010.10504*. [Online]. Available: <http://arxiv.org/abs/2010.10504>
- [26] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-Y. Lee, "Audio albert: A lite bert for self-supervised learning of audio representation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 344_350.
- [27] J. G. Wilpon and C. N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 349_352.
- [28] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. EUROSPEECH*, 1997.
- [29] S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson, "Recognition of elderly speech and voice-driven document retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 1999, pp. 145_148.

- [30] A. Potamianos, A. Potamianos, S. Narayanan, and S. Member, "Robust recognition of children's speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 603_616, Nov. 2003.
- [31] S. Kwon, S.-J. Kim, and J. Y. Choeh, "Preprocessing for elderly speech recognition of smart devices," *Comput. Speech Lang.*, vol. 36, pp. 110_121, Mar. 2016.
- [32] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220_239, May 2017.
- [33] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2506_2510.
- [34] F. Biadisy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," in *Proc. Interspeech*, Sep. 2019, pp. 4115_4119.
- [35] J.-W. Kim, H.-Y. Jung, and M. Lee, "Vocoder-free end-to-end voice conversion with transformer network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1_8.
- [36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5998_6008.
- [37] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech*, Oct. 2020, pp. 4676_4680.
- [38] R. Liu, X. Chen, and X. Wen, "Voice conversion with transformer network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, p. 7759.
- [39] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 5410_5419.
- [40] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, "LRSpeech: Extremely low-resource speech synthesis and recognition," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2802_2812.
- [41] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," 2016, *arXiv:1611.01144*. [Online]. Available: <http://arxiv.org/abs/1611.01144>
- [42] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," 2016, *arXiv:1611.00712*. [Online]. Available: <http://arxiv.org/abs/1611.00712>
- [43] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006, pp. 369_376.
- [44] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyriannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006_4010.
- [45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Tech-nol.*, Minneapolis, MI, USA, vol. 1, Jun. 2019, pp. 4171_4186. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423>
- [46] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: <http://arxiv.org/abs/1607.06450>
- [47] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236_243, Apr. 1984.
- [48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.