علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۵ شماره ۱ (الف)، ۱۳۸۶
صفحات ۱۸-۲۷
مقاله کوتاه

فارس بیان: سنتز کننده گفتار فارسی مبتنی بر روش انتخاب واحد

محمد مهدی همایونپور　　مجید نم نبات

دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

در سالهای اخیر روش سنتز با انتخاب واحد مبتنی بر دادگان بزرگ بدلیل تولید صحبت طبیعی با کیفیت مناسب، ساده بودن ایده و داشتن پتانسیل بالقوه بالا مورد توجه بسیاری از محققان قرار گرفته است. اجزای اصلی این روش یک دادگان بزرگ شامل نمونه های صوتی مختلف، دو معیار برای ارزیابی نمونه ها بنامهای هزینه هدف و اتصال و در انتها یک الگوریتم جستجو برای انتخاب بهترین نمونه ها می باشد. در این مقاله، ساختار یک سیستم سنتز با انتخاب واحد پیاده سازی شده برای زبان فارسی، بنام فارس بیان شرح داده می شود. بدین منظور زیرهزینه های تشکیل دهنده توابع هزینه، روشهای مختلف تعیین وزنهای زیرهزینه ها و الگوریتمهای پیرایش مورد استفاده برای کاهش فضای جستجوی نمونه ها در این پژوهش شرح داده می شوند. کیفیت خروجی سیستم به نحو قابل توجهی طبیعی می باشد. برای ارزیابی سیستم از تست MOS استفاده شده است که مقدار MOS برای معیار کیفیت کلی ۳،۸ بدست آمده است.

کلمات کلیدی: تبدیل متن به گفتار، سنتز پیوندی، سنتز با انتخاب واحد، زبان فارسی، دادگان سنتز، هزینه هدف، هزینه اتصال، جستجوی ویتربی، تست MOS.

۱- مقدمه

تاکنون روشهای سنتز متفاوتی همچون روشهای سنتز پیوندی با TD-PSOLA یا HNM، سنتز با فیلتر MLSA و ... ارائه شده است. سنتز با انتخاب واحد یکی از روشهای ارائه شده برای سنتز صحبت می باشد که از یک دادگان بزرگ شامل نمونه های مختلف استفاده می نماید. این روش سنتز از جمله روشهای سنتز پیوندی می باشد که در آن برای سنتز هر رشته آوایی، بهترین نمونه ها از میان نمونه های موجود انتخاب و گفتار خروجی با متصل نمودن این نمونه ها به یکدیگر ساخته می شود. این روش امروزه به دلیل تولید صحبت طبیعی با کیفیت مناسب، ساده بودن ایده و داشتن پتانسیل بالقوه بالا مورد توجه محققان قرار گرفته است. نامهای دیگر این روش سنتز به کمک دادگان بزرگ[1]، یا سنتز بروش دنبال هم چینی مجدد[2] می باشد. در سنتز پیوندی نرمال، بازای هر واحد یک نمونه صوتی وجود دارد، در حالیکه در این روش از یک دادگان بزرگ شامل نمونه های مختلف استفاده می شود. در سنتز پیوندی نرمال پروسه اصلی سیستم انجام عملیاتهای اصلاح نوا بر روی واحدها می باشد که باعث کاهش میزان طبیعی بودن خروجی سیستم می گردد. در حالیکه در سنتز با انتخاب واحد، مسئله مهم انتخاب واحدها و نمونه های مناسب صوتی از میان تعداد زیادی از واحدها و نمونه های صوتی میباشد. در این روش زمان سنتز و حجم محاسبات از جمله مسائل بحرانی هستند. در این حالت سعی میشود نمونه هایی انتخاب شوند که به رشته آوایی هدف تا حد زیادی نزدیک باشند و تا آنجا که ممکن است از انجام الگوریتمهای اصلاح نوا روی نمونه ها اجتناب می شود.

این روش سنتز ابتدا با استفاده از واحدهای غیر هم اندازه در سیستم $ATR-\nu-TALK$ برای زبان ژاپنی ارائه گردید [۱]. آقایان بلک و کمپل[3] با توسعه این روش و استفاده از این روش برای سنتز زبان انگلیسی در سیستم CHATR خروجی با کیفیت بسیار طبیعی تولید نمودند [۲، ۳]. محققان AT&T با ترکیب بخش NLP از سیستم فستیوال [۴، ۵] و بخش DSP از سیستم CHATR و بهینه سازی این بخشها سیستم تبدیل متن به گفتار خود را ارائه نمودند که ارزیابی های مختلف حاکی از قرار گرفتن این سیستم در زمرهٔ بهترین سیستمهای تبدیل متن به گفتار برای زبان انگلیسی می باشد [۶، ۷]. در سالهای اخیر این روش سنتز مورد توجه بسیاری از محققان قرار گرفته است و جنبه های مختلفی از آن مورد بررسی و بهینه سازی قرار گرفته است که از آن جمله می توان به ارائه الگوریتمهای بهینه برای کوچک سازی دادگان و پیرایش آن [۸، ۹، ۱۰] و یا خوشه بندی آن [۱۱]، متن انتخاب شده برای دادگان [۱۲] ، افزایش سرعت سیستم سنتز [۱۳] اشاره نمود. علاوه بر این کارهای زیادی برای بهینه سازی توابع هزینه شامل هزینه اتصال [۱۴] و هزینه هدف [۱۵، ۱۶، ۱۷، ۱۸] و ارائه یک معیار کمی با ضریب همبستگی مناسب با ارزیابی های شنوایی [۱۹] انجام شده است مرجع [۲۰] برای زبان ژاپنی با بهره گیری از دادگانهای بسیار بزرگ (در حدود ۱۰-۲۰ ساعت) سیستمهای تبدیل متن به گفتار با کیفیت خروجی بسیار طبیعی ارائه نموده است.

در این مقاله ساختار موتور سنتز با انتخاب واحد پیاده سازی شده برای زبان فارسی شرح داده می شود. در این سیستم سعی شده است علاوه بر پیاده سازی الگوریتم اولیه انتخاب واحد، روند توسعه این روش نیز مورد توجه قرار داده شود و نتایج حاصل از بهینه سازی بخشهای مختلف نیز مورد توجه قرار بگیرد.

در بخش ۲ ، توصیف کلی سنتز با انتخاب واحد، در بخش ۳ دادگان مورد استفاده، در بخش ۴ ساختار توابع هزینه شرح داده شده اند. در بخش ۵ روشهای مختلف تعیین وزنها و در بخش ۶ الگوریتم جستجوی ویتربی و انواع الگوریتمهای پیرایش بیان شده است. نتایج حاصل از ارزیابی در بخش ۷ آورده شده است و بخش ۸ نیز به نتیجه گیری و ارائه پیشنهادات می پردازد

۲- توصیف کلی سنتز با انتخاب واحد

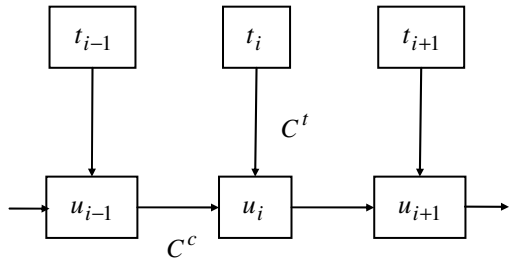
همانطور که اشاره شد این سیستم از یک دادگان بزرگ بهره می برد. در ابتدا لازم است که دادگان سیستم به نمونه های سازنده آن تقطیع و سپس برای هر نمونه یک بردار ویژگی استخراج شود. عموما ویژگیها به دو دسته ویژگیهای آوایی و سیگنالی (مربوط به نوا) تقسیم بندی می شوند. بطور مشابه برای رشته آوایی هدف که گفتار سنتز شده برای آن می خواهد تولید شود، لازم است که خصوصیات نوایی آن شامل کشش، انرژی و منحنی پیچ مشخص باشد. برای سنجش میزان مناسب بودن استفاده از یک نمونه دادگان بجای یک واحد هدف، دو تابع هزینه، هزینه هدف C^t و هزینه اتصال[۴] C^c تعریف میشود. هزینه هدف، نشان دهنده میزان شباهت یک نمونه از دادگان و یک واحد از رشته آوایی هدف می باشد. مقدار این هزینه با استفاده از مجموع خطی وزندار یکسری زیرهزینه مشخص می شود. هر زیر هزینه مربوط به یک ویژگی می باشد که مقدار آن با استفاده از تابع متریک (D_j^t) انتخاب شده برای آن ویژگی (j) محاسبه می شود. در صورتیکه t_i واحد هدف و u_i یک نمونه از دادگان، W_t^j وزن تابع فاصله ویژگی j و P تعداد زیرهزینه ها باشد. هزینه هدف بصورت زیر محاسبه میشود:

$$C^t = \sum_{j=1}^{P} w_j^t D_j^t(t_i, u_i) \qquad (۱)$$

هزینه اتصال، مشخص کننده میزان گسستگی طیف، انرژی و پیچ است که هنگام اتصال یک نمونه از دادگان به نمونه انتخاب شده قبلی آن بوجود می آید. برای محاسبه مقدار این هزینه، مشابه هزینه هدف، یک سری ویژگی نشان دهنده میزان گسستگی انتخاب و سپس برای هر ویژگی یک تابع فاصله (D^c) انتخاب میشود. تابع هزینه اتصال بصورت جمع خطی وزندار مقادیر فاصله محاسبه میشود. در صورتیکه u_{i-1} نمونه انتخاب شده قبلی و u_i یک نمونه از دادگان، q تعداد زیرهزینه های انتخاب شده و W_j^c را وزن تابع فاصله مربوط به ویژگی j در نظر بگیریم، فرمول محاسبه این هزینه بصورت زیر می باشد:

$$C^c = \sum_{j=1}^{q} w_j^c D_j^c(u_{i-1}, u_i) \qquad (۲)$$

در شکل ۱ مفهوم انواع هزینه شرح داده شده است. بعنوان یک حالت خاص، در صورتیکه u_{i-1} و u_i در دادگان نمونه های مجاور هم باشند، چون اتصال میان واحدهای مجاور صورت میگیرد، هزینه اتصال صفر در نظر گرفته میشود. این شرط امکان انتخاب نمونه های مجاور از دادگان و بواقع انتخاب نمونه های غیر هم اندازه را فراهم میکند.



شکل ۱- مفهوم انواع هزینه ها

البته تاکنون پژوهشهای مختلفی برای یافتن یک فرم مناسبتر همانند استفاده از مجذور میانگین مربعات زیرهزینه ها (RMS) [۲۱] برای محاسبه هزینه ها بجای فرم جمع خطی وزندار صورت گرفته است. در صورتیکه رشته آوایی هدف را زنجیره ای از واحدهای هدف بصورت $t_1^n = (t_1, ..., t_n)$ در نظر بگیریم. وظیفه پروسه انتخاب واحد، پیدا کردن مجموعه ای بهینه از نمونه های دادگان $u_1^n = (u_1, ..., u_n)$ از میان تمامی مجموعه های موجود می باشد که بیشترین شباهت را به رشته آوایی هدف داشته باشند. برای هر مجموعه از نمونه های دادگان هزینه کل برابر با مجموع هزینه های هدف و اتصال آنها بصورت زیر تعریف میشود:

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \qquad (۳)$$

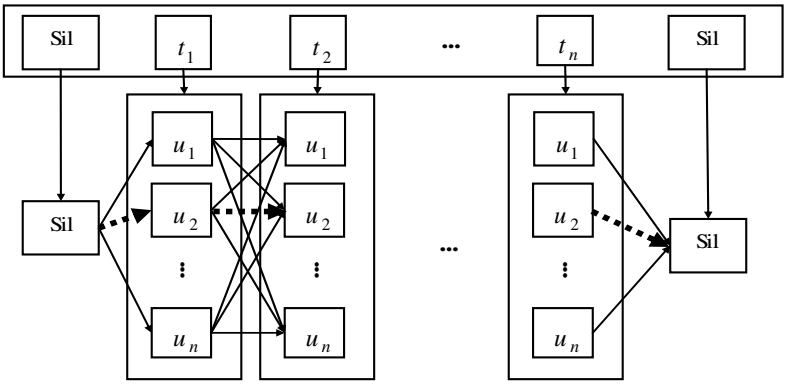
نماد S مشخص کننده سکوت و $C^c(S, u_1)$ و $C^c(u_n, S)$ هزینه اتصال واحد اول و آخر را به سکوت مشخص می نمایند. الگوریتم انتخاب واحد سعی می کند مجموعه واحدهایی $\bar{u}_n^1$ را انتخاب نماید که مقدار کل هزینه برای آنها مینیمم شود.

$$\bar{u}_n^1 = \min_{u_1, ..., u_n} C(t_1^n, u_1^n) \qquad (۴)$$

انتخاب بهترین مجموعه با ساختن یک گراف از تمام نمونه های دادگان صورت می گیرد. چون بطور بالقوه ممکن است هر دو نمونه از دادگان بهم متصل شوند، اتصالات گراف کامل در نظر گرفته میشود. در این گراف هزینه عبور از هر یال برابر با هزینه اتصال نمونه های مربوط به دو طرف یال و هزینه هر گره معادل هزینه هدف آن نمونه در نظر گرفته میشود. با کمک الگوریتم جستجوی دینامیک ویتربی در این گراف میتوان مجموعه بهینه از نمونه های دادگان را برای یک رشته آوایی پیدا نمود. البته در پیاده سازی، بجای در نظر گرفتن تمام نمونه های دادگان بصورت یک گراف با اتصالات کامل، هنگام سنتز یک رشته آوایی هدف، فقط گرافی از نمونه های موجود برای واحدهای سازنده هدف ساخته میشود و مجموعه بهینه نمونه ها با پیدا کردن کوتاه ترین مسیر در این گراف پیدا می شوند. نمونه ای از این گراف در شکل ۲ نشان داده شده است. در این گراف، یک مسیر فرضی بعنوان کوتاه ترین مسیر بصورت خط چین نشان داده شده است.

۳- دادگان مورد استفاده

دادگان یکی از مهمترین اجزای سازنده سیستمهای سنتز با انتخاب واحد می باشند.



شکل ۲- ساختار گراف جستجو برای یافتن مجموعه بهینه از نمونه ها

تحقیقات نشان داده است که هر چه این دادگان بزرگتر باشد و همچنین شیوه بیان و ریتم صحبت گوینده آن یکنواخت تر باشد کیفیت خروجی سیستم بهتر خواهد بود. برای ساخت دادگان این سیستم، از دو ساعت صحبت ضبط شده توسط یک گوینده مرد در اتاقک ضد صدا با فرکانس نمونه برداری ۱۶ کیلوهرتز استفاده شده است. متن صحبت برگرفته از بخشهایی از متون دادگان فارس دات بزرگ می باشد. جملات این متن از نظر فرکانس وقوع هر آوا، متوازن نمی باشند. این دادگان با کمک سیستم تقطیع اتوماتیک پیاده سازی شده [منتشر نشده] در سطح واج برچسب گذاری شده است و سپس برچسبهای یکساعت از آن بطور دستی اصلاح شده است. در انتها ۶۳۰۰۰ نمونه واج بعنوان واحد سنتز بدست آمده است. همانطور که اشاره شد برای هر یک از این نمونه ها یک بردار ویژگی شامل ویژگیهای همچون مقادیر انرژی، پیچ و کشش زمانی و ... استخراج می گردد. بردار ویژگی مربوط به تمامی نمونه ها در یک دیتابیس متنی ذخیره می گردد.

۴- ساختار توابع هزینه

همانطور که اشاره شد برای تعیین میزان شباهت نمونه های یک دادگان به واحدهای سازنده هدف، لازم است که یکسری ویژگیهای مختلف گسسته و سیگنالی انتخاب شوند. در این بخش ابتدا ویژگیهای انتخاب شده برای هزینه هدف و سپس ساختار هزینه اتصال شرح داده می شوند.

۴-۱ ساختار هزینه هدف

از مشخصات آواهای کناری آوای اصلی، در یک همسایگی بطول ۱ حول آوای اصلی از آواها و همچنین موقعیت آوای در هجا بعنوان ویژگیهای گسسته برای هزینه هدف استفاده شده است. برای هر آوای کناری سه مشخصه کلاس آوایی آن، محل تولید آوا و صدادار یا واکدار یا بیواک بودن مد نظر قرار گرفته است. برای صدادارها، کلاس آوایی متمایز کننده حروف صدادار کوتاه و کشیده از یکدیگر و محل تولید آوا نیز مشخص کننده میزان شباهت این حروف از نظر مشخصات ثانویه می باشد. برای ویژگی موقعیت در هجا نیز سه حالت ابتدا، وسط و انتهای هجا بودن، در نظر گرفته شده است. البته ویژگیهای گسسته دیگری نیز همچون موقعیت در کلمه، عبارت و جمله یا جهت کلی تغییر پیچ و انرژی نیز می توانست مورد استفاده قرار گیرد [۲۲]. تابع فاصله ویژگیهای گسسته، مساوی بودن می باشد. در صورتیکه مقدار دو ویژگی یکسان باشد مقدار زیرهزینه مربوط صفر و گرنه یک منظور می شود.

ویژگیهای سیگنالی هزینه هدف، به کمک خصوصیات نوایی آوا همچون انرژی، میزان کشش زمانی آن و مقادیر پیچ آن تعیین می شوند. در این سیستم یک ویژگی برای میزان کشش زمانی و یک ویژگی نیز برای انرژی در نظر گرفته شده است. ویژگی اول با استفاده از لگاریتم مقدار کشش محاسبه می شود. در صورتیکه شروع آوا از نمونه U_S و شماره نمونهٔ پایان آن U_F در سیگنال صحبت باشد و همچنین S_i مقدار نمونهٔ i ام در سیگنال صحبت باشد، ویژگی مربوط به انرژی بصورت زیر محاسبه می شود:

$$E = \log\left(\sqrt{\frac{\sum_{i=U_S}^{U_F}(S_i)^2}{U_F - U_S}}\right) \qquad (۵)$$

علاوه بر این، سه ویژگی برای مقادیر پیچ در نظر گرفته شده که مقدار آنها با تقسیم هر آوا به سه قسمت مساوی و محاسبه لگاریتمِ میانگینِ مقادیرِ پیچِ هر قسمت بدست می آیند. دلیل استفاده از لگاریتم برای محاسبه تمامی ویژگیهای سیگنالی، فراهم نمودن توزیع نرمال بهتر برای مقادیر هر ویژگی می باشد. برای تمامی این ویژگیها، از متریک قدر مطلق تفاضل بعنوان تابع فاصله استفاده شده است. استخراج پیچ یکی از مسائل مهم و مشکل در مباحث پردازش گفتار می باشد، علاوه بر این استخراج مقادیر پیچ بصورت دقیق موجب بهبود کیفیت خروجی سیستم سنتز می شود. در این پژوهش از ابزار Praat برای استخراج مقادیر پیچ استفاده شده است.

۴-۲ ساختار هزینه اتصال

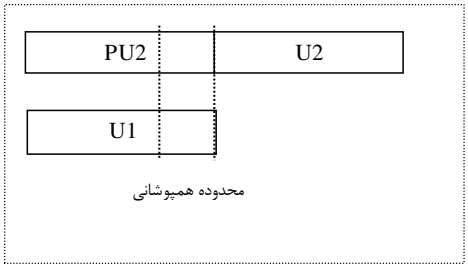
وظیفه هزینهٔ اتصال، محاسبه میزان گسستگی می باشد که در صحبت خروجی سیستم سنتز بر اثر اتصال دو نمونه از دادگان بوجود می آید. این گسستگی میتواند ناشی از گسستگی در منحنی انرژی، پیچ و یا طیف باشد. برای همین برای پاندازه گیری هزینهٔ اتصال، میزان گسستگی انرژی، پیچ و طیف مورد بررسی قرار می گیرد. میزان گسستگی انرژی و پیچ، با متریک قدر مطلق تفاضل مقادیر لگاریتم انرژی و پیچ دو نمونه در مرز خود اندازه گیری میشود. ویژگی مربوط به انرژی بصورت مشابه مطابق فرمول ۵ در محدوده یک فریم بطول ۳۰ میلی ثانیه مجاور مرز محاسبه می شود.

برای اندازه گیری میزان گسستگی طیف، تاکنون معیارهای کمّی متفاوتی توسط محققان مورد ارزیابی و درصد وابستگی آنها به ارزیابی های کیفی و شنوایی ارائه شده است. در این سیستم از ضرایب کپستروم مل برای سنجش میزان

گسستگی استفاده شده است. برای محاسبه گسستگی در سیستمهای انتخاب واحد عموما از دو روش استفاده شده است. در روش اول میزان شباهت فریمهای دو طرف مرز اتصال مورد ارزیابی قرار می گیرد [۲۳] و در روش دوم میزان شباهت بخش انتهایی واج سمت چپ مرز با محدودهٔ مرزی محتوای قبلی واجِ سمتِ راستِ اتصال اندازه گیری می شود. در این روش ممکن است که بخشهای مرزی مورد مقایسه هر یک به چند فریم تقسیم و سنجش شباهت میان فریمهای متناظر انجام شود. بطور مثال در صورتیکه U1 و U2 دو نمونه از دادگان و PU2 نمونه قبلی U2 در دادگان باشد، با توجه به شکل ۳، با اندازه گیری میزان شباهت دو بخش مرزی واقع در محدوده همپوشانی، میزان گسستگی طیف ناشی از اتصال این دو نمونه با یکدیگر مشخص می شود [۲۲].

در این سیستم برای محاسبهٔ میزان گسستگی از روش دوم و از ضرایب نرمالیزه سازی شدهٔ MFCC و مشتق اول آن، بعنوان معیار سنجش استفاده شده است. محدوده همپوشانی ۳۰ میلی ثانیه و یک فریم در نظر گرفته شده است. مقدار گسستگی طیف بصورت زیر محاسبه می شود [۲۰]:

$$Dis_{Spec}(U_1,U_2)=\frac{1}{N}\sum_{i}(\hat{x}_i^1-\hat{x}_i^2)^2 \quad 0\le i\le N \qquad (۶)$$



شکل ۳- روش محاسبه میزان گسستگی طیف

البته برای انتخاب این روش حالتهای دیگری نیز به شرح زیر مورد ارزیابی قرار گرفته است که در نهایت با توجه به میزان حافظه مورد نیاز و همچنین ارزیابی های شنیداری غیر استاندارد، این روش انتخاب شده است:

- استفاده از روش اول: بررسی میزان شباهت فریمهای دو طرف مرز اتصال.
- عدم نرمالیزه سازی ضرایب MFCC و عدم استفاده از مقادیر مشتق اول به دلیل کوچک بودن آنها نسبت به ضرایب MFCC و تاثیر کم آنها در مقدار فاصله.
- در نظر گرفتن محدوده همپوشانی به اندازه ۴۵ میلی ثانیه و تقسیم آن به دو فریم ۳۰ میلی ثانیه با سرعت حرکت ۱۵ میلی ثانیه. در این حالت مقدار نهایی گسستگی طیف برابر با مجموع وزنی مقدار فاصله هر فریم از فریم متناظر خود می باشد. در سیستم کنونی، برای مقدار فاصلهٔ فریمِ نزدیکتر به مرز، وزن ۰٫۶ و برای دیگری، وزن ۰٫۴ انتخاب شده است.

برای محاسبهٔ ضرایبِ MFCC و مشتق اول آن، قبل از فریم بندی پیش تاکید با ضریب ۰٫۹۷۵ انجام شده است. برای پنجره بندی نیز از تابع همینگ استفاده شده است. قابل ذکر است در بعضی از سیستمها زیر هزینه های دیگری نیز با توجه به محتواهای آوایی مجاور اتصال در نظر گرفته اند.

۳-۴ نرمالیزه کردن ویژگیهای سیگنالی

همانطور که اشاره شد توابع هزینه، بواقع جمع وزندار مقدار متریکهای ویژگیهای مختلف هستند. بدلیل اینکه این ویژگیها نسبت به یکدیگر محدوده تغییر متفاوتی دارند، متریکهای آنها نیز دامنه های تغییر متفاوتی نسبت به یکدیگر پیدا می کنند و این امر تعیین مقدار وزنها را دشوار می کند. لذا تمامی ویژگیهای سیگنالی به فرم نرمالیزه با میانگین صفر و انحراف معیار واحد تبدیل می شوند. برای تبدیل یک مقدار ویژگی x به فرم نرمالیزه با مقدار میانگین μ و انحراف معیار σ بصورت زیر عمل می شود.

$$Norm(x)=\frac{x-u}{\sigma} \qquad (۷)$$

با توجه به اینکه برای توزیعهای نرمال، در حدود ۹۵٫۵٪ داده ها در محدوده $[\mu-2\sigma,\mu+2\sigma]$ واقع می شوند، محدوده تغییر ویژگیهای نرمالیزه شده را میتوان $[-2,2]$ دانست. در اینحالت مقدار تابع فاصله تمام ویژگی ها در محدوده $[0,4]$ تغییر می نماید. برای نرمالیزه سازی ضرایب MFCC و مشتق اول آن، هر ضریب نسبت به ضرایب دیگر مستقل در نظر گرفته شده و میانگین و انحراف معیار آن محاسبه می شود. برای نرمالیزه سازی کشش زمانی نیز، میانگین و انحراف معیار نمونه های هر واج بطور جداگانه محاسبه و از این مقادیر برای نرمالیزه سازی نمونه های واج استفاده شده است. دلیل این امر را می توان در شبیه بودن میزان کشش نمونه های یک واج نسبت به یکدیگر به دلیل کشش ذاتی آن واج و تفاوت زیاد میان میزان کشش ذاتی واجهای مختلف دانست. بطور مثال عموما بیشتر نمونه های واج /A/ دارای کشش زیاد می باشند در حالیکه نمونه های واج /r/ کوتاه می باشند. در اینحالت استفاده از یک مقدار میانگین و انحراف معیار کل برای نرمالیزه سازی ویژگی کشش زمانی موجب کوچک شدن این زیر هزینه نسبت به زیر هزینه های دیگر می گردد و لذا بهتر است که برای هر واج یک مقدار میانگین و انحراف کل در نظر گرفته شود. با توجه به نرمالیزه کردن ویژگیها می توان گفت متریک انتخابی برای تمامی ویژگیهای سیگنالی بجز زیرهزینهٔ گسستگی طیف، فاصلهٔ اقلیدسی نرمالیزه شده بعنوان یک حالت خاص از متریک Mahalanobis می باشد. برای ارزیابی و مقایسه محدوده تغییر هزینه های هدف و اتصال و همچنین هر یک از زیرهزینه ها نسبت به یکدیگر، یک رشته آوایی هدف شامل ۲۰۰۰ واج بطول ۱۶۰ ثانیه سنتز و مقادیر میانگین (μ)، انحراف معیار (σ) و پارامتر $\mu+2\sigma$ بعنوان مقدار ماکزیمم، برای هر هزینه یا زیرهزینه محاسبه و در جدول ۱ ارائه شده اند. علاوه بر این برای ارزیابی دقیقتر هر هزینه یا زیرهزینه توزیع تقریبی مقادیر آن تخمین و در جدول ۲ نشان داده شده است. با توجه به این داده ها می توان گفت که محدوده تغییر هزینه ها و زیرهزینه ها بجز زیرهزینه مربوط به انرژی تقریبا یکسان و قابل قبول می باشد. کم بودن محدوده تغییر زیرهزینه هدف مربوط به انرژی احتمالا ناشی از شرایط یکسان ضبط صدا می باشد.

۵- تعیین وزنها

کیفیت نمونه های انتخاب شده برای یک رشته آوایی هدف، به مقدار زیادی به وزنهایی که برای زیرهزینه های مختلف در هزینه هدف و اتصال انتخاب میشود، بستگی دارد. تعیین وزنها برای هزینه اتصال، بدلیل کم بودن تعداد وزنها (سه تا) بصورت تجربی و با سعی و خطا ممکن می باشد. در جدول ۳ مقادیر وزن در نظر گرفته شده برای هزینه اتصال نشان داده شده است.

جدول ۱- اطلاعات آماری هزینه ها و زیر هزینه ها

دسته	زیرهزینه مربوط به ویژگی	μ	σ	$\mu+2\sigma$
هزینه اتصال	گسستگی پیچ	۰,۹۲	۱,۰۲	۲,۹۷
	گسستگی انرژی	۰,۸۸	۰,۷۰	۲,۲۷
	گسستگی طیف	۱,۳۱	۰,۶۸	۲,۶۸
	کل هزینه	۱,۱۰	۰,۵۴	۲,۱۹
هزینه هدف	پیچ، بخش اول واج	۰,۹۳	۰,۸۳	۲,۵۹
	پیچ، بخش دوم واج	۰,۹۳	۰,۸۲	۲,۵۶
	پیچ، بخش سوم واج	۰,۹۵	۰,۸۲	۲,۵۹
	انرژی	۰,۵۹	۰,۴۹	۱,۵۶
	کشش زمانی	۰,۸۷	۰,۶۶	۲,۲
	کل هزینه	۱,۲۹	۰,۴۷	۲,۲۳

جدول ۲- توزیع تقریبی هزینه ها و زیرهزینه ها

دسته	زیرهزینه مربوط به ویژگی	درصد مقادیر موجود در محدودهٔ				
		۱-۰	۲-۱	۳-۲	۴-۳	۴<
هزینه اتصال	گسستگی پیچ	٪۷۱	٪۱۸	٪۳	٪۸	٪۰
	گسستگی انرژی	٪۶۴	٪۲۸	٪۷	٪۰,۸	٪۰,۲
	گسستگی طیف	٪۳۹	٪۴۵	٪۱۴	٪۲	٪۰,۱
	کل هزینه	٪۵۰	٪۴۲	٪۸	٪۰,۳	٪۰
هزینه هدف	پیچ، بخش اول واج	٪۶۴	٪۲۸	٪۴	٪۴	٪۰
	پیچ، بخش دوم واج	٪۶۴	٪۲۸	٪۴	٪۴	٪۰
	پیچ، بخش سوم واج	٪۶۲	٪۲۹	٪۵	٪۳	٪۰
	انرژی	٪۸۲	٪۱۶	٪۲	٪۰,۱	٪۰
	کشش زمانی	٪۶۴	٪۲۹	٪۶	٪۱	٪۰,۱
	کل هزینه	٪۲۶	٪۶۶	٪۷	٪۰	٪۰

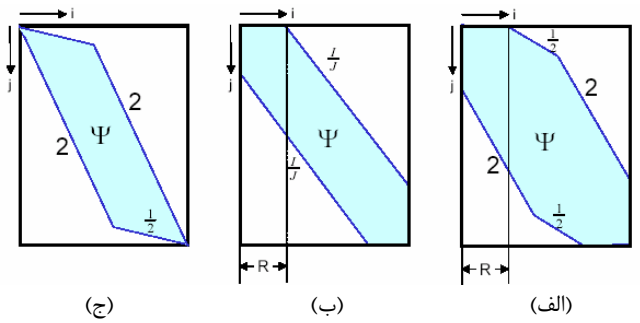
جدول ۳- مقادیر وزن برای هزینه اتصال

وزن مربوطه	میزان گسستگی	نوع اتصال
۰,۴	پیچ	بین واکدارها
۰,۲۵	انرژی	
۰,۳۵	طیف	
۰,۴	انرژی	بقیهٔ حالتها
۰,۶	طیف	

برای هزینه اتصال دو دسته وزن در نظر گرفته شده است. در صورتیکه در اتصال مورد بررسی هر دو نمونه واکدار باشند از مجموعه وزنهای دسته اول و در غیر اینصورت از مجموعه وزنهای دسته دوم استفاده میشود. اگر هر دو نمونه واکدار باشند ولی مقدار پیچ مرزی برای یک نمونه موجود و برای نمونه دیگر ناموجود باشد، مقدار گسستگی پیچ برابر با ماکزیمم مقدار فاصله و در صورتیکه مقدار پیچ مرزی برای هر دو نمونه ناموجود باشد، مقدار گسستگی پیچ برابر با یک دوم ماکزیمم مقدار فاصله در نظر گرفته می شود. برای اینکه محدوده تغییر هزینه اتصال همیشه $[0,4]$ باشد، مقدار وزنهای هزینه اتصال همواره بگونه ای انتخاب میشود که مجموع آن ۱ باشد.

۵-۱ تعیین وزنهای هزینه هدف

برای هزینه هدف، بدلیل زیاد بودن تعداد ویژگیهای در نظر گرفته شده برای آن، بهتر است که از یک پروسه اتوماتیک برای تعیین مقادیر بهینه وزنها استفاده شود. برای یافتن مقادیر بهینه وزنها بصورت اتوماتیک پروسه ای مورد نیاز است که در آن کیفیت خروجی سیستمهای سنتز بصورت ادراکی ارزیابی شود. دو روش کلی برای تعیین وزنهای بهینه وجود دارد. روش اول جستجو در فضای وزنها،[۵] و روش دوم آموزش به کمک رگرسیون[۶] می باشد. هر دو این روشها برای پیدا کردن وزنهای بهینه از صحبت طبیعی موجود در دادگان، شامل رشته آوایی آن و ویژگیهای محاسبه شده برای آن، بعنوان رشته آوایی هدف استفاده می کنند. در هر دو روش تعیین وزنها، هدف مینیمم نمودن تفاوت میان سیگنال صحبت سنتز شده و سیگنال صحبت طبیعی مربوط به رشته آوایی هدف می باشد. بدین منظور لازم است که یک تابع فاصله یا معیار هدف مناسب با ضریب همبستگی بالا با ادراک انسان برای مشخص کردن میزان اختلاف میان صحبت سنتز شده و صحبت طبیعی تعریف شود. انتخاب تستهای ادراکی همچون تست MOS بعنوان معیار هدف بدلیل مشکل و خطاپذیر بودن این تستها و همچنین عدم امکان انجام این تستها برای متون طولانی و در دفعات متعدد ناممکن می باشد. لذا سعی میشود از معیارهایی کمّی که ضریب همبستگی بالایی با ادراک انسان دارند بعنوان معیار هدف استفاده شود. یکی از معیارهای هدف مناسب، میانگین فاصله اقلیدسی ضرایب MFCC میان بردارهای فریمهای تراز شده زمانی[۷] یک نمونه با بردارهای فریمهای واحد هدف می باشد. در این پژوهش برای تراز کردن زمانی بردارها با یکدیگر از الگوریتم پیچش زمانی پویا[۸] استفاده شده است. تراز بندی زمانی با استفاده از الگوریتم DTW بدون محدود کردن فضای جستجو ممکن است که بدرستی صورت نگیرد. لذا عموما برای این الگوریتم یک فضای جستجو مشابه یکی از حالتهای شکل ۴ تعریف میشود [۲۴]. فاصله با بردار ویژگیهای خارج از این فضای جستجو بی نهایت در نظر گرفته می شود.



شکل ۴- نمونه هایی از محدوده جستجوهای تعریف شده

در این پژوهش از محدوده جستجویی مشابه نمونه (ب) شکل ۴ استفاده می شود. بدین منظور لازم است که دو شرط زیر برای اندیس میان دو بردار ویژگی که با یکدیگر مقایسه می شوند درست باشد و گرنه فاصله میان این دو بردار ویژگی بی نهایت در نظر گرفته میشود:

$$\begin{cases} j+1-n_y/n_x*(i+1)-n_y/2-0.5<0 \\ j+1-n_y/n_x*(i+1)+n_y/2+0.5>0 \end{cases} \quad (۸)$$

این دو شرط بواقع برگرفته از دو معادله خطی هستند که به فاصله مساوی از خط $j=n_y/n_x*i$ قرار گرفته اند و یکی عرض از مبدا $n_y/2$ و دیگری عرض از مبدا $-n_y/2$ دارد که از نقطه $(n_x/2,0)$ عبور می کند. تنها این دو شرط نسبت به این دو خط بگونه ای در نظر گرفته شده اند که محدوده جستجوی بزرگتری بسازند و همچنین شامل اندیسهایی که روی این دو خط می افتند، نیز باشند.

در روش اول برای هر وزن، یکسری مقادیر ممکن در نظر گرفته میشود. سپس بازای هر ترکیب ممکن از مقادیر وزنها، مجموعه متنوعی از گویشها[9] با استفاده از این مقادیر سنتز میشوند و صحبت سنتز شده برای گویشها ساخته میشود. سپس با کمک معیار هدف، تفاوت میان صحبت سنتز شده و صحبت طبیعی برای گویشهای مختلف سنجیده میشود و در انتها آن ترکیبی از مقادیر وزنها که معیار هدف محاسبه شده برای خروجی آنها بازای گویشهای مختلف پایدارتر و مینیمم باشد، بعنوان مجموعه بهینه وزنها انتخاب میشوند. این روش نیاز به حجم محاسبات بسیار زیادی دارد و با اضافه شدن تعداد پارامترهای وزن و یا مقادیر ممکن بازای هر وزن زمان محاسبات بصورت توانی افزایش پیدا می کند. از این روش میتوان برای تعیین مجموعه وزنهای بهینه کل تابع هزینه، شامل مجموع هزینه هدف و اتصال استفاده نمود. بررسیها نشان می دهد که این روشِ آموزش، اهمیت بیشتری برای تابع هدف قائل میشود که موجب ایجاد گسستگی های واضح در سیستم و کاهش کیفیت خروجی می شود. در این روش برای محاسبه معیار هدف میان صحبت طبیعی و صحبت خروجی سیستم، از فاصله ضرایب MFCC فریمهای این دو سیگنال که با یکدیگر تراز شده اند، میانگین گیری میشود. بدلیل کوچک بودن محل وقوع گسستگیها در اتصالات میان هر دو واحد سنتز نسبت به زمان کشش واحدها، تاثیر گسستگیها در میزان میانگین نهایی ناچیز بوده و لذا هنگام تعیین وزنها، برای زیرهزینه هایِ هزینه هدف اهمیت بیشتری نسبت به هزینه اتصال قائل می شود.

روش دوم از رگرسیون خطی متعدد[10] و معیار هدف برای تعیین وزنهای هزینه هدف استفاده می کند. در این روش هر آوا یا هر کلاسی از آواها میتوانند مقادیر وزنهای متفاوتی داشته باشند. الگوریتم کار بصورت زیر می باشد:

۱- برای هر نمونه از دادگان متعلق به کلاس آوای فعلی که ضرایب آن تعیین میشود، کارهای زیر انجام شود.
   a. با این نمونه همچون یک نمونه هدف رفتار شود.
   b. با استفاده از معیار هدف، اختلاف سیگنالی و صوتی میان این نمونه با نمونه های دیگر کلاس آوای فعلی محاسبه میشود.
   c. N تا از بهترین نمونه ها که با کمترین میزان فاصله که شباهت بیشتری دارند انتخاب شود. (مثلا N=20)
   d. متریکهای مربوط به تابع هدف برای تمام این N نمونه با در نظر گرفتن نمونه هدف محاسبه شود.
۲- با استفاده از تمام نمونه های هدف و نمونه های مشابه آن، مجموعه ای از معیار هدف و زیرهزینه های تابع هدف متناظر تشکیل شود.
۳- با استفاده از رگرسیون خطی متعدد سعی می شود که معیار با استفاده از مجموع وزنی زیر هزینه های محاسبه شده پیشگویی شود. وزنهای بدست آمده از رگرسیون، بعنوان وزنهای این مجموعه آوا استفاده میشود.
۴- مراحل بالا برای هر مجموعه آوا تکرار میشود.

این روش نسبت به روش جستجوی فضای وزنها، مزیتهای زیادی دارد. اولا با کمک این روش می توان برای هر نوع واحد و یا هر کلاس واحد، یک دسته وزن داشته باشیم. دوما زیاد شدن پارامترها زمان محاسبه را بصورت توانی افزایش نمی دهد و زمان آموزش نسبت به روش اول بسیار کوتاهتر میباشد[3].

در ابتدا در این سیستم برای تعیین مجموعه بهینه وزنها برای هر واج از روش دوم استفاده گردید. از میانگین فاصله اقلیدسی ۱۲ ضریب MFCC میان بردارهای تراز شده زمانی یک نمونه با بردارهای هدف بعنوان معیار هدف استفاده شد. برای محاسبه ضرایب MFCC اندازه پنجره ۳۰ میلی ثانیه و سرعت حرکت فریم ۱۵ میلی ثانیه انتخاب شده است. همچنین قبل از فریم بندی، پیش تاکید با ضریب ۰,۹۷۵ انجام شده است و برای پنجره بندی نیز از تابع همینگ استفاده شده است. بعد از استخراج تمامی بردارهای آموزشی مربوط به تمام نمونه های یک واج، محدودهٔ معیارِ هدف به [0,4] تغییر داده میشود. هنگام آموزش با رگرسیون، بواقع سعی می شود مجموعه وزنهای بهینه بگونه ای انتخاب شوندکه هزینه هدف همیشه مقدارِ معیارِ هدف را پیشگویی نماید. لذا با تغییر محدوده معیار هدف به [0,4] میتوان امیدوار بود که محدوده تغییر هزینهٔ هدف نیز [0,4] شود. در صورتیکه معیار هدف OM، میانگین آن μ ،انحراف معیار آن σ و مقدار مینیمم آن $OM_{\min}$ در نظر گرفته شود، تغییر محدوده مطابق فرمول زیر صورت می گیرد:

$$OM_{range} = \frac{OM - OM_{\min}}{(\mu + 2\sigma - OM_{\min})} \qquad (۹)$$

برای انجام رگرسیون از ابزار ols مربوط به کتابخانه Speech_Tools سیستم فستیوال استفاده شده است [۲۵]. ضریب همبستگی میان ویژگیهای ورودی و معیار هدف هنگام رگرسیون بسیار ضعیف و برای واجهای مختلف در محدوده ۱۵٪ تا ۴۰٪ می باشد. این مساله می تواند ناشی از نامناسب بودن معیار هدف و همچنین ناکافی بودن ویژگیهای انتخاب شده برای هزینه هدف باشد. ارزیابی های شنیداری غیررسمی حاکی از نامناسب بودن کیفیت خروجی سیستم هنگام استفاده از وزنهای حاصل از روش رگرسیون می باشند. لذا در نهایت برای هزینه هدف از دو دسته وزن تجربی استفاده شده است. از یک دسته وزن برای واجهای واکدار و برای در نظر گرفتن ویژگیهای مربوط به پیچ و از دسته دیگر برای واجهای بیواک استفاده می شود.

۶- الگوریتم انتخاب واحد

ورودیِ سیستمِ سنتز یک رشته آوایی هدف همراه با ویژگیهای پروزودیک آن شامل مقادیرکشش زمانی و انرژی هر واج و منحنی پیچ رشته آوایی می باشد. بدلیل عدم وجود یک سیستم تبدیل متن به گفتار کامل و عدم داشتن مدلهای پیشگویی کننده نوای متون و رشته های آوایی برای این گوینده، مشخصات پروزودیک رشته آوایی هدف از گفتار طبیعی آن که توسط گوینده دادگان بیان شده است استخراج میشود که به اینکار سنتز مجدد[11] گفته میشود. برای سنتز ورودی، ابتدا رشته آوایی هدف به گویشهای سازنده آن تقسیم بندی میشود. سنتز هر گویش بطور جداگانه و با ساختن یک گراف از نمونه های موجود برای واحدهای سازنده آن گویش صورت می گیرد. در این گراف، گره های ابتدایی و انتهایی سکوت دو سوی گویش می باشد. همچنین هزینه اشغال هر گره، هزینه هدف نمونه و هزینه عبور از یال هر گراف برابر با هزینه اتصال دو نمونه در دو سوی یال در نظر گرفته میشود. هزینه اتصال نمونه های ابتدا و انتهایی گویش به سکوت دو طرف آن صفر در نظر گرفته میشود. نمونه ای از این گراف در شکل ۲ نشان داده شده است. هدف پیدا کردن مسیری از نمونه ها در این گراف می باشد که کمترین هزینه را داشته باشد. بدلیل مشابهت این گراف با مدل مخفی مارکف، این مساله مشابه پیدا کردن مسیر با بیشترین احتمال وقوع در مدل مخفی مارکف می باشد. لذا برای پیدا کردن این مسیر میتوان مشابه مدل مخفی مارکف از الگوریتم جستجوی ویتربی استفاده نمود. با در نظر گرفتن ک ضریب W_C برای کل هزینه های اتصال برای ایجاد انعطاف بیشتر در سیستم، هزینه کل یک مسیر بطول n بصورت زیر محاسبه می شود:

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + W_c \sum_{i=2}^{n} C^c(u_{i-1}, u_i) \qquad (۱۰)$$

با انتخاب مقادیر بزرگتر از ۱ برای این ضریب، سیستم مسیرهایی که بیشتر نمونه های آن در دادگان، همجوار بوده یا اتصالات هموارتری ایجاد می کنند،

انتخاب میکند. در این سیستم برای حفظ توازن میان هزینه هدف و اتصال مقدار این ضریب ۱ در نظر گرفته شده است.

اگر تعداد واحدهای هدف T تا باشد، جستجوی ویتربی در T مرحله، کوتاهترین مسیر را پیدا می کند. فرض کنید گراف حاصل کلا شامل N نمونه یا گره (v) باشد که هزینه هدف برای هر نمونه یا هزینه اشغال هر گره با $C_i^t \quad i=1,...,N$ مشخص شود و این گره ها بوسیله M یال (e) به یکدیگر متصل می باشند که هزینه عبور از آنها با $C_{ij}^c \quad i,j=1,...N$ مشخص شود. متغیر $\delta_t(i)$ هزینه کوتاهترین مسیر در مرحله t ام، از گره ابتدایی تا گره i ام در نظر گرفته شده است. همچنین فرض شود که هر گره دارای یک متغیر $\omega_t(i)$ برای ذخیره کوتاهترین مسیر باشد. در اینصورت الگوریتم جستجوی ویتربی دارای سه مرحله بصورت زیر می باشد:

۱. مقدار دهی اولیه:

$$\delta_1(i)=C_i^c \qquad 1\le i\le N \tag{۱۱}$$

$$\omega_t(i)=0 \qquad 1\le i\le N \tag{۱۲}$$

۲. روال برگشتی:

$$\delta_t(j)=\min_{1\le i\le N}[\delta_{t-1}(i)+C_{ij}^c]+C_j^t \qquad 2<t<T,\ 1<j<N \tag{۱۳}$$

$$\omega_t(j)=\operatorname*{argmin}_{1\le i\le N}[\delta_{t-1}(i)+C_{ij}^c] \qquad 2<t<T,\ 1<j<N \tag{۱۴}$$

۳. خاتمه

$$P^*=\min_{1\le i\le N}[\delta_T(i)] \tag{۱۵}$$

مسیر بهینه با برگشت به عقب بصورت زیر مشخص میشود:

$$q_t^*=\omega_{t+1}(q_{t=1}^*) \qquad t=T-1,T-2,...,1 \tag{۱۶}$$

P^* مقدار هزینه کوتاهترین مسیر و q_t^* اندیس گره های موجود در این مسیر را مشخص می نماید. پیچیدگی محاسباتی این الگوریتم از مرتبه $O(N^2T)$ می باشد. در هر مرحله از روالِ برگشتی در جستجوی ویتربی برای یافتن مسیر با هزینه مینیمم تا یک نمونه خاص، لازم است که هزینه اتصال آن نمونه با تمام نمونه های لایه قبل محاسبه گردد. در اینحالت به دلیل اینکه در تمامی هزینه های اتصال مورد بررسی نمونه سمت راست مرز یکسان می باشد، برای افزایش سرعت سیستم از یک حافظه با سرعت دسترسی بالاتر[۱۲] برای ذخیرهٔ ویژگیهای نمونه سمت راست مربوط به هزینه اتصال بجای محاسبه یا بارگذاری مداوم آنها از فایل استفاده شده است.

۶-۱ الگوریتمهای پیرایش[۱۳]

یکی از مسائل بحرانی در سیستمهای سنتز انتخاب واحد، بالا بودن حجم پردازش و کند بودن زمان سنتز می باشد. دلیل این مشکل تعداد زیاد نمونه هایی می باشد که باید بررسی شوند تا بهترین آنها انتخاب شود. عموما بخش بحرانی و زمانبر این سیستمها الگوریتم جستجوی ویتربی و محاسبه هزینه اتصال میان نمونه ها در هنگام جستجو می باشد. لذا برای کاهش زمان محاسباتی سیستم بهتر است، بجای بررسی تمامی نمونه های دادگان بازای هر واحد هدف، تنها نمونه های مناسب مورد بررسی قرار گیرند و بدین ترتیب تعداد نمونه های گراف محدود شوند. برای اینکار از الگوریتمهای پیرایش استفاده میشود. در حالت ایده آل، این الگوریتمها باید بگونه ای طراحی گردند که بدون کاهش کیفیت خروجی سیستم و حفظ نمونه های موجود در مسیر بهینه، از تعداد نمونه های مورد پردازش و زمان محاسباتی سیستم بکاهند. البته در عمل استفاده از این الگوریتمها بسته به موقعیت بکارگیری و نوع الگوریتم موجب کاهش کیفیت خروجی سیستم می گردند. در سیستمهای انتخاب واحد، عموما پیرایش نمونه ها در سه مرحله صورت می گیرد. مرحله اول پیرایش، هنگام جستجو میان نمونه های یک واحد هدف و قبل از اضافه کردن آنها به گراف انجام می شود. در این مرحله عموما پیرایش بر اساس مشابهت های آوایی صورت می گیرد. در این سیستم به دلیل اینکه محاسبه هزینه هدف پیچیدگی محاسباتی بسیار کمی دارد و همچنین مشکلات پیرایش بر اساس مشابهت های آوایی، پیرایشی صورت نمی گیرد. مرحله دوم بعد از محاسبه هزینه هدف برای تمام نمونه های یک واحد هدف و با توجه به مقدار هزینه هدف نمونه ها صورت می گیرد. در این حالت تعدادی از نمونه ها با هزینه هدفِ مینیمم نگهداری و بقیه نمونه ها حذف می شوند. در این سیستم در این مرحله از پیرایش حداکثر ۵۰ نمونه بازای هر واحد هدف نگهداری می شود. برای انجام این پیرایش، تمام نمونه ها بر اساس مقدار هزینه هدف بصورت افزایشی مرتب می شوند. اگر تعداد کل نمونه های لیست N تا باشد، تعداد نمونه هایی که در این پیرایش حفظ میشوند، بصورت زیر محاسبه می شود:

$$N_{Keep}=\begin{cases} N & if\ \ N\le 25 \\ .1*(N-25)+25 & if\ \ 25<N<275 \\ 50 & if\ \ N\ge 275 \end{cases} \tag{۱۷}$$

مرحله سوم پیرایش هنگام جستجوی ویتربی و بعد از تمام شدن هر مرحله از روالِ برگشتی انجام می شود. در این روش که پیرایش با پهنای دید[۱۴] نیز نامیده می شود، بعد از اتمام هر مرحله، درصدی از مسیرها با هزینهٔ مینیمم نگهداری و بقیه دور ریخته می شوند. در سیستم کنونی، تعداد مسیرهایی که در پیرایش با پهنای دید نگه داشته می شوند با توجه به تعداد کل مسیرها در هر مرحله N توسط رابطه زیر مشخص میشوند.

$$N_{Keep}=\begin{cases} N & if\ \ N\le 10 \\ .25(N-10)+10 & if\ \ N>10 \end{cases} \tag{۱۸}$$

البته پیرایش با پهنای دید بر اساس مقدار هزینه مسیرها نیز می تواند صورت گیرد. در اینحالت در هر مرحله مسیرهایی که هزینه آنها از مقدار خاصی بزرگتر باشد، کنار گذاشته می شوند. در سیستم CHATR در مرحله پیرایش با هزینه هدف ۲۵ تا ۵۰ نمونه و در پیرایش هنگام جستجوی ویتربی ۱۰ تا ۲۰ مسیر نگهداری میشود [۳].

۷- ارزیابی سیستم

بعد از انتخاب نمونه های بهینه، سیگنال خروجی با متصل کردن سیگنال این نمونه ها به یکدیگر ساخته می شود. برای بهبود کیفیت خروجی، در مرزهای اتصال بطور ساده یک هموار سازی ساده صورت می گیرد. برای ارزیابی کیفیت خروجی سیستم سنتز از تست MOS برای سنجش چهار معیار کیفیت کلی صحبت، میزان طبیعی بودن، قابل فهم بودن و خوشایند بودن صحبت سنتز شده استفاده شده است. بدین منظور ۵۰ نمونه صحبت طبیعی از گوینده دادگان و خارج از نمونه های موجود در دادگان سیستم با طول ۵ تا ۱۵ میلی ثانیه انتخاب، پارامترهای نوایی آنها استخراج و توسط سیستم، سنتزِ دوباره می شوند. برای کاهش نمونه ها در هر تست، این ۵۰ نمونهٔ صحبتِ سنتز شده، بطور تصادفی به چهار گروه هر یک متشکل از ۲۲ نمونه تقسیم بندی می شوند. هر یک از نمونه ها

می تواند حداکثر در دو گروه وجود داشته باشد. برای ارزیابی سیستم، معیارهای کیفیت کلی، میزان طبیعی بودن، قابل فهم بودن و خوشایند بودن صحبت برای نمونه های هر گروه توسط شش گوینده (جمعا ۲۴ گوینده برای تمام گروهها) مورد ارزیابی قرار می گیرد. به هنگام تست، شنوندگان به هر نمونه حداکثر دوبار می توانند گوش فرا دهند و علاوه بر این امکان برگشت به نمونه های قبلی و تصحیح نمرات نیز وجود ندارد. شنوندگان بعد از گوش دادن به هر نمونه، به هریک از معیارها امتیازی بین ۱ تا ۵ بصورت صحیح یا اعشاری می دهند.

برای سنجش میزان دقت و سختگیری شنوندگان به هنگام ارزیابی، به هر گروه بصورت تصادفی ۵ نمونه از نمونه های صحبت طبیعی اولیه نیز اضافه شده است. با اضافه نمودن نمونه های طبیعی می توان شنوندگانی که نمرات کمی به نمونه های طبیعی می دهند را به علت سختگیری بیش از حد کنار گذاشت. علاوه بر این می توان با محاسبهٔ ضریب همبستگی نمرات داده شده به نمونه های طبیعی برای یک شنونده، در صورتیکه این مقدار از حد آستانهٔ خاصی کمتر باشد به علت بی دقت بودن شنونده امتیازات وی را در محاسبات نهایی در نظر نگرفت [۷]. در این ارزیابی ۲۸ شنونده سیستم را مورد ارزیابی قرار داده اند که در انتها امتیازات ۴ شنونده به دلایل فوق الذکر در نظر گرفته نشده است. از نظر آماری ۷۵٪ شنوندگان مرد و بقیه زن هستند. همچنین ۹۲٪ شنوندگان در محدوده سنی ۲۰-۲۷ قرار دارند و بقیه بزرگتر از این محدوده سنی می باشند. مقدار میانگین و انحراف معیارِ امتیازات داده شده به هر معیار برای نمونه های طبیعی و نمونه های سنتز شده در جدول ۴ نشان داده شده است. همانطور که مشاهده می شود معیار میزان قابل فهم بودن بیشترین میانگین امتیاز و میزان خوشایند بودن کمترین میانگین امتیاز را بدست آورده اند. مقدار MOS برای کیفیت کلی سیگنال نیز ۳،۸ بدست آمده است. مقادیر انحراف معیار حاکی از وجود بیشترین اختلاف نظر در تعریف و چگونگی امتیازدهی برای معیارهای میزان خوشایند بودن و میزان طبیعی بودن بین شنوندگان می باشد. به همین ترتیب بنظر می آید کمترین اختلاف نظر بین شنوندگان در تعریف و چگونگی امتیازدهی به معیارِ میزان قابل فهم بودن وجود داشته است.

جدول ۴- نتایج حاصل از ارزیابی سیستم توسط تمام شنوندگان

	نمونه های طبیعی		نمونه های سنتز شده	
	مقدار MOS	مقدار انحراف معیار	مقدار MOS	مقدار انحراف معیار
کیفیت کلی سیگنال	۴،۷	۰،۳۴	۳،۸	۰،۵۶
میزان طبیعی بودن	۴،۷	۰،۳۲	۳،۶	۰،۵۸
میزان قابل فهم بودن	۴،۹	۰،۲۳	۳،۹	۰،۴۶
میزان خوشایند بودن	۴،۵	۰،۴۵	۳،۵	۰،۵۹

با تفکیک شنوندگان بر اساس میزان آشنایی آنها با کیفیت خروجی دیگر سیستمهای سنتز و همچنین پردازش صحبت به دو دستهٔ حرفه ای و غیر حرفه ای، ۷ شنونده جزء دستهٔ حرفه ای و بقیه جزء دستهٔ غیر حرفه ای قرار می گیرند. در جداول ۵ و ۶ میانگین و انحراف معیار امتیازات داده شده توسط شنوندگان هر دسته نشان داده شده است. همانطور که مشاهده می شود، امتیازات شنوندگان حرفه ای ثبات بیشتری دارد و همچنین نسبت به امتیازات گروه دیگر، هم برای داده های طبیعی و هم برای نمونه های سنتز شده میانگین بالاتری دارد. هر چند که بنظر می آید اختلاف نظر زیادی بین شنوندگان هر دو دسته در چگونگی امتیازدهی به معیارهای میزان خوشایند بودن و میزان طبیعی بودن صحبت وجود دارد.

جدول ۵- نتایج حاصل از ارزیابی سیستم توسط شنوندگان حرفه ای

	نمونه های طبیعی		نمونه های سنتز شده	
	مقدار MOS	مقدار انحراف معیار	مقدار MOS	مقدار انحراف معیار
کیفیت کلی سیگنال	۴،۷۹	۰،۳۱	۴،۰۲	۰،۴۱
میزان طبیعی بودن	۴،۷۶	۰،۲۹	۳،۸۸	۰،۵۵
میزان قابل فهم بودن	۴،۹۴	۰،۰۸	۴،۰۱	۰،۳۳
میزان خوشایند بودن	۴،۷۵	۰،۲۵	۳،۷۵	۰،۶۲

جدول ۶- نتایج حاصل از ارزیابی سیستم توسط شنوندگان غیرحرفه ای

	نمونه های طبیعی		نمونه های سنتز شده	
	مقدار MOS	مقدار انحراف معیار	مقدار MOS	مقدار انحراف معیار
کیفیت کلی سیگنال	۴،۷	۰،۳۵	۳،۶۷	۰،۵۹
میزان طبیعی بودن	۴،۶	۰،۳۳	۳،۵۰	۰،۵۷
میزان قابل فهم بودن	۴،۸۵	۰،۲۷	۳،۸۷	۰،۵۱
میزان خوشایند بودن	۴،۴۷	۰،۴۹	۳،۴۳	۰،۵۷

۸- نتیجه گیری و پیشنهادات

در این مقاله، ساختار یک موتور سنتز با روش انتخاب واحد برای زبان فارسی شرح داده شد. در این بررسی ساختار هزینه های اتصال و هدف و ویژگیهای مورد استفاده برای انتخاب بهترین نمونه ها و همچنین الگوریتم جستجو مورد بررسی قرار گرفت. برای ایجاد یک فضای رقابتی عادلانه به هنگام یافتن مسیر شامل نمونه های بهینه، ویژگیهای پیوسته به فرم نرمالیزه مورد استفاده قرار می گیرند. همچنین روشهای مختلف تعیین وزنهای توابع هزینه هدف و اتصال علی الخصوص برای هزینه هدف بدلیل داشتن زیرهزینه های بیشتر معرفی گردید که در نهایت بدلیل مشکلاتی که این روشها داشتند، دسته وزنها بصورت تجربی و بصورت سعی و خطا تعیین گردیدند.

یکی از مسائل بحرانی این سیستمها پیچیدگی محاسباتی بالای الگوریتم جستجو می باشد. بواقع محاسبه هزینه های اتصال به هنگام جستجوی ویتربی، زمانبرترین بخش می باشد که برای کاهش این زمان و انجام سنتز بصورت برخط از الگوریتمهای پیرایش استفاده می شود. پیرایش در سه مرحله مختلف، هنگام پیدا کردن نمونه ها، بعد از محاسبه هزینه هدف و با استفاده از آن و در انتها هنگام جستجوی ویتربی صورت می گیرد. استفاده از الگوریتمهای پیرایش برای افزایش سرعت، موجب افت کیفیت خروجی سیستم می گردد. بررسیها نشان می دهند که الگوریتمهای پیرایش پیاده سازی شده به اندازه کافی مناسب نمی باشند و لذا لازم است این الگوریتمها برای بهبود کارایی مورد بازنگری قرار گیرند. در نهایت برای ارزیابی خروجی سیستم از تست MOS از نظر چهار معیار کیفیت کلی، میزان

قابل فهم بودن، طبیعی بودن و خوشایند بودن استفاده شده است که به ترتیب برای این معیارها مقادیر ۳/۸، ۳/۹، ۳/۶ و ۳/۵ آمده است. این در حالی است که نتایج ارزیابی کیفیت کلی برای یکسری از نمونه های طبیعی به هنگام تست MOS، مقدار ۴,۷ بدست آمده است. علاوه بر این با فرض اینکه سیستم توسط شنوندگان حرفه ای بهتر مورد ارزیابی قرار گرفته است، شنوندگان حرفه ای به معیار کیفیت کلی خروجی سیستم امتیاز ۴,۰۱ را داده اند که تمامی این موارد حاکی از کیفیت بسیار مناسب و خوب این سیستم سنتز می باشد. انتظار می رود که با بهینه سازی روشهای آموزش وزنها و استفاده از روشهای اتوماتیک کیفیت خروجی سیستم بهبود یابد. علاوه بر این می توان امیدوار بود که استفاده از الگوریتمهای اصلاح نوا همچون الگوریتم TD-PSOLA موجب بهبود کیفیت خروجی سیستم گردد. یکی از پارامترهای بسیار مهم در تعیین میزان کیفیت خروجی، اندازه دادگان و میزان دقت تقطیع آنها می باشد. یکی از مشکلات عمده سیستم کنونی عدم اصلاح دستی تقطیع یک ساعت از دادگان می باشد که موجب ایجاد بعضا خروجیهایی با تلفظ آوایی غلط می گردد و لذا لازم است که تمامی دادگان بصورت دستی اصلاح گردد.

مراجع

[1] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR – ν – TALK speech synthesis system," *ICSLP*, vol. 1, pp. 483–486, 1992.

[2] N. Campbell and A. Black, "CHATR: a multi-lingual speech re-sequencing synthesis system," *Institute of Electronic, Information and Communication Engineers*, 1996.

[3] Hunt, and A. Black, "Unit Selection in A Concatenative Speech Synthesis System Using A Large Speech Database," *ICASSP*, pp. 373-376, 1996.

[4] P. Taylor, A. W. Black, and R. Caley, "The architecture of the Festival speech synthesis system," *The Third ESCA Workshop in Speech Synthesis*, pp. 147–151, 1998.

[5] Clark, R. A. J., Richmond, K., and King, S., *Festival 2 - Build Your Own General Purpose Unit Selection Speech Synthesiser*, *5th ISCA Speech Synthesis Workshop*, pp. 173-178, 2004.

[6] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou and A. Syrdal, "The AT&T Next-Gen TTS System," *137th Acoustical Society of America Metting*, 1999.

[7] Black and K. Tokuda, "Blizzard Challenge-2005: Evaluating Corpus-Based Speech Synthesis on Common Datasets," *Interspeech*, 2005.

[8] P. Rutten, M. Aylett, J. Fackrell, and P. Taylor, "A Statistically Motivated Database Pruning Technique for Unit Selection Synthesis," *ICSLP*, pp. 125-128, 2002.

[9] W. Hamza1, and R. Donovann, "Data-Driven Segment Preselection in the IBM Trainable Speech Synthesis System," *ICSLP*, pp. 2609-2612, 2002.

[10] Kishore, S. P., and Black, A. W., "Unit Size in Unit Selection Speech Synthesis," *ICASSP*, pp. 1317-1320, 2003.

[11] W. Black, and P. Taylor, "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis," *ICASSP*, vol. 2, pp. 601-604, 1997.

[12] Black and K. Lenzo, "Optimal data selection for unit selection synthesis," *4th ESCA Workshop on Speech Synthesis*, 2001.

[13] M. Beutnagel, M. Mohri, and M. Riley, "Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis," *Speech Communication and Technology*, vol. 2, pp. 607-610, 1999.

[14] T. Toda, H. Kawaiz, and M. Tsuzakiz, "Optimizing Sub-Cost Functions for Segment Selection Based on Perceptual Evaluations in Concatenative Speech Synthesis," *ICASSP*, vol. 1, pp. 657-700, 2004.

[15] Di´az, F. C., Alba, J. L., Banga, E. R., "A Neural Network Approach for the Design of the Target Cost Function in Unit-Selection Speech Synthesis," *INTERSPEECH*, pp. 2533-2536, 2005.

[16] Rouibia, S., Rosec, O., "Unit Selection for Speech Synthesis Based on a New Acoustic Target Cost," *INTERSPEECH*, pp. 2565-2568, 2005.

[17] Di´az, F. C., Banga, E. R., "On the Design of Cost Functions for Unit-Selection Speech Synthesis," *EUROSPEECH*, pp. 289-292, 2003.

[18] Park, S. S., Kim, C. K., and Kim, N. S., "Discriminative Weight Training for Unit-Selection Based Speech Synthesis," *ICASSP*, pp. 281-284, 2003.

[19] J. Vepa, S. King, "Subjective Evaluation of Join Cost & Smoothing Methods," *5th ISCA Speech Synthesis Workshop*, pp. 7-12, 2004.

[20] Nukaga, N., Kamashida, R. and Nagamatsu, K., "Unit Selection Using Pitch Synchronous Correlation for Japanese Concatenative Speech Synthesis," *5th ISCA Speech Synthesis Workshop*, pp. 43-48, 2005.

[21] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Segment Selection Considering Local Degradation of Naturalness in Concatenative Speech Synthesis," *ICASSP*, vol. 1, pp. 696-699, 2003.

[22] Black, and N. Campbell, "Optimizing Selection of Units from Speech Databases for Concatenative Synthesis," *ICSLP*, pp. 581-584, 1995.

[23] Conkie A., "A Robust Unit Selection System for Speech Synthesis," *137th meet. ASA/Forum Acusticum*, 1999.

[24] K. Wang and T. Gasser, "Alignment of Curves by Dynamic Time Warping," *The Annals of Statistics*, vol. 25, no. 3, pp. 1251-1276, 1997.

[25] P. Taylor, R. Caley, A. Black, S. King, *Edinburgh Speech Tools Library System Documentation Edition 1.2*, 15th June 1999.

دکتر محمد مهدی همـایونپور 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 (۱۳۷۴) در دانشگاه سوربون جدید در فرانسه و همزمان دورهٔ دکترای خود را در دانشگاه پاریس ۱۱ در زمینه مهندسی برق (۱۳۷۴) بپایان 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. ایشان علاوه بـر تـدریس، راهنمـایی پـروژه هـای کارشناسـی، کارشناسی ارشد و دکتری در زمینه های مهندسی کامپیوتر و فناوری اطلاعات و نیز هدایت تعداد زیادی پروژه های صنعتی و ملی را عهده‌دار بوده است. نـامبرده عضو انجمن های علمی کامپیوتر، ارتباطات و فناوری اطلاعات و رمز می باشـد و مسئولیت های اجرایی متعدد از جمله ریاست و معاونت های آموزشی و پژوهشی دانشکده مهندسی کامپیوتر و فناوری اطلاعـات دانـشگاه صـنعتی امیـر کبیـر و شرکت در برگزاری چندین کنفرانس و مسابقه علمی را بر عهده داشـته و موفـق به انتشار بیش از ۱۵۰ مقاله علمی- پژوهشی در مجلات و کنفرانس های علمـی داخل و خارج از کشور گردیده است.

آدرس پست الکترونیکی ایشان عبارت است از:

homayoun@aut.ac.ir

مجید نم نبات در سال ۱۳۸۲ مدرک کارشناسی را در رشته سخت افزار کامپیوتر از دانشگاه شاهد و در سال ۱۳۸۵ مدرک کارشناسی ارشد در رشته هوش مصنوعی را از دانشگاه امیرکبیر دریافت نمود. ایشان مدت ۲ سال مسئولیت بخش پردازش سیگنال سیستم تبدیل متن به گفتار در شرکت ایران شگرف را بر عهده داشته است. ایشان بیش از ۱۸ مقاله کنفرانسی و مجله ای در زمینه های مختلف پردازش گفتار ارائه نموده است. زمینه های تخصصی مورد علاقه ایشان سیستمهای تبدیل متن به گفتار علی الخصوص بخشهای پردازش نوایی و سنتز گفتار می باشد.

آدرس پست الکترونیکی ایشان عبارت است از:

maj.nam@gmail.com

---

[1] Corpus-based Speech Synthesis
[2] Speech Re-sequencing Synthesis
[3] Black and Campbell
[4] Target Cost and Concatenation (Transition) Cost
[5] Weight space search
[6] Regression training
[7] Time Aligned
[8] Dynamic Time Warping (DTW)
[9] Utterance
[10] Multiple Linear Regression
[11] Re-synthesis
[12] Cache
[13] Pruning
[14] Beam Width Constraint