بهبود تولیدکننده‌های گفتار سازه‌ای و پیوندی با الهام از عملکرد فشرده‌سازهای گفتار

نوشین مقصودی و محمدمهدی همایونپور

چکیده: این مقاله به پیاده‌سازی و بهبود بخش تولید گفتار از یک سیستم تبدیل متن به گفتار می‌پردازد. با این هدف، روش تولید پیوندی مبتنی بر روش جمع هم‌پوشان با پریود صحیح و تحریک چندباندی و روش تولید سازه‌ای برای زبان فارسی پیاده‌سازی شده و به‌منظور بهبود در کیفیت خروجی از قدرت فشرده‌سازهای گفتاری استفاده شده است. به‌عبارت دیگر در ایده مطرح‌شده در این مقاله برای رفع مشکلات تولیدکننده‌های گفتار از فشرده‌سازهای موجود استفاده شده است. به این ترتیب که از فشرده‌ساز STRAIGHT[1] برای هموارسازی طیفی در تولیدکننده پیوندی و از فشرده‌ساز پیشگویی خطی با تحریک ترکیبی در تولید گفتار به روش سازه‌ای بهره گرفته شده است. نتایج ارزیابی‌ها نشان داده که استفاده از این فشرده‌سازها به کاهش ناپیوستگی‌ها در تولیدکننده پیوندی و افزایش معیارهای قابلیت فهم و طبیعی‌بودن در تولیدکننده سازه‌ای کمک کرده است.

کلید واژه: **STRAIGHT، تحریک چندباندی، روش پیوندی، روش سازه‌ای، فشرده‌ساز.**

۱- مقدمه

در سال‌های اخیر و با ورود رایانه به عرصه‌های متفاوت زندگی، تسهیل ارتباط بین انسان و ماشین با استفاده از گفتار به‌عنوان کانال ارتباطی، علاقه زیادی را به سوی خود جلب نموده است. دو تکنولوژی اصلی مورد نیاز برای پردازش گفتار، بازشناسی گفتار و تولید[2] گفتار می‌باشند. پیشرفت‌های اخیر در تولید گفتار منجر به توسعه تولیدکننده‌هایی با هوشمندی قابل توجه شده اما در بسیاری از موارد ضعف این سیستم‌ها از نظر کیفیت صدا، طبیعی‌بودن و وضوح گفتار تولیدشده غیر قابل اغماض است. تبدیل متن به گفتار در کاربردهای زیادی می‌تواند مفید واقع شود. اولین کاربرد که در آن از تبدیل متن به گفتار استفاده شد، نرم‌افزار خواندن متون برای نابینایان بوده است. امروزه این سیستم‌ها در کاربردهای دیگری مانند آموزش الکترونیکی، ترجمه گفتار به گفتار، خواندن پیام مناسب در مکان‌های عمومی، سیستم‌های راهنمای تلفنی پیشرفته و خواندن نامه‌های الکترونیکی نیز استفاده می‌شوند. با صرف نظر از روش‌های مکانیکی و الکتریکی اولیه که در تولید گفتار استفاده می‌شوند، دو روش مهم برای تولید گفتار، روش پیوندی[3] [۱] و روش سازه‌ای[4] [۲] می‌باشند. این روش‌ها با توجه به ویژگی‌های خاص خود می‌توانند در کاربردهای متفاوت مورد استفاده واقع شوند. تولیدکننده سازه‌ای با وجود سابقه طولانی آن، به‌دلیل آن که نرم‌افزار آن بسیار کم‌حجم بوده و از سرعت تولید بالایی برخوردارست، امروزه همچنان مورد توجه و استفاده بوده و بسیاری از سیستم‌های تبدیل متن به گفتار مدرن در کنار روش‌های جدیدتر تولید گفتار همچون تولید مبتنی بر انتخاب واحد و تولیدکننده هارمونیک – نویزی، از امکان تولید سازه‌ای نیز برخوردارند. همچنین این روش تولید گفتار به دلایلی که اشاره شد، برای استفاده در تلفن‌های همراه مناسب می‌باشد. اما زمانی که کیفیت و طبیعی‌بودن گفتار خروجی از اهمیت بیشتری برخوردار باشد، روش تولید پیوندی که از واحدهای به‌دست آمده از گفتار طبیعی بهره می‌برد، شانس بیشتری برای انتخاب دارد. در این مقاله به پیاده‌سازی و بهبود عملکرد دو روش تولید گفتار به روش سازه‌ای و روش پیوندی مبتنی بر جمع هم‌پوشان هم‌زمان با گام (TPMBROLA)[5] با پریود صحیح و تحریک چندباندی برای زبان فارسی پرداخته شده است. برای بهبود کارایی این تولیدکننده‌ها از ایده‌های موجود در فشرده‌ساز[6]های گفتاری شامل فشرده‌ساز MELP[7] و فشرده‌ساز STRAIGHT استفاده کردیم. فشرده‌سازها جهت فشرده‌سازی سیگنال در سیستم‌های مخابراتی و در دو مرحله آنالیز و تولید سیگنال گفتار عمل می‌کنند. به این ترتیب که در مرحله آنالیز در گیرنده، پارامترهایی از سیگنال گفتار به‌دست آمده و به‌طور مناسب کد شده و فشرده‌سازی انجام می‌شود و سپس در مرحله تولید در گیرنده این پارامترها کدگشایی شده و برای دست‌یابی به سیگنال اولیه استفاده می‌شوند. از فشرده‌سازهای مطرح می‌توان به روش‌های پیشگویی خطی با تحریک کد[8] [۳]، پیشگویی خطی با تحریک ترکیبی[9] [۴]، فشرده‌سازهای مبتنی بر تحریک چندباندی[10] [۵] و STRAIGHT [۶] اشاره کرد. روش STRAIGHT که قدرت زیادی در مدیریت و کنترل پارامترها دارد، در مرحله آنالیز، پارامترهای منبع شامل اطلاعات گام گفتار و مشخصات سازه‌های گفتار را با استفاده از یک طیف هموارشده استخراج می‌کند. این پارامترها پس از اعمال تغییرات مطلوب، توسط مؤلفه تولید STRAIGHT به گفتار با مشخصات دلخواه تبدیل می‌شوند [۶]. در این مقاله برای رفع ناهمواری‌های طیفی در روش تولید پیوندی، از شیفت سازه‌ها با توجه به اطلاعات حاصل از خروجی مرحله آنالیز در STRAIGHT استفاده شده و

این مقاله در تاریخ ۲۹ شهریور ماه ۱۳۸۹ دریافت و در تاریخ ۲۸ مهر ماه ۱۳۹۰ بازنگری شد.

نوشین مقصودی، آزمایشگاه پردازش هوشمند سیگنال و گفتار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران (email: n_maghsoodi@aut.ac.ir).
محمدمهدی همایونپور، آزمایشگاه پردازش هوشمند سیگنال و گفتار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران (email: homayoun@aut.ac.ir).

1. Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram
2. Synthesis
3. Concatenative Synthesis
4. Formant Synthesis
5. True Priode MulitiBand Resynthesis OverLap Add
6. Vocoder
7. Mixed Excitation Linear Predictive
8. Code Excited Linear Predictive
9. Mixed Excitation
10. Multiband Excitation

سپس سیگنال جدید با توجه به ویژگی‌های تغییریافته، در مرحله تولید STRAIGHT تولید می‌شود. تاکنون از STRAIGHT در تولید گفتار با اهدافی چون افزایش تقویت سازه‌های گفتار [۷] و اعمال تغییرات نوایی مطلوب [۸] و [۹] استفاده شده است. همچنین با هدف بهبود کیفیت تولیدکننده سازه‌ای، سیگنال تحریک ساده با سیگنال تحریک ترکیبی، مشابه آنچه که در فشرده‌ساز MELP صورت می‌گیرد، جایگزین شده است. در روش پیشنهادی ابتدا دادگانی از اطلاعات قدرت واکداری و پارامترهای واکدار و بی‌واکی باندهای فرکانسی مربوط به واج‌های واکدار مشابه آنچه در کدکننده MELP صورت می‌گیرد به‌دست آمده و ذخیره می‌شود و سپس در زمان تولید، اطلاعات مورد نیاز از دادگان استخراج شده و مشابه آنچه که در کدگشای MELP صورت می‌گیرد، برای تولید سیگنال تحریک مناسب استفاده می‌شود.

در ادامه در بخش ۲ روش‌های تولید گفتار شرح داده شده و سپس در بخش ۳ چگونگی پیاده‌سازی سیستم‌های تولید گفتار و اعمال پیشنهادها مورد بررسی قرار می‌گیرد. نتایج آزمایش‌ها و ارزیابی‌های انجام‌شده نیز در بخش ۴ بیان گردیده و در خاتمه بخش ۵ به نتیجه‌گیری و ارائه پیشنهاد برای کارهای آینده می‌پردازد.

۲- معرفی سیستم‌های تولید گفتار

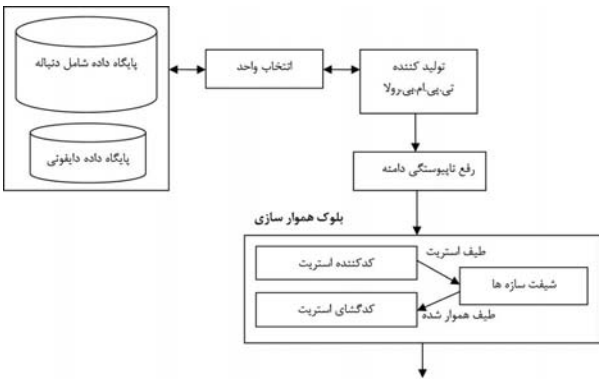
در روش‌های پیوندی برای تولید گفتارِ متناظر با ورودی، دنباله‌ای از واحدهای مناسب از دادگان از پیش ضبط شده گفتار استخراج شده و سپس این واحدها به هم متصل می‌شوند. در صورتی که اتصال به‌صورت خام و ساده انجام شود، در محل اتصال واحدها، ناپیوستگی نامطلوبی حس خواهد شد که منجر به افت کیفیت گفتار می‌شود. از طرفی با این روش امکان ایجاد نوا و احساسات در گفتار تولیدشده امکان‌پذیر نخواهد بود. لذا روش‌های پیوندی دیگری پدید آمدند که اساس آنها ایجاد تغییر و اصلاح روی واحدهای دادگان برای رفع ناپیوستگی‌ها و ایجاد نوای مطلوب در گفتار است. مطرح‌ترین این روش‌ها مبتنی بر الگوریتم‌های جمع هم‌پوشان هم‌زمان با گام (PSOLA)[1] هستند [۱۰]. ایده اصلی روش PSOLA در حوزه زمان یا TDPSOLA[2] این است که بتوان با حذف یا تکرار برخی از از پنجره‌های هم‌زمان با گام، پریود گام گفتار واکدار و سرعت زمانی سیگنال را به نحو مطلوب تغییر داد. در یک نگاه کلی، در این روش واحدهای گفتاری از دادگان بازیابی می‌شوند، سپس بنابر عوامل گوناگونی نظیر نوع واج‌های مجاور، نحوه ادای کلمه، جایگاه کلمه در جمله و نوع جمله (خبری، پرسشی و ...) تغییرات زمانی و گامی روی آن واحد انجام می‌شود. در انتها همه واحدهای اصلاح‌شده را به روش جمع هم‌پوشان با هم ترکیب می‌کنیم تا سیگنال گفتار مطلوب حاصل شود. مشکل این روش از محل اتصال بین دو واحد ناشی می‌شود. به عبارت دیگر وقتی دو واحد از کلمات مختلف و از بافت‌های متفاوت استخراج می‌شوند، ممکن است در هنگام اتصال از نظر پریود گام، فاز و ویژگی‌های طیفی فاقد تطابق لازم باشند که این امر منجر به ایجاد ناهمواری‌های قابل تشخیص در گفتار خروجی می‌شود [۱۱]. برای غلبه به مشکلات یادشده دوتوآ در [۱۲] روشی ارائه کرده که در آن از تولید مجدد بخش‌های واکدار در پایگاه داده قطعات گفتار استفاده می‌شود تا عدم تطابق در محل اتصال دو واحد به حداقل برسد. نتیجه، روش تولید بهبودیافته MBROLA[3] نامیده شده که از راندمان محاسباتی الگوریتم پایه PSOLA به همراه انعطاف‌پذیری روش MBE[4] استفاده می‌کند. در این روش سعی شده تا طی تولید مجدد، عدم تطابق فاز و پریود گام با اعمال پریود گام ثابت و مقداردهی مجدد فاز در ابتدای هر پریود گام تا حد زیادی برطرف شود. یک مزیت مهم تولید مجدد این است که با اعمال گام ثابت، مشخص‌کردن نشانگرهای گام در سیگنال به راحتی امکان‌پذیر بوده و این نشانگرها می‌توانند فاصله نسبی دلخواه نسبت به هم داشته باشند. اما در این روش نیز علاوه بر این که به علت ثابت‌گرفتن فاز و گام در فریم‌ها، خروجی همراه با صدای وزوز[5] خواهد بود، استفاده از واحدهای بزرگ‌تر از دایفون امکان‌پذیر نیست. از آنجایی که در واحدهای بزرگ‌تر از دایفون غالباً تغییرات زیادی در گام وجود دارد، فرض ثابت‌بودن گام در واحدهای گفتاری دادگان، فرضی نادرست است. لذا در این‌گونه دادگان‌ها ثابت در نظر گرفتن طول فریم نیز مشکل‌زا خواهد بود. از این رو در رویکردی جدید که TPMBROLA نامیده شده، در هر دوره تناوب، ابتدا فاز نخستین هارمونیک محاسبه می‌شود و سپس، پردازش‌های بعدی به‌طور هم‌زمان با این فاز انجام می‌شود. با این کار عملیات مرحله تولید نیز هم‌زمان می‌شود. در این الگوریتم برخلاف MBROLA، تنها تغییری که روی گام انجام می‌شود این است که دوره تناوب گام به یک عدد صحیح گرد می‌شود [۱۳].

روش دیگر پیاده‌سازی شده بر مبنای مدل تولید سازه‌ای است. اساس کار روش سازه‌ای مدل‌سازی مجرای گفتار به‌عنوان فیلتر از یک سو و ساختن سیگنال تحریک از سوی دیگر است. با در دست داشتن این دو پارامتر- مثلاً برای هر واج از گفتار- می‌توان سیگنال گفتار را تولید نمود. در واقع این کار به نوعی شبیه‌سازی مکانیزم تولید گفتار در انسان است. از آنجایی که سیگنال گفتار کاملاً ناایستان است، فرض اساسی این‌گونه تولیدکننده‌ها این است که در بازه‌های بسیار کوتاهی به نام فریم، سیگنال گفتار را می‌توان ایستان در نظر گرفت. لذا در این روش‌ها پردازش در فریم‌هایی به طول ۵ تا ۳۰ میلی‌ثانیه انجام می‌شود و در ابتدای هر فریم می‌بایست پارامترهای مدل به‌روز شوند. در این روش با توجه به دنباله واجی ورودی به سیستم و با استفاده از قوانین از پیش تعریف شده، کانتور گام[6] و اطلاعات سازه‌ای هر فریم استخراج شده و به مؤلفه تولیدکننده سازه‌ای داده می‌شود. یک نمونه از این تولیدکننده سازه‌ای که از ۳۹ پارامتر کنترلی برای مدل‌سازی گفتار استفاده نموده و برای تولید واج‌ها از دو نوع پیکربندی سری و موازی تشدیدسازها[7] بهره می‌برد، تولیدکننده کلات[8] است که بسیاری از تولیدکننده‌های تجاری نیز به‌عنوان الگوریتم پایه از این تولیدکننده استفاده می‌کنند [۲]. در تولیدکننده معرفی‌شده در این مقاله نیز برای پیاده‌سازی تولیدکننده سازه‌ای بر مبنای روش کلات عمل شده است.

۳- پیاده‌سازی

در این بخش پیاده‌سازی روش‌های تولید مطرح‌شده در بخش قبل و راه‌کارهای اعمال‌شده برای بهبود کیفیت خروجی آنها شرح داده می‌شود.

1. Pitch Synchronous OverLap and Add
2. Time Domain Pitch Synchronous OverLap and Add
3. MulitiBand Resynthesis OverLap and Add
4. MultiBand Excitation
5. Buzziness
6. Pitch Contour
7. Resonator
8. Klatt



شکل ۱: ساختار کلی تولیدکننده پیوندی پیاده‌سازی شده.

۳-۱ استفاده از STRAIGHT در تولیدکننده پیوندی

در پیاده‌سازی تولیدکننده پیوندی می‌توان سه گام اصلی در نظر گرفت: تهیه دادگان، پیاده‌سازی تولیدکننده گفتار و پردازش‌های ثانویه جهت بهبود کیفیت. در این مقاله برای تهیه دادگان که نقش زیادی در کیفیت نهایی گفتار دارد، در کنار واحد دایفون که می‌تواند حجم کم پایگاه داده را تضمین کند، از واحدهای غیر یکنواخت استفاده شده است. بنابراین یک بخش از مجموعه داده اولیه، متنی است که بتوان از آن دایفون‌ها را استخراج نمود. در تهیه این متون باید توجه داشت که همه دایفون‌های ممکن در زبان فارسی که ۹۶۶ دایفون می‌باشند، در متن حضور داشته باشند. اما با توجه به این که با داشتن واحدهای دایفون تعداد نقاط اتصال زیادی برای ساخت هر کلمه در زمان تولید گفتار خواهیم داشت و با علم به این مطلب که افزایش تعداد نقاط اتصال، پیوستگی گفتار نهایی و در نتیجه کیفیت آن را تهدید می‌کند، این ایده استفاده شده که علاوه بر دایفون از بعضی واحدهای تولید بزرگ‌تر از دایفون برای تولید بخش‌هایی از گفتار که در زبان رخداد بالایی دارند، استفاده شود. به‌عنوان مثال دنباله "می" در افعال مضارع و یا پسوندهایی چون "ایم"، "اند" و کلماتی چون "بود"، "است"، "شد" و مانند آن در زبان فارسی بسیار استفاده می‌شوند. استفاده مستقیم از این‌گونه واحدهای بزرگ‌تر از دایفون، تأثیر به‌سزایی در بهبود کیفیت تولید گفتار و طبیعی‌تر شدن آن خواهد داشت. به این ترتیب بخش دوم از تهیه متون اولیه، انتخاب جملاتی است که شامل دنباله‌های حروف پرتکرار در زبان فارسی که پس از یک سری آمارگیری‌ها به‌دست آمده، باشند. بنابراین ابتدا نیاز به انجام یک سری پردازش‌های آماری روی زبان با استفاده از یک پیکره بزرگ که بتواند اطلاعات آماری زبان را به خوبی منعکس کند، خواهیم داشت. با استفاده از آمارگان در دسترس که شامل تعداد تکرارهای همه دنباله‌های حروف از یک پیکره بزرگ می‌باشد، دنباله‌های پرتکرار حروف استخراج می‌شوند. واحدهای غیر یکنواخت مذکور شامل تعداد محدودی از دنباله‌های حروف پرتکرار زبان فارسی با طول حداقل ۲ و حداکثر ۹ می‌باشد که در جدول ۱ تعداد دنباله حروف متناظر با هر طول ذکر شده است. در مرحله بعد و پس از تهیه پیکره‌ای به اندازه کافی بزرگ برای استخراج این دنباله حروف پرتکرار، جملات منتخب از این پیکره که شامل دنباله حروف پرتکرار هستند به همراه جملات شامل دایفون‌ها توسط یک گوینده با صدای مناسب، در محیط عاری از نویز و تا حد امکان به‌صورت یکنواخت و به دور از تغییرات شدید گام ضبط شد.

این پایگاه داده ترکیبی در شکل ۱ که شمای کلی سیستم را نشان می‌دهد، مشاهده می‌شود. در زمان تولید گفتار برای انتخاب از پایگاه داده، اولویت با واحدهای با طول بیشتر خواهد بود که درباره نحوه این انتخاب

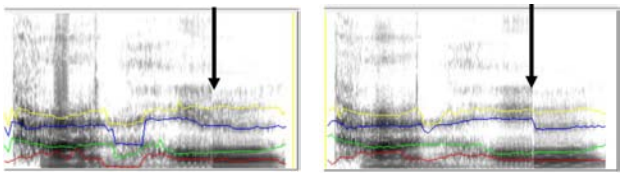
جدول ۱: تعداد دنباله‌های حروف با طول‌های مختلف

تعداد	طول دنباله
۹۰	۲
۱۸۵	۳
۲۷۰	۴
۳۵۰	۵
۲۴۴	۶
۱۵۶	۷
۷۰	۸
۴۵	۹

در ادامه بیشتر توضیح داده می‌شود. بنابراین به‌کار بردن دنباله حروف بزرگ‌تر با اولویت بیشتر در زمان تولید گفتار، باعث کاهش تعداد نقاط اتصال در گفتار تولید شده و در نتیجه با کاهش ناپیوستگی‌ها به بهبود کیفیت خروجی کمک می‌کند. در تهیه پایگاه داده برای تأمین تنوع در ویژگی‌های گامی و طول واحدها، از واحدهای پایه دایفون و دنباله‌های پرتکرار حروف، بیش از یک نمونه نگهداری شده است. این کار موجب می‌شود که در زمان تولید گفتار نیاز به روشی برای انتخاب بهترین نمونه از هر واحد داشته باشیم. این مؤلفه در شکل ۱ نشان داده شده است.

مؤلفه انتخاب واحد به این صورت عمل می‌کند که با شروع از ابتدای دنباله واجی ورودی سعی می‌کند از واحدهای با طول بیشتر با اولویت بیشتری استفاده کند و در صورت نیافتن آن به تدریج از طول واحد هدف می‌کاهد. در نهایت اگر واحد مورد نظر در بین دنباله‌های پرتکرار یافت نشد، از دایفون‌ها استفاده خواهد شد. به این ترتیب جمله‌ای با حداقل تعداد ناپیوستگی‌ها خواهیم داشت. در مورد هر واحد همه نمونه‌های موجود از آن واحد از پایگاه داده استخراج می‌شود. یافتن بهترین دنباله واحدها از بین این نمونه‌ها متناظر با یافتن کوتاه‌ترین مسیر در یک گراف جهت‌دار می‌باشد. به این ترتیب که گراف به سطوح مشخص تقسیم می‌شود و سطح اول متناظر با همه واحدهای موجود در پایگاه داده است که می‌توانند متناظر با اولین واحد از دنباله واج‌های ورودی برای تولید باشند و این روند برای سطوح بعدی تعمیم می‌یابد [۱۴]. در این گراف هر رأس از یک سطح به همه رئوس سطح بعد وصل می‌شود و یال‌های بین رئوس، هزینه پیوند بین دو واحد متوالی از گراف را نشان می‌دهد. همچنین هر رأس در بردارنده هزینه هدف می‌باشد. به‌عبارت دیگر انتخاب هر رأس از گراف، هزینه هدف متناظر با آن رأس را به هزینه مسیر رئوسی که تاکنون انتخاب شده‌اند، اضافه می‌نماید. مسیری با کمترین هزینه از این گراف با استفاده از الگوریتم جستجوی ویتربی به‌عنوان دنباله واحدهای بهینه محاسبه شود. در یافتن مسیر بهینه، به‌عنوان تابع هزینه هدف[1] از زیرهزینه‌های طول بازه زمانی مطلوب[2] و متوسط فرکانس گام مطلوب و به‌عنوان تابع هزینه پیوند[3] از زیرهزینه‌های اختلاف طیفی، اختلاف فرکانس گام و اختلاف انرژی در فریم‌های مرزی استفاده شده است. در صورتی که هزینه کل مسیر بهینه به‌دست آمده بیش از یک حد آستانه مشخص باشد، مرحله یافتن دنباله واحدها مجدداً اجرا می‌شود تا مجموعه‌ای با ترکیب واحدهای متفاوت استخراج شود. در این مرحله، ترکیبی با تعداد واحدهای برابر با اولین مقدار ممکن بعد از تعداد حداقل واحدها جستجو می‌شود. حد آستانه استفاده‌شده در این مرحله، با توجه به

1. Target Cost
2. Desired Duration
3. Join Cost



(الف) (ب)

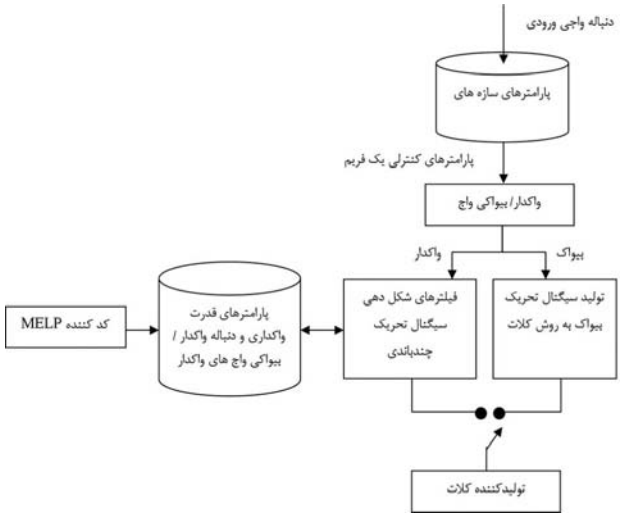
شکل ۲: اسپکتروگرام یک سیگنال، (الف) قبل و (ب) بعد از هموارسازی با استریت.

مقدار متوسط هزینه مجموعه‌ای از جملات خروجی با کیفیت مطلوب تعیین شده است. هزینه‌ها با توجه به طول هر جمله نرمال‌سازی شده‌اند. پس از به‌دست آمدن ترکیب جدید مجدداً مسیر بهینه به‌دست می‌آید. در صورتی که هزینه نهایی مسیر دوم نسبت به مسیر اول بهتر بود، این مسیر به‌عنوان مسیر بهینه انتخاب می‌شود و در غیر این صورت مسیر قبلی در جایگاه مسیر بهینه باقی می‌ماند. البته باید در نظر داشت که به علت سربار زمانی حاصل از تکرار الگوریتم بهتر است تعداد دفعات این تکرار را محدود نمود. پس از مرحله یافتن دنباله واحدهای بهینه، نوبت به تولید واحدهای انتخاب‌شده به روش جمع هم‌پوشان هم‌زمان با گام می‌رسد. به این ترتیب خروجی اولیه تولید می‌شود.

پس از تولید خروجی، مشکل اصلی وجود ناپیوستگی‌های محسوس در گفتار تولیدشده است. بخشی از این ناپیوستگی‌ها می‌تواند ناشی از اختلاف دامنه یک واحد نسبت به واحدهای اطرافش باشد. برای رفع این حالت، دامنه متوسط هر واحد با دامنه متوسط واحدهای اطرافش مقایسه می‌شود. در صورتی که نسبت دامنه متوسط واحد نسبت به هر دو واحد اطرافش از یک حد آستانه مشخص کمتر و یا از حد آستانه مشخص دوم بیشتر باشد، دامنه نمونه‌های واحد مذکور در ضریبی که با نسبت یادشده متناسب است، ضرب می‌شود. به این ترتیب نوعی هنجارسازی در مواردی که دامنه یک واحد نسبت به واحدهای همسایه تفاوت بارزی دارد، ایجاد می‌شود. در مرحله بعد ناپیوستگی‌های طیفی باید هموارسازی شود. در بخش مقدمه گفته شد که یکی از پیشنهادهای ارائه‌شده در این مقاله برای ایجاد بهبود در تولیدکننده، شیفت‌دادن سازه‌ها به‌صورت مستقل از هم برای رفع اختلاف بین سازه‌های دو واحد مجاور می‌باشد. برای اعمال این پیشنهاد، از قابلیت موجود در فشرده‌ساز STRAIGHT که قابلیت تغییر در مقدار یک سازه خاص را فراهم می‌کند، استفاده شده است. در روش STRAIGHT طیف سیگنال گفتار توسط پوش آن بازنمایی می‌شود. سپس پوش طیف هم در جهت فرکانس و هم در جهت زمان هموار می‌شود. در مرحله بعد با نرخ نمونه‌برداری نسبتاً بالایی (در هر ۱ میلی‌ثانیه) در هر دو محور زمان و فرکانس عمل نمونه‌برداری انجام می‌شود. البته این کار روی فریم‌های زمان کوتاه سیگنال که ۱۰۲۴ نمونه دارند، انجام می‌شود تا بتوان از تبدیل فوریه سریع استفاده نمود. این بازنمایی از آن جهت که هموار و سپس با نرخ بالایی نمونه‌برداری شده، می‌تواند منشأ گستره وسیعی از تغییرات بلادرنگ در مراحل بازسازی مجدد سیگنال گفتار باشد. این روش بار محاسباتی بالایی را می‌طلبد و شاید بتوان همین بار محاسباتی بالا را ضعف عمده آن دانست. در مرحله آنالیز STRAIGHT، اطلاعات مربوط به فاز استخراج نمی‌شود اما در مرحله بازسازی مجدد آن، برای جلوگیری از وزوزی‌شدن خروجی، سیگنال گفتار بازسازی‌شده از یک سری فیلترهای تمام‌گذر[1] عبور داده می‌شود. برای ساخت سیگنال تحریک نیز از روش تحریک چندباندی (مجموع وزن‌دار بخش‌های نویز و قطار ضربه) استفاده می‌شود [۶]. روش پیاده‌سازی شده در این مقاله به این صورت عمل می‌کند که ابتدا طیف هموارشده به روش STRAIGHT که خروجی مرحله آنالیز STRAIGHT است، محاسبه می‌شود. این طیف که به‌صورت یک ماتریس سه‌بعدی (شامل تغییرات فرکانس، زمان و دامنه) می‌باشد، به تابعی که وظیفه شیفت‌دادن سازه‌ها را دارد، داده می‌شود. پس از شیفت‌دادن سازه‌ها، بخش تولید STRAIGHT طیف جدید را به خروجی تبدیل می‌کند. تابع شیفت سازه‌ها به این صورت عمل می‌کند که در هر نقطه پیوند، در صورتی که اختلاف هر یک از سازه‌های دو فریم حول نقطه پیوند از یک حد آستانه بیشتر بود، آن سازه مشخص در ۴ فریم از سمت چپ و ۴ فریم از سمت راست دچار شیفت در فرکانس در جهت مخالف می‌شود تا مقدار فرکانس سازه‌ها به هم نزدیک شود. مقدار این شیفت از فریم مرزی به سمت فریم‌های دورتر کاهش می‌یابد تا از تغییر ناگهانی مقادیر فرکانس سازه‌ها جلوگیری شود. به این ترتیب با شیفت مناسب سازه‌ها در این روش می‌توان ناپیوستگی طیفی در نقاط پیوند را کنترل نمود. همچنین برای افزایش سرعت با توجه به خاصیت مل که مبتنی بر قدرت شنوایی در انسان است، عمل شده است. به این ترتیب که با توجه به این خاصیت که قدرت سیستم شنوایی انسان در تشخیص تمایز بین دو تون فرکانسی با افزایش فرکانس کاهش می‌یابد، نتیجه می‌گیریم که عدم تطابق طیفی در سیگنال خروجی تولیدکننده در سازه‌های با فرکانس پایین‌تر از نظر شنونده محسوس‌تر بوده و در نتیجه چشم‌پوشی از هموارسازی سازه‌های با فرکانس بالاتر چندان تأثیرگذار نخواهد بود. بنابراین برای تعیین حد آستانه اختلاف بین فرکانس سازه‌ها در مرز دو واحد به این نکته توجه می‌شود که برای سازه‌های با فرکانس پایین‌تر مقدار حد آستانه کمتر و برای سازه‌های با فرکانس بالاتر مقدار بزرگ‌تر حد آستانه مناسب بوده و به افزایش سرعت فرآیند هموارسازی کمک می‌کند. اسپکتروگرام یک سیگنال آزمایشی قبل و بعد از هموارسازی توسط STRAIGHT در شکل ۲ نشان داده شده است. نزدیک‌شدن سازه سوم واحدهای حول نقطه پیوند در شکل کاملاً مشهود است. سایر سازه‌ها به‌دلیل اختلاف کم هر سازه با سازه متناظر در قطعه کناری نیاز به هموارسازی نداشته‌اند. ملاحظه می‌شود که با این روش می‌توان هر سازه را به‌صورت مستقل از سایر سازه‌ها تغییر داد. همین امر علاوه بر امکان مدیریت بیشتر سازه‌ها، از ایجاد نتیجه‌ای بدتر از پیش از هموارسازی جلوگیری می‌کند. در حالی که در روش‌های دیگر در برخی از مواقع ممکن است به نتیجه‌ای نامطلوب‌تر از سیگنال اولیه برسیم که همین موضوع نقطه ضعف این روش‌ها محسوب می‌شود [۱۵]. البته باید توجه داشت که پس از آنالیز و تولید توسط STRAIGHT، مانند هر روش آنالیز و تولید گفتار دیگر، کمی افت در طبیعی‌بودن گفتار خواهیم داشت که به‌علت از دست دادن بخشی از اطلاعات سیگنال پس از استفاده از فشرده‌ساز می‌باشد.

برای رفع تغییرات ناگهانی گام نیز از روش معرفی‌شده در [۱۳] که با جابه‌جایی گام از ناپیوستگی‌های کانتور گام می‌کاهد، استفاده شده است. این الگوریتم سعی می‌کند حالت بهینه‌ای را بیابد که با حداقل‌کردن یک تابع هزینه و با جابه‌جایی مقدار گام برخی از واحدها، ناپیوستگی‌های کانتور گام کاهش یابد. تابع هزینه به‌صورت مجموع ناپیوستگی‌های موجود پس از جابه‌جایی عمودی منحنی گام و با در نظر گرفتن امتیاز منفی برای حداقل‌کردن تعداد جابه‌جایی‌ها محاسبه می‌شود. امتیاز منفی یادشده برابر با جمع تعداد جابه‌جایی‌ها تقسیم بر طول واحدهاست. این امتیاز منفی تضمین می‌کند که احتمال جابه‌جایی واحدهای کوچک‌تر بیشتر از واحدهای بزرگ‌تر باشد. به این ترتیب تابع هزینه زمانی حداقل

1. All-Pass Filter



شکل ۳: نمودار بلوکی تولیدکننده سازه‌ای پیاده‌سازی شده.

می‌شود که با کمترین تعداد جابه‌جایی بیشترین کاهش ناپیوستگی را داشته باشیم.

۳-۲ استفاده از تحریک چندباندی در تولیدکننده سازه‌ای

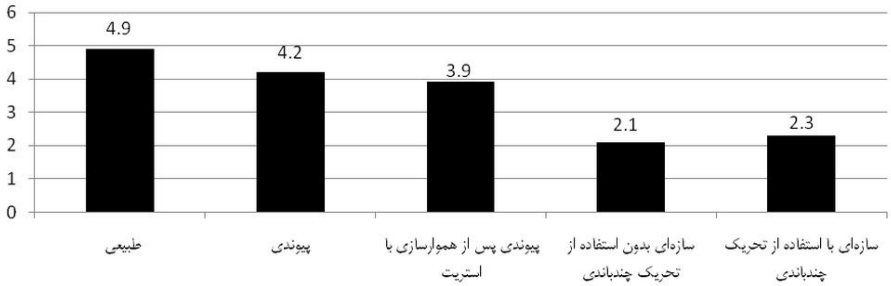
توسعه یک تولیدکننده سازه‌ای قابل تفکیک به دو مرحله است: مرحله استخراج پارامترهای سازه‌ها و مرحله پیاده‌سازی تولیدکننده گفتار. در مرحله اول تحلیل صوتی واج‌های مختلف انجام می‌شود. پارامترهای صوتی شامل فرکانس، پهنای باند و دامنه سازه‌هاست که تعیین این پارامترها برای سه سازه اول باید با دقت بیشتر انجام شود. در محاسبه این پارامترها، واج‌ها از متون از پیش ضبط شده استخراج شده و اثر حاصل از کنار هم قرار گرفتن واج‌ها کنار هم تحلیل می‌شود و در نهایت مقادیر سازه‌ها با استفاده از روش معرفی‌شده در [۱۶] محاسبه می‌شود. به این ترتیب که پس از به‌دست آمدن قطب‌ها، قطب‌هایی که مقادیر حقیقی دارند، حذف می‌شوند. سپس مجموعه فرکانس و پهنای باند به‌دست آمده با توجه به قوانین پذیرش قطب، پردازش می‌شوند. اولین قانون بررسی می‌کند که مقادیر فرکانس به‌دست آمده از حداکثر فرکانس مجاز برای آن سازه بیشتر نباشد و قانون دوم محدوده پهنای باند را بررسی می‌کند. در صورتی که پهنای باند به‌دست آمده به‌ازای یک قطب از حد آستانه مربوط به پهنای باند بیشتر باشد، آن قطب حذف می‌شود. به این ترتیب سعی می‌شود از پذیرش قطب‌های ناشی از اثر نویز به‌عنوان یک سازه جلوگیری شود. پس از طی این روال ممکن است تعداد سازه‌های قابل قبول از تعداد سازه‌های مورد نیاز کمتر باشد. در این صورت الگوریتم استخراج قطب‌ها مجدداً با افزایش مرتبه ضرایب پیشگویی خطی اجرا می‌شود و این روال تا زمانی که تعداد سازه‌های مورد نیاز استخراج شوند، تکرار می‌شود.

مدل‌های منبع- فیلتر که تولید سازه‌ای بر مبنای آن عمل می‌کند، از مدل بسیار ساده‌شده‌ای برای منبع صوت استفاده می‌کنند. به‌طوری که از قطار ضربه به‌عنوان سیگنال واکدار و از نویز برای حالت بی‌واک بهره می‌برند. اگرچه این مدل مناسب و ساده است ولی کاهش کیفیت خروجی تولیدکننده که ناشی از استفاده از این مدل ساده سازی شده می‌باشد، غیر قابل اجتناب است. به‌ویژه حالت وزوزی ایجادشده در بخش‌های واکدار گفتار خروجی به‌عنوان یک مشکل عمده جلب توجه می‌کند. چون در حقیقت یک سیگنال واکدار در همه باندهای فرکانسی حالت پریودیک کامل ندارد. به این ترتیب وقتی تنها از قطار ضربه بدون حضور نویز استفاده می‌کنیم، آنچه در نهایت تولید می‌شود با گفتار طبیعی تفاوت محسوسی خواهد داشت. با توجه به آنچه گفته شد، در این مقاله از مجموعه‌ای از باندهای نویز و قطار ضربه به‌عنوان منبع صوت صدای واکدار استفاده شده است. برای این منظور از سیستم کدکننده MELP که با بهره‌گیری از این نوع سیگنال تحریک، توانایی تولید گفتاری شبیه به سیگنال گفتار واقعی را دارد، استفاده شده است. سیستم کدکننده MELP یک نوع سیستم کدکننده گفتار است که برای برطرف‌نمودن برخی از اشکالات موجود در روش‌های کدکردن پیشین توسعه یافته است. ایده اصلی مورد استفاده در این نوع سیستم کدکننده، استفاده از یک سیگنال تحریک ترکیبی به‌عنوان سیگنال ورودی به فیلتر تولید گفتار است. کلمه ترکیبی به این نکته اشاره دارد که از ترکیبی از باندهای فرکانسی طیف پریودیک و نویز برای ساخت سیگنال تحریک استفاده می‌گردد. به این ترتیب با پرداخت هزینه محاسباتی بیشتر، می‌توان به سیگنال تحریکی که تقریب دقیق‌تری از سیگنال تحریک گفتار طبیعی است، دست یافت. در این روش ابتدا یک بانک فیلتر، گفتار ورودی را به پنج باند تقسیم نموده و از هر باند برای یافتن ضریب قدرت واکداری متناظر با آن باند استفاده می‌کند. این پارامترها به‌عنوان اطلاعات مرحله کدکننده ذخیره شده و سپس در زمان کدگشایی با توجه به این اطلاعات باندهای سیگنال تحریک تولید می‌شوند و در نهایت گفتار اولیه بازسازی می‌شود. علاوه بر این، سیستم MELP با استفاده از سیگنال نیمه واکدار در مرز بین فریم‌های واکدار و بی‌واک از یک روش جدید درون‌یابی برای هموارنمودن حالات گذر بین فریم‌ها استفاده می‌کند. ویژگی دیگر روش MELP این است که با درون‌یابی پارامترهای استخراج‌شده از گفتار به هموارسازی انتقال بین فریم‌ها کمک می‌کند [۳]. بنابراین تولیدکننده پیاده‌سازی شده در این مقاله بر مبنای تولیدکننده سازه‌ای کلات است با این تفاوت که سیگنال تحریک در واج‌های واکدار به‌صورت چندباندی و به روش تولید MELP تولید می‌شود. اکنون در صورتی که بخواهیم روش MELP را با یک تولیدکننده گفتار که قابلیت تولید هر ورودی دلخواه را دارد، ترکیب کنیم، نیاز داریم که مرحله آنالیز برای به‌دست آوردن پارامترهای مورد نیاز یک بار انجام شده و در یک دادگان ذخیره شود تا در هر بار تولید گفتار پارامترهای مفید استخراج شده و در اختیار تولیدکننده قرار بگیرد. پارامترهایی که در مرحله آنالیز استخراج می‌شوند، پارامترهای قدرت واکداری و دنباله واکدار/بی‌واکی مربوط به پنج باند واج‌های واکدار می‌باشد. لازم به یادآوری است که با توجه به این که روش MELP در حالت استاندارد برای کاهش تعداد بیت ارسالی به گیرنده از نرخ نمونه‌برداری ۸۰۰۰ هرتز استفاده می‌کند، برای افزایش کیفیت در سیستم تولید گفتار پیاده‌سازی شده این نرخ به ۱۶۰۰۰ هرتز تغییر داده شد که در نتیجه آن پهنای باند هر باند فرکانسی نیز به دو برابر تغییر یافت. با این توضیحات نمودار بلوکی تولیدکننده سازه‌ای مشابه شکل ۳ خواهد بود.

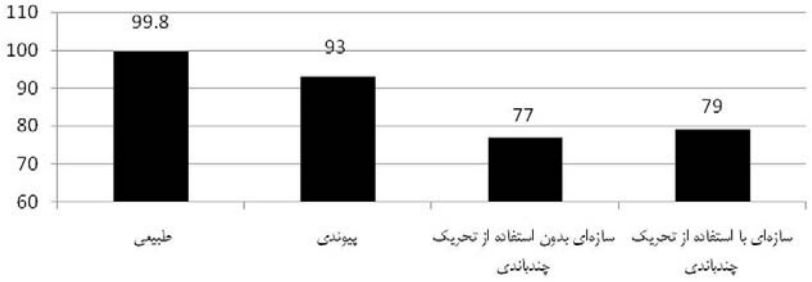
۴- نتایج ارزیابی

یک سیستم تبدیل متن به گفتار می‌تواند از جنبه‌های قابلیت درک، طبیعی و روان‌بودن[1] (عدم ناپیوستگی) مورد ارزیابی قرار گیرد. هر یک از این جنبه‌ها متناسب با کاربرد سیستم ممکن است اهمیت بیشتری داشته باشند. بیشتر روش‌های ارزیابی مطرح و مورد استفاده روش‌های نظری[2] و وابسته به آگاهی و قضاوت انسان می‌باشند و از میان آنها دو روش

1. Fluidity
2. Subjective



شکل ۴: مقایسه روش‌های تولید گفتار پیاده‌سازی شده از نظر طبیعی‌بودن گفتار خروجی بر اساس معیار میانگین امتیازات نظرخواهی.



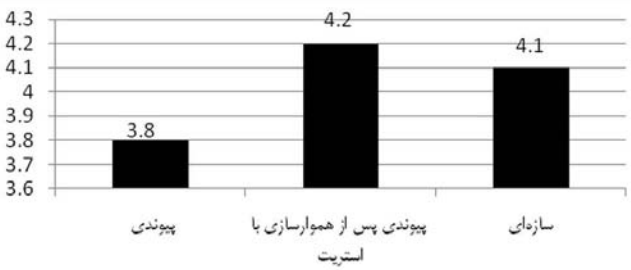
شکل ۵: مقایسه روش‌های تولید گفتار از نظر قابلیت فهم با توجه به تست تشخیص قافیه.

معمول‌تر که برای ارزیابی تولیدکننده‌ها مورد استفاده قرار گرفته، روش‌های تست تشخیص قافیه (DRT)[1] و تست میانگین امتیازات نظرخواهی (MOS)[2] هستند. روش تست تشخیص قافیه قابلیت فهم یک واحد خاص از خروجی را ارزیابی می‌کند. این تست از یک مجموعه از کلمات برای ارزیابی مفهوم‌بودن هم‌خوان‌ها در خروجی و در نتیجه قدرت سیستم در تولید هر هم‌خوان استفاده می‌کند. به این ترتیب که مجموعه آموزشی شامل مجموعه‌ای از کلمات است که هر جفت در هم‌خوان ابتدایی با هم متفاوت هستند. هر بار شنونده یک کلمه از این مجموعه را می‌شنود و مشخص می‌کند که این خروجی کدام یک از جفت کلمات بوده است. در نهایت نتیجه ارزیابی با محاسبه نرخ خطا در پاسخ‌های داده‌شده به‌دست می‌آید. این آزمون در حالت استاندارد شامل ۹۶ جفت کلمه است که برای زبان فارسی توسط گروه پردازش گفتار مرکز تحقیقات الکترونیک دانشگاه شریف تدوین شده است [۱۷]. در روش تست میانگین امتیازات نظرخواهی، ۲۰ جمله که تا حد امکان ترکیب‌های متفاوت واج‌ها را شامل باشند، انتخاب شده و با استفاده از هر یک از روش‌های متفاوت تولید گفتار ضبط می‌شوند. سپس شنونده به جمله خروجی با توجه به سطوح کیفیت از پیش تعیین شده‌ای امتیاز می‌دهد. از این امتیازها که در حالت استاندارد با ۵ سطح بد، ضعیف، متوسط، خوب و عالی مشخص شده‌اند، میانگین‌گیری می‌شود و خروجی به‌عنوان نتیجه ارزیابی مورد استفاده قرار می‌گیرد. برای این ارزیابی‌ها از ۲۰ شنونده دعوت شد.

در ارزیابی طبیعی‌بودن خروجی تولیدکننده‌ها در هر حالت تولید و برای هر شنونده پخش شد. شکل ۴ نتایج این تست را برای روش‌های تولید گفتار در حالت با استفاده و بدون استفاده از فشرده‌سازها در مقایسه با حالت طبیعی نشان می‌دهد. همان‌طور که در شکل ملاحظه می‌شود، هموارسازی از میزان طبیعی‌بودن جملات خروجی می‌کاهد. این کاهش کیفیت پس از طی مراحل آنالیز و تولید روش STRAIGHT و به‌علت پردازش‌های انجام‌شده روی سیگنال گفتار اصلی ایجاد می‌شود. کاهش کیفیت و طبیعی‌بودن گفتار با زنگ‌دار شدن آن نمود پیدا می‌کند. می‌توان گفت در انتخاب روش هموارسازی باید بین این کاهش کیفیت و میزان بهبود در عدم ناپیوستگی تعادل برقرار کرد و مناسب‌ترین حالت را برگزید. همچنین با توجه به شکل می‌توان دریافت که روش پیوندی افت کمتری از نظر طبیعی‌بودن داشته در حالی که امتیاز کسب‌شده توسط روش سازه‌ای به مراتب کمتر است. می‌دانیم که هر چه در یک روش تولید گفتار بیشتر از گفتار طبیعی استفاده شود و این گفتار کمتر دستخوش تغییر شود، افت کیفیت در خروجی کمتر خواهد بود. بنابراین از آنجا که در روش تولید پیوندی از واحدهای از پیش ضبط شده گفتار استفاده می‌شود در حالی که در روش سازه‌ای گفتار به‌صورت مصنوعی مدل‌سازی می‌شود، واضح است که گفتار حاصل از تولید پیوندی کیفیت بهتری خواهد داشت. همچنین در این شکل می‌بینیم که روش سازه‌ای در حالت استفاده از تحریک ترکیبی به روش MELP اندکی بهبود کیفیت دارد که این به‌علت مدل‌سازی دقیق‌تر سیگنال تحریک در این روش می‌باشد. البته این بهبود کمتر از حد انتظار بوده است.

در آزمایش بعدی با استفاده از تست تشخیص قافیه قابلیت فهم خروجی روش‌ها مورد ارزیابی قرار گرفت. نتایج ارزیابی در شکل ۵ نشان می‌دهد که شنونده‌ها در هر یک از روش‌های تولید گفتار برخی از هم‌خوان‌ها را درست تشخیص نداده‌اند که این نشان‌دهنده ضعف روش‌های تولید گفتار در مدل‌سازی این هم‌خوان‌ها می‌باشد. با توجه به نتایج مشهود است که روش تولید پیوندی به‌طور متوسط قادر به مدل‌سازی مناسب نزدیک به ۹۳٪ از کلمات آزمایش‌شده می‌باشد. از بین هم‌خوان‌هایی که نادرست تشخیص داده شده‌اند، بیشتر هم‌خوان‌های از نوع انسدادی به‌ویژه /p/، /b/ و /d/ به چشم می‌خورند. به نظر می‌رسد با توجه به کوتاه‌بودن مرحله انفجار در این هم‌خوان‌ها برچسب‌زنی این واج‌ها گاهی با دقت بالا انجام نشده و در نتیجه در زمان تولید گفتار طول این مرحله از حالت عادی نیز کوتاه‌تر شده به‌طوری که شنونده قادر به تشخیص درست آن نبوده است. اما در روش سازه‌ای میزان خطای تشخیص به مراتب بیشتر بوده است که این موضوع بیشتر به ضعف روش در مدل‌سازی برخی از واج‌ها که پارامترهای نزدیک به هم دارند، برمی‌گردد. در دو حالت استفاده از تحریک چندباندی و عدم استفاده از آن

1. Diagnostic Rhyme Test
2. Mean Opinion Score



شکل ۶ مقایسه روش‌های تولید گفتار پیاده‌سازی شده از نظر عدم ناپیوستگی.

نیز مشاهده می‌شود که تفاوت چندانی وجود ندارد و بهبود نه چندان قابل توجه در حالت استفاده از تحریک چندباندی نیز به‌علت مدل‌سازی بهتر برخی از انسدادی‌های واکدار توسط این روش به نظر برخی از شنونده‌ها بوده است.

در نمودار شکل ۶ نیز روش پیوندی بدون هموارسازی با خروجی به‌دست آمده پس از اعمال هموارسازی، روش پیوندی بدون استفاده از دنباله حروف و روش سازه‌ای از نظر روان‌بودن مقایسه شده است. با توجه به نمودار درمی‌یابیم در روش سازه‌ای که گفتار به‌صورت مصنوعی مدل‌سازی می‌شود، به‌علت تغییر تدریجی پارامترها در فریم‌های کوچک و درون‌یابی در مرز واج‌ها، ناپیوستگی محسوسی وجود ندارد و از این حیث این روش برتری خود را نسبت به روش پیوندی نشان می‌دهد. اما پس از هموارسازی انجام‌شده با استفاده از STRAIGHT و با کم‌شدن ناپیوستگی‌ها روان‌بودن گفتار خروجی بهبود یافته است. بنابراین با انتخاب یک روش هموارسازی مناسب و با تحمل کمی افت در طبیعی‌بودن گفتار می‌توان ناپیوستگی‌های گفتار خروجی را تا حدی کاهش داد.

۵- نتیجه‌گیری و پیشنهادها

تولید گفتار فرآیندی پیچیده شامل چندین مرحله است که هر گام خود نیازمند دانش نظری لازم برای انتقال به تولیدکننده می‌باشد تا در نهایت بتوان به خروجی نزدیک به گفتار طبیعی دست یافت. پس از تولید واحدها به روش جمع هم‌پوشان، مشکلی که جلب توجه می‌کند ناپیوستگی محسوسی است که در برخی از نقاط پیوند وجود دارد. برای رفع این مشکل استفاده از روش هموارسازی مناسب می‌تواند تأثیرگذار باشد. با اعمال روش شیفت‌دادن سازه‌ها در این مقاله و پس از یکپارچه‌سازی سیستم با فشرده‌ساز STRAIGHT که امکان مدیریت مستقیم پارامترهای سازه‌ها را می‌دهد، نتایج تست شنیداری نشان داد که استفاده از روش STRAIGHT و شیفت سازه‌ها در کاهش ناپیوستگی گفتار تولیدشده مؤثر بوده است، هرچند که منجر به کاهش طبیعی‌بودن گفتار شده است. به این ترتیب با انتخاب روش مناسب برای هموارسازی با تحمل هزینه کاهش طبیعی‌بودن گفتار می‌تواند تا حدودی روان‌بودن و پیوستگی در سیگنال خروجی را افزایش داد. بنابراین نکته‌ای که باید در نظر گرفت این است که روش‌های هموارسازی برای تغییر ویژگی‌های سیگنال گفتار به‌منظور هموارسازی نیاز به یک مرحله کدکردن و یک مرحله کدگشایی دارند که در طی این مراحل از طبیعی‌بودن گفتار کاسته می‌شود. بنابراین استفاده از روشی برای کدکردن و کدگشایی که منجر به حداقل کاهش کیفیت شود، اهمیت زیادی دارد.

در پیاده‌سازی روش تولید سازه‌ای می‌توان گفت استخراج ویژگی‌های هر واج از زبان، قلب پیاده‌سازی این نوع تولیدکننده است. استخراج این ویژگی‌ها که وابسته به نحوه ادای هر واج در زبان و تأثیر واج‌های مجاور بر آن می‌باشد، مرحله‌ای زمان‌بر و نیازمند دقت بالاست. مشکل عمده این روش که نتایج حاصل از ارزیابی نیز آن را تأیید کرده امتیاز پایین آن از نظر طبیعی‌بودن گفتار تولیدشده است. با این فرض که یکی از عوامل این عدم طبیعی‌بودن، ناکارآمدی روش تولید سیگنال تحریک است، از روش موجود در فشرده‌ساز MELP برای تولید سیگنال تحریک چندباندی استفاده شد. پس از پیاده‌سازی، نتایج ارزیابی بهبود در کاهش فلزی‌بودن[1] گفتار را نشان داده است. لازم به ذکر است که تست‌های شنیداری نشان داد که روش تولید سازه‌ای به‌علت مدل‌سازی گفتار به‌صورت مصنوعی و در نتیجه تسلط در تغییر پارامترها از یک فریم به فریم بعدی، از نظر روان‌بودن و پیوستگی امتیاز بیشتری نسبت به روش پیوندی کسب کرده است. به‌عنوان پیشنهادی برای ادامه کار می‌توان به راهکار ترکیب دو تولیدکننده پیوندی و سازه‌ای اشاره نمود. با این روش می‌توان ضعف یک روش را با قدرت روش دیگر پوشش داد. به‌عنوان مثال با توجه به مشکل ناپیوستگی در نقاط اتصال در تولیدکننده پیوندی، می‌توان با ترکیب آن با روش سازه‌ای در مرز بین دو واحد به‌جای تغییر در ویژگی‌های فریم‌های مرزی، فریم‌های جدیدی که ویژگی‌های مطلوب داشته و می‌توانند فاصله طیفی بین دو فریم مرزی را پر کنند، با استفاده از روش سازه‌ای تولید و در مرز واحدهای دارای ناپیوستگی محسوس اضافه نمود.

مراجع

[1] D. O'Shaughnessy, *Speech Communication: Human and Machine*, NewYork, Addison - Wesley, 1990.

[2] D. Klatt, "Software for a cascade/parallel formant synthesizer," *J. of Acoustical Society of America*, vol. 67, no. 3, pp. 971-995, Mar. 1980.

[3] P. Kabal, Code Excited Linear Prediction Coding of Speech at 4.8 kb/s. Technical Report 87-36, INRS-Telecommunications, University of Quebec, 1987.

[4] T. Moriya and M. Honda, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 4, pp. 242-250, Jul. 1986.

[5] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1223-1235, Aug. 1988.

[6] H. Kawahara, I. Masuda - Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch - adaptive time - frequency smoothing and an instantaneous - frequencybased F0 extraction," *Speech Communication*, vol. 27, no. 3, pp. 187-207, Apr. 1999.

[7] H. Zen and T. Toda, "An overview of Nitech HMM - based speech synthesis system for blizzard challenge 2005," in *Proc. of Interspeech*, pp. 93-96, Sep. 2005.

[8] H. Matsui and H. Kawahara, "Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system," in *Proc. 8th European Conf. on Speech Communication and Technology*, pp. 2113-2116, 1-4 Sep. 2003.

[9] T. Yonezawa, N. Suzuki, K. Mase, and K. Kogure, "Gradually changing expression of singing voice based on morphing," in *Proc. of Interspeech*, pp. 541-544, Sep. 2005.

[10] F. Charpentier and M. G. Stella, "Diphone synthesis using an overlap add technique for speech waveform concatenation," *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 2015-2018, Apr. 1986.

[11] T. Dutoit, An Introduction to Text - to - Speech Synthesis, The Netherlands: Kluwer, 1997.

[12] T. Dutoit and H. Leich, "MBR- PSOLA: text to speech synthesis based on a MBE re - synthesis of the segments database," *Speech Communication*, vol. 13, no. 3, pp. 435-440, Nov. 1993.

[13] B. Bozkurt, T. Dutoit, C. D'Alessandro, V. Pagel, and R. Prudon, "Improving quality of MBROLA synthesis for non-uniform units synthesis," in *Proc. IEEE Workshop Speech Synthesis*, pp. 7-9, 11-13 Sep. 2002.

[14] A. Mihelic and J. Zganec-Gros, "Efficient unit-selection in text-to-speech synthesis," in *Proc. of the 11th Int. Conf. on Text, Speech, and Dialogue*, pp. 411-418, 2008.

1. Metallic

[15] T. David, J. Chappell, and H. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication*, vol. 36, no. 3-4, pp. 343-373, Mar. 1998.

[16] R. C. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 2, pp. 129-134, Apr. 1993.

[۱۷] ح. قادری، *تولید گفتار فارسی از روی دنباله آوایی از طریق مدل‌کردن ساختار گویای انسان*، پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر، دانشگاه صنعتی شریف، ۱۳۷۷.

نوشین مقصودی در سال ۱۳۸۶ مدرک کارشناسی مهندسی کامپیوتر گرایش نرم افزار خود را از دانشگاه صنعتی شریف و در سال ۱۳۸۸ مدرک کارشناسی ارشد مهندسی کامپیوتر با گرایش هوش مصنوعی خود را از دانشگاه صنعتی امیرکبیر دریافت نمود. زمینه‌های علمی مورد علاقه نام‌برده عبارتند از: پردازش زبان طبیعی، پردازش گفتار و الگوریتم‌های تکاملی.

دکتر محمد مهدی همایونپور در سال ۱۳۳۹ در شهر شیراز متولد شد. تحصیلات تا مقطع دیپلم را در شهر شیراز سپری و دیپلم متوسطه خود را در سال ۱۳۵۸ دریافت کرد. وی تحصیلات خود در مقطع کارشناسی را در رشته مهندسی برق (الکترونیک) در دانشگاه صنعتی امیرکبیر (سال ۱۳۶۶)، کارشناسی ارشد را در رشته برق (مخابرات)، از دانشگاه خواجه نصیرالدین طوسی (سال ۱۳۶۹)، کارشناسی ارشد دوم خود را در زمینه فونتیک (سال ۱۳۷۴) در دانشگاه سوربون جدید در فرانسه و هم‌زمان دورهٔ دکترای خود را در دانشگاه پاریس ۱۱ در زمینه مهندسی برق (۱۳۷۴) به‌پایان رسانید. نام‌برده از سال ۱۳۷۴ در سمت عضو هیأت علمی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیر کبیر به تدریس و تحقیق مشغول می‌باشد. زمینه‌های تخصصی مورد علاقه ایشان شامل پردازش سیگنال‌های دیجیتال، بازشناسی گفتار و گوینده، تبدیل متن به گفتار، کدینگ گفتار، پردازش زبان طبیعی, تشخیص نفوذ در سیستم ها و شبکه های کامپیوتری، اتوماسیون صنعتی، چند رسانه‌ای و طراحی سخت افزار می‌باشد.