بهبود کیفیت سیستمهای تبدیل گفتار[1] مبتنی بر مدل GMM[2]

مهدی اسلامی[1]، ابوالقاسم صیادیان[2*]

۱- دانشجوی دکتری، دانشکده مهندسی برق، دانشگاه صنعتی امیر کبیر

۲- دانشیار گروه مخابرات، دانشکده مهندسی برق، دانشگاه صنعتی امیر کبیر

* تهران، صندوق پستی: ۴۴۱۳- ۱۵۸۷۵

m_eslami@aut.ac.ir

چکیده- در سیستمهای تبدیل گفتار، گوینده A جملاتی را بیان می‌کند و هدف عبارت است از تغییر متکلم جملات بیان شده، از گوینده A (مبدأ) به گوینده مورد نظر B (مقصد). البته در مواردی به‌جای تبدیل گفتار از عبارت تبدیل گوینده[3] نیز استفاده می‌شود. تغییر گوینده باید به قسمی انجام پذیرد که سیگنال تغییر یافته کیفیت مطلوب و طبیعی، اما با صدای گوینده B داشته باشد. سه روش مورد استفاده در این سیستمها عبارتند از: روشهای مبتنی بر چندی سازی برداری[4]، روشهای مبتنی بر تبدیلهای LMR[5] و روشهای مبتنی بر مدل مخلوط گاوسی (مدل آماری). در تمامی روشهای ذکر شده برای انطباق زمانی جفت کلمات یا جفت جملات متناظر دو گوینده (در مرحلهٔ یادگیری) از روش پیچش زمانی پویا[6] استفاده می‌شود. در طی این تحقیق ضمن بررسی روشهای موجود، از تکنیک انطباق زمانی DTW برای طراحی توابع تبدیل جفت واجهای متناظر دو گوینده (به‌جای جفت کلمات یا جملات) استفاده می‌شود. این کار موجب انطباق بیشتر دو گوینده در کوچکترین واحد زبانی (یعنی واج) می‌شود. همچنین به منظور کاهش خطا، از تبدیلهای خطی موقتی وابسته به واج در مرحله یادگیری استفاده می‌شود. با اصلاحات مناسب دیگری که در روش یادگیری و طراحی تبدیلهای خطی مورد نیاز انجام شده، به عملکرد بسیار مناسبی در تبدیل گفتار در مقایسه با روشهای رایج نائل شده‌ایم.

کلید واژگان: تبدیل گفتار، تغییر گوینده، نگاشت طیفی[7]، مدل مخلوط گاوسی.

۱- مقدمه

الگوریتمهای تبدیل گفتار (یا تبدیل گوینده)، در ابتدا به منظور انطباق (یا وفق دادن) گوینده جدید در سیستمهای بازشناسی گفتار[8] ارائه شده و مورد استفاده قرار گرفته‌اند[۱]. از جمله کاربرد این الگوریتمها، انطباق بیشتر الگوهای مرجع به مشخصات گوینده جدید به منظور افزایش دقت عملکرد سیستمهای بازشناسی گفتار بوده است. با توسعه الگوریتمهای تبدیل گفتار، این روشها در سیستمهای تبدیل متن به گفتار[9] یا TTS نیز مورد استفاده قرار گرفته است[۳،۲]. در سیستمهای TTS کلیه الگوهای از قبل ذخیره شده، نوعاً متعلق به یک یا دو گوینده (مرد یا زن) می‌باشند. چنانچه بخواهیم توسط این سیستمها صدای گوینده دیگری (غیر از دو گوینده اصلی) را تولید کنیم، ناگزیر از الگوریتمهای تبدیل گفتار استفاده می‌کنیم.

1.Voice Conversion (VC)
2. Gaussian Mixture Model (GMM)
3. Speaker Transformation
4. Vector Quantization
5. Linear Multivariate Regression
6. Dynamic Time Warping (DTW)
7. Spectral Transformation
8 Speech Recognition Systems
9. Text to Speech

علاوه بر دو کاربرد عمده ذکر شده، در سالهای اخیر تلاش زیادی شده تا از الگوریتمهای تبدیل گفتار در کاربردهای صداگذاری[1] برای فیلمها و انیمیشن‌ها استفاده شود. در کاربرد اخیر هدف آن است که فرد صداگذار، توانایی تولید صدای گویندگان متعدد را با کیفیت طبیعی داشته باشد. تحقیق گزارش شده در این نوشتار در راستای تحقق هدف اخیر انجام شده است.

بلوک دیاگرام سیستم تبدیل گفتار متداول در شکل ۱ نشان داده شده است که مطابق آن، ابتدا نواحی واکدار، بی واک و سکوت سیگنال گفتاری تعیین و در قسمت بعد، سیگنال گفتار به بخش تحلیل کننده وارد می‌شود. در این قسمت پارامترهای طیفی و عروضی[2] قابهای سیگنال گفتار استخراج و در مرحله بعد، با استفاده از توابع نگاشت، ویژگیهای طیفی و همچنین ویژگیهای عروضی گفتار گوینده مبدأ و مقصد به یکدیگر تبدیل می‌شوند. در نهایت با استفاده از پارامترهای جدید، گفتار مورد نظر بازسازی می‌شود. بنابراین، هر سیستم تبدیل گفتار دو مرحله عمده به شرح زیر دارد: الف– مرحله یادگیری (طراحی توابع تبدیل دو گوینده) ب– مرحله تبدیل گفتار (یا تبدیل گوینده). هدف از یادگیری، طراحی تعداد مناسبی از توابع تبدیل تناظر[3] یا کتابهای کد تناظر[4] است.

برای اجرای هدف فوق، به تعدادی کلمات یا جملات مناسب از گوینده مبدأ و مقصد نیاز است. در بیشتر روشها، برای انطباق زمانی دو کلمه یا دو جمله متناظر دو گوینده (در مرحله آموزش)، از روش DTW استفاده می‌شود. پس از انطباق زمانی، فرایند تخمین توابع تبدیل یا کتابهای کد ارتباط، به روشهای مختلفی انجام می‌شود. در مرحله تبدیل گفتار، ابتدا قابهای سیگنال[5] تلفظ شده توسط گویندهٔ A، با استفاده از کتاب کد آن شماره‌گذاری می‌شود. در مرحله بعد، به کمک کتابهای کد تناظر، تبدیل مناسب گوینده B برای هر کلمه کد A انتخاب و به پارامترهای طیفی و تحریک آن اعمال می‌شود. اگر چه روشهای فعلی تبدیل گفتار در تولید گفتار قابل فهم برای گوینده جدید، موفقیتهایی نسبی کسب کرده‌اند، اما تا وصول به گفتار با کیفیت طبیعی و به دور از صداهای ناخوشایند[6]، هنوز راه زیادی در پیش است. آنچه در این تحقیق انجام شده، تلاش در جهت بهبود عملکرد این روشها به منظور تولید گفتار طبیعی‌تر است. در بخش دوم این نوشتار ضمن تشریح مختصر روشهای موجود، به نقاط قوت و ضعف آنها اشاره خواهیم کرد. در بخش سوم مبانی روش پیشنهادی در این تحقیق ارائه خواهد شد. در بخش چهارم نتایج شبیه‌سازی و پیاده سازی برای سه روش موجود و همچنین روش جدید پیشنهادی، مطرح خواهد شد.

در بخش پنجم جمع‌بندی و نتیجه‌گیری از این تحقیق بیان خواهد شد.

۲- بررسی روشهای مطرح در تبدیل گفتار

روشهای مطرح و موجود در تبدیل گفتار عبارتند از:

الف– VQ-VC

ب– LMR-VC

ج– GMM-VC

برای بررسی نحوه استفاده از نقاط قوت هر یک از روشهای مذکور، در اینجا شرح مختصری از هر یک ارائه می‌شود.

۲-۱- تبدیل گفتار به روش VQ_VC [۲-٤]

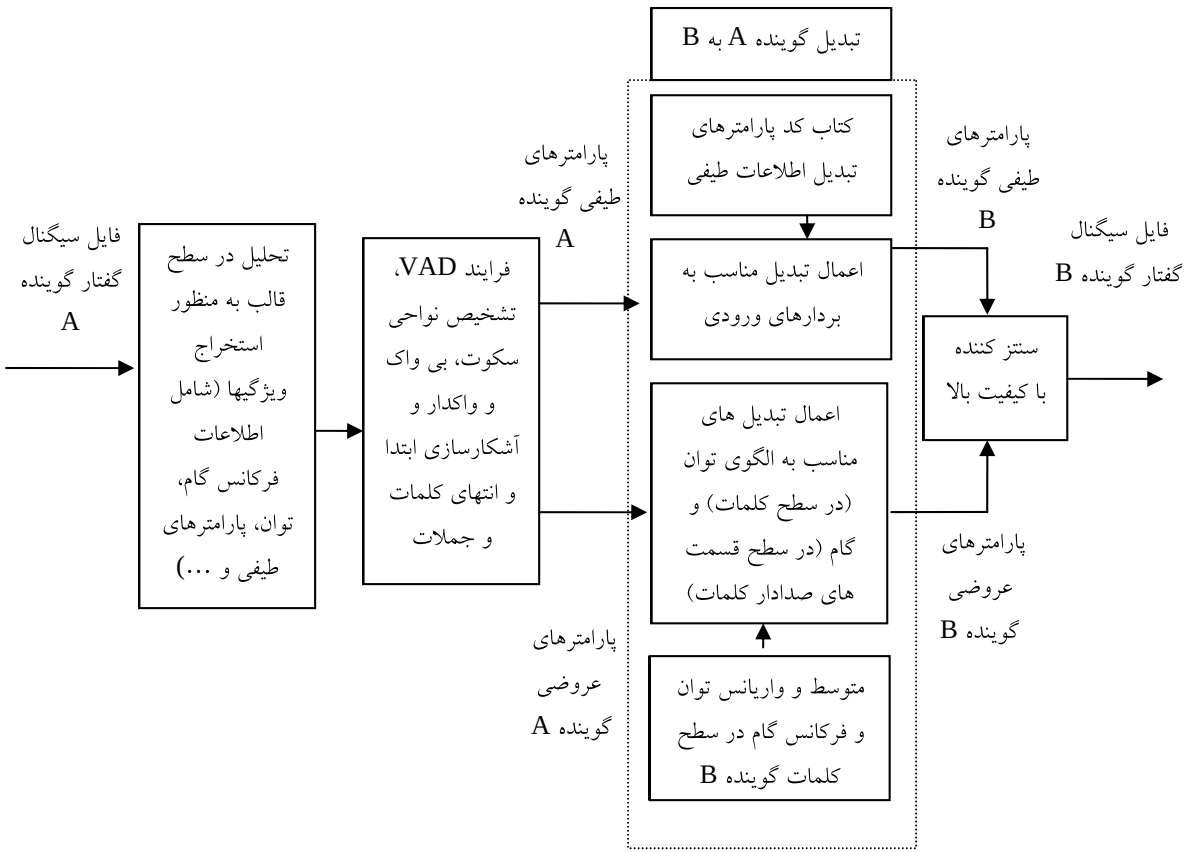
در این روش، تبدیل گفتار بدون جداسازی پارامترهای آوایی انجام می‌شود. ایده اولیه این روش، استفاده از کتابهای کد برای پارامترهای آوایی است. این کتابهای کد تمامی اطلاعات مربوط به هویت گوینده گفتار را در بر دارند. بنابراین تبدیل خصیصه‌های آوایی یک گوینده به

1. Dubbing
2. Prosodic
3. Correspondence Transform Function
4. Correspondence Codebooks
5. Signal Frames
6. Artifact

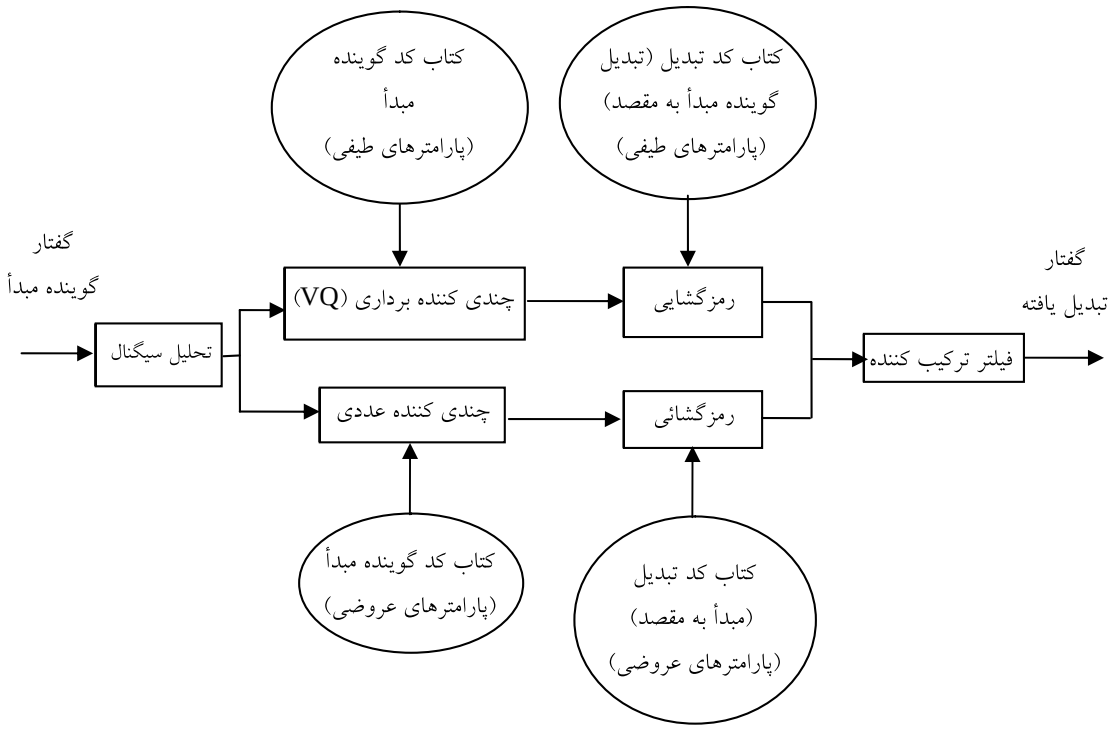
دیگری، به مسأله نگاشت کتابهای کد دو گوینده کاهش می‌یابد. این روش نیز مانند سایر روشها دو مرحله دارد: مرحله آموزش، یک مرحله تبدیل، ساخت و سنتز گفتار. مرحله یادگیری، فرایندی برای تولید کتابهای کد است. مرحله تبدیل و ساخت، فرایندی است برای ساختن گفتار با استفاده از کتاب نگاشت کد. ذکر این نکته ضروری به‌نظر می‌رسد که در تبدیل گفتار به روش VQ_VC، سه کتاب کد موجود است: کتاب کد گوینده مبدأ، کتاب کد گوینده مقصد و کتاب کدی که نگاشت میان دو کتاب قبلی را نشان می‌دهد. در شکل۲ بلوک دیاگرام سیستم تبدیل گوینده مبدأ به گوینده مقصد با استفاده از تبدیل کتاب کد نشان داده شده است.

در ادامه، مراحل یادگیری و تبدیل گفتار با استفاده از این روش خواهیم داد.

مرحله یادگیری: در این مرحله گوینده A و B مجموعه لغات آموزشی را (بیش از ۱۰۰ کلمه) تلفظ می‌نمایند. با استفاده از مجموعه لغات آموزشی، یک کتاب کد برای گوینده A و یک کتاب کد برای گوینده B طراحی می‌شود که معمولاً شامل ۲۵۶ تا ۱۰۲۴ بردار مرجع است. سپس کلیه قابهای سیگنالها با شاخص کتاب کد متناظر گوینده جایگزین می‌شود. تناظر بین بردارهای هر جفت کلمه یکسان از دو گوینده، به روش DTW تعیین می‌شود. فرایند فوق برای تمامی کلمات متناظر دو گوینده انجام می‌شود. از طرفی وابستگی بین بردارهای مرجع گوینده A و بردارهای مرجع گوینده B به‌صورت هیستوگرام تجمعی در نظر گرفته می‌شود. برای مثال بردار مرجع شماره ۱۰۰ کتاب کد گوینده A، به سه بردار شماره ۲۱، ۳۲ و ۲۰۰ کتاب کد گوینده B با تکرار ۲، ۵ و ۸ نسبت داده می‌شود.



شکل ۱ بلوک دیاگرام سیستم کلاسیک تبدیل گفتار



شکل ۲ بلوک دیاگرام تبدیل گوینده مبدأ به گوینده مقصد به روش VQ-VC

بدین ترتیب با استفاده از هیستوگرام حاصل برای هر بردار مرجع A، یک تابع وزنی نرمالیزه شده برای ترکیب خطی بردارهای متناظر از گوینده B به‌دست می‌آید. با ترکیب خطی بردارهای متناظر از گوینده B برای هر بردار مرجع A، یک بردار تبدیل ساخته می‌شود. مجموعه بردارهای تبدیل را که به شرح بالا حاصل می‌شود کتاب کد تبدیل[1] نامیده می‌شود. در مرحله تبدیل گفتار، کلمات کد گوینده A، توسط کلمات کد کتاب تبدیل تعویض می‌شود. چنانچه اعوجاج متوسط بین مجموعه سیگنالهای دو گوینده از سطح آستانه‌ی نسبی کمتر نشد، فرایند DTW بار دیگر اعمال می‌شود (این بار با شاخص کتاب کد تبدیل فعلی و کتاب کد گوینده B). فرایند DTW و کلیه مراحل ذکر شده تا وصول به همگراییمطلوب و مورد نظر ادامه یافته و در پایان فرایند بهینه سازی، کتاب کد تبدیل نهایی حاصل می‌شود. برای فرکانس گام[2] و توان[3] (قابهای متناظر از دو گوینده)، دو کتاب کد اسکالر متمایز به شرح بالا طراحی می‌شود.

مرحله تبدیل گفتار: در این مرحله، گوینده A جمله مورد نظر را بیان می‌کند. سپس قابهای سیگنال توسط کتاب کد گوینده A چندی‌سازی می‌شود. برای هر بردار مرجع در سیگنال چندی شده، بردار مرجع متناظر از کتاب کد تبدیل جایگزین و آنگاه سیگنال جدید توسط پارامترهای بردارهای تبدیل جایگزین شده تولید می‌شود. در این صورت، مشخصه گوینده سیگنال سنتز شده به گوینده B شبیه خواهد بود.

۲-۲- تبدیل گفتار به روش LMR –VC[۸-٤]

مرحله یادگیری: مانند روش قسمت قبل، دو گوینده A و B تعدادی کلمه یا جمله آموزشی را تلفظ می‌کنند. فضای پارامترهای طیفی و تحریک دو گوینده به روش VQ به M کلاس خوشه بندی می‌شوند (مقدار پارامتر M بین ۶۴ تا ۲۵۶ انتخاب می‌شود). به روش

1. Mapping Codebook
2. Pitch Frequency
3. Energy

DTW تناظر کلاسهای دو گوینده به‌دست می‌آید. بدین ترتیب مانند روش VQ_VC، کتاب کد تبدیل گوینده B قابل محاسبه است. فرض می‌کنیم که $C_A(i)$ بردار مرجع کلاس i گوینده A و $C_B(j)$ بردار مرجع تبدیل j متناظر با آن باشد. همچنین فرض می‌کنیم که $\sum_A(i)$ و $\sum_B(j)$ ماتریس کواریانس کلیه بردارهای آموزشی متعلق به کلاس i و j از گوینده A و B باشند. در این صورت، تبدیل خطی بین کلاس (i , j) به شرح زیر به‌دست خواهد آمد:

$$(i , j) = \Gamma(i , j).\sum_A^{-1}(i) \qquad (۱)$$

T

در این رابطه، $\Gamma(i , j)$ ماتریس کواریانس متقابل بین بردارهای آموزش کلاس i (از گوینده A) و کلاس j (از گوینده B) است. کلیه بردارهای آموزشی متعلق به کلاس iام گوینده A (مانند $X_A(i)$) را با تبدیل زیر به کلاس jام از گوینده B شبیه‌تر می‌کنیم:

$$X_B(j) = T(i , j) .[X_A(i) - C_A(i)] + C_B(j) \qquad (۲)$$

تبدیل فوق برای کلیه بردارها و برای تمامی کلاسهای گوینده A انجام می‌شود. با اعمال تبدیلهای LMR به کلیه بردارهای سیگنال گوینده A و با استفاده مکرر از روش DTW، تناظر جدید بین کلاسها به‌دست می‌آید. فرایند مذکور تا وصول به همگرایی مطلوب در تابع اعوجاج تجمعی DTW ادامه می‌یابد.

مرحله تبدیل گفتار: در این مرحله، از گوینده A خواسته می‌شود تا جمله مورد نظر را تلفظ نماید. سپس قابهای سیگنال توسط کتاب کد گوینده A کلاس‌بندی می‌شود. برطبق رابطه (۲) کلیه بردارهای متعلق به هر کلاس گوینده، توسط بردار متوسط و ماتریس تبدیل متناظر انتقال یافته و چرخش لازم اعمال می‌شود. سیگنال جدید توسط پارامترهای تبدیل یافته سنتز می‌شود که در نتیجه این سیگنال، حاوی مشخصه گوینده B خواهد بود.

۲-۳ – تبدیل گفتار به روش GMM [۹]

مدل مخلوط گاوسی، روشی کاملا پایدار برای نشان دادن ویژگی‌های آوایی گوینده است. این مدل ترکیبی از چندین مدل گاوسی است. به‌طور ساده می‌توان گفت که قله‌های گاوسی در چگالی طیف این مدل، همان محل تجمع بردارهای مربوط به یک آوای خاص است. این موضوع در واقع یکی از دلایل عمده ای است که ما را به استفاده از GMM به منظور بیان فضای آوایی گوینده تشویق می‌نماید. این مشاهدات نشان می‌دهند که ترکیب خطی توابع گاوسی پایه، توانایی توصیف دسته بزرگی از توزیع‌ها را دارد. در آموزش مدل گاوسی، هدف است که پارامترهای مدل با استفاده از داده‌های آموزشی موجود تخمین زده شود تا بهترین تطبیق بر روی بردارهای ویژگی گوینده به‌دست آید.

برای تخمین پارامترهای مدل گاوسی روشهای متعددی وجود دارد. در این میان یکی از متداول ترین روشها، تخمین بیشینه درستنمایی[1] است. هدف از تخمین ML یافتن پارامترهای مدل است به گونه ای که شباهت GMM را به بردارهای آموزشی بیشینه سازد. در ادامه مراحل یادگیری و تبدیل گفتار با استفاده از این روش را بررسی خواهیم کرد. شکلهای ۳ و۴ نحوه یادگیری را برای روش GMM نشان می‌دهند.

مرحله یادگیری: در این روش فضای ویژگیهای هر گوینده توسط مدل آماری GMM به‌صورت زیر نمایش داده می‌شود:

$$P(X) = \sum_{i=1}^{M} \alpha_i . N(X; \mu_i, \sum_i) \qquad (۳)$$

که در آن $P(X)$ توزیع احتمال بردار X برای گوینده A است. تابع توزیع $N(X; \mu_i, \sum_i)$ توزیع نرمال با بردار متوسط μ_i و ماتریس کواریانس $\sum_i$ است. α_iها مقادیر اسکالر مثبت نرمالیزه شده‌ای هستند که به‌عنوان تابع وزنی کلاسهای مختلف فضای ویژگیها مورد استفاده قرار

1. Maximum Likelihood

می‌گیرند ($\sum_{i=1}^{M}\alpha_i=1$). نحوه مدلسازی GMM برای ویژگیهای هر گوینده در [۹،۱۰] به تفصیل آورده شده است. در مدل GMM، هر کلاس توسط بردار متوسط μ_i و ماتریس کواریانس Σ_i توصیف می‌شود. توابع وزنی مخلوط‌ها $\{\alpha_i\}$ها نماینده فرکانس نسبی هر کلاس در مجموعه بردارهای آموزشی سیستم است. احتمال شرطی اینکه بردار X به کلاس خاص C_i تعلق داشته باشد، از رابطه زیر به‌دست می‌آید:

$$P(c_i \mid X)=\frac{\alpha_i.N(X;\mu_i,\Sigma_i)}{\sum_{j=1}^{M}\alpha_j.N(X;\mu_j,\Sigma_j)} \qquad (۴)$$

پارامترهای مدل GMM با استفاده از الگوریتم EM[1] تخمین زده می‌شوند[۱۰]. این روش برای تخمین پارامترهای مدل به‌کار می‌رود. در شکل ۳ نحوه پیاده سازی تابع تبدیل نشان داده شده است.

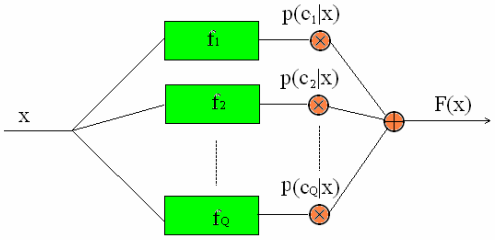
مقدار اولیه $\alpha=\frac{1}{M}$ و برای μ برابر با M بردار کد تولید شده توسط الگوریتم VQ و در نهایت کواریانس Σ برابر ماتریس یکه در نظر گرفته می‌شود. سپس با استفاده از الگوریتم EM، تا زمانی که $P_{GMM}(X;\alpha,\mu,\Sigma)$ مقدار بیشینه خود را به دست آورد یا تعداد دفعات اجرای الگوریتم از سطح آستانه‌ای بیشتر شود، ادامه می‌یابد.

در طی اجرای EM، باید مراقب باشیم تا ماتریس کواریانس، به مقادیر تکین نزدیک نشود. این کار با افزودن یک ماتریس قطری با مقادیر ثابت، پس از هر بار تکرار انجام می‌شود. هدف از مرحله یادگیری روش GMM_VC، تخمین تابع تبدیل به شکل زیر است:

$$CF(X)=\sum_{i=1}^{M}P(c_i \mid X).[V_i+\Gamma_i.\Sigma_i^{-1}.(X-u_i)] \qquad (۵)$$

1. Expectation Maximization

که در آن $V_i=E\{Y\}$ بردار متوسط بردارهای آموزشی گوینده B متناظر با کلاس i گوینده A است.



شکل ۳ نمایش نحوه پیاده سازی تابع تبدیل توسط مجموعه‌ای از مخلوطهای وزن دار شده

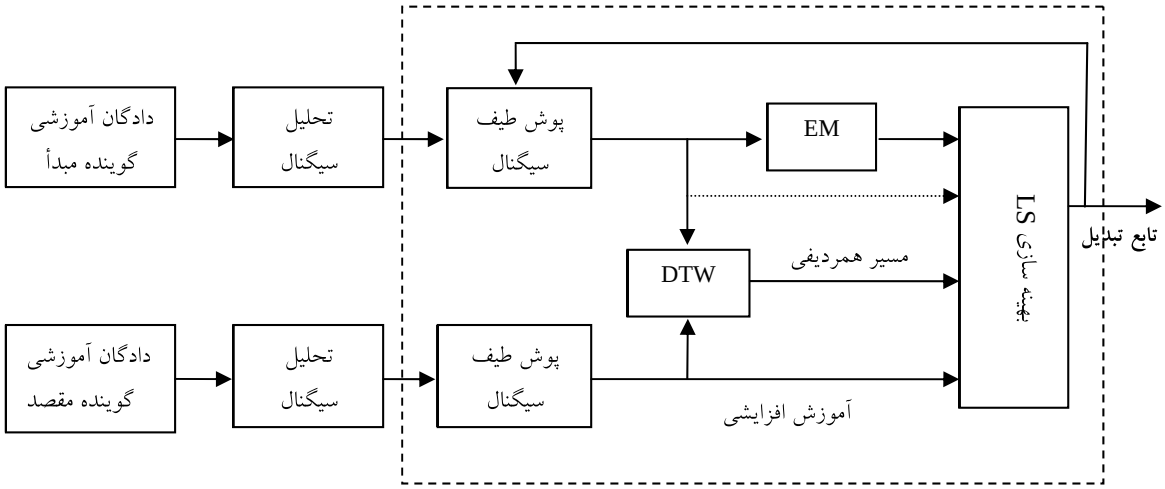
Γ_i ماتریس کواریانس متقابل بین مجموعه بردارهای آموزشی دو کلاس متناظر از دو گوینده A وB است:

$$\Gamma_i=E\{(Y-V_i).(X-\mu_i)^T\} \qquad (۶)$$

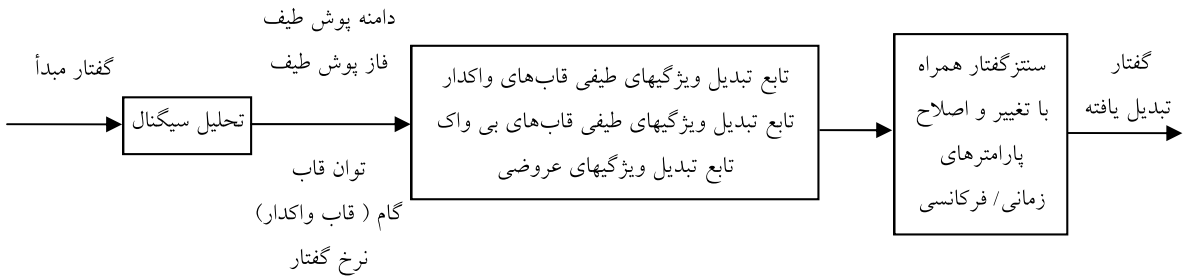
روش متناظر کردن، مانند سایر روشها، توسط روش DTW انجام می‌شود. ملاحظه می‌شود که تابع تبدیل روش GMM (رابطه ۵) مشابه روش LMR–VC است؛ لیکن در بازنمایی توابع تبدیل، پارامترهای تمامی کلاسها (C_iها) به نسبت α_iها دخالت دارند (برخلاف روش LMR که برای هر کلاس از گوینده B صرفاً پارامترهای یک کلاس ازA گوینده دخالت دارند). روش بهینه سازی برای وصول به بهترین تناظر و محاسبه V_iها، μ_iها، Σ_iها و Γ_iها در مرجع [۹] به تفصیل بیان شده است.

مرحله تبدیل گفتار: چنانچه پارامترهای تابع تبدیل مدل GMM در مرحله یادگیری تخمین زده شود، مرحله تبدیل گفتار توسط رابطه (۵) به سهولت قابل انجام است. بدین ترتیب هر بردار X از سیگنال متعلق به گوینده A توسط رابطه (۵) به برداری معادل برای گوینده B تبدیل می‌شود. فرایند بالا برای کلیه بردارهای جمله تلفظ شده توسط گوینده A انجام می‌شود. سیگنال جدیدی که توسط پارامترهای بردارهای تبدیل یافته سنتز شود، حاوی مشخصه گوینده B خواهد بود. بلوک

دیاگرام سیستم مورد استفاده در تبدیل گفتار به روش مدل GMM در شکل ۵ نشان داده شده است.



شکل ۴ نحوه یادگیری برای روش GMM_VC



شکل ۵ بلوک دیاگرام مورد استفاده در سیستم تبدیل گفتار برای روش GMM_VC

۳- روشی جدید برای تبدیل گفتار

روش VQ_VC به‌عنوان روش پایه در تبدیل گفتار، مزایایی از نظر سادگی تحلیل و هزینه محاسباتی در مرحله یادگیری و در مرحله تبدیل گفتار دارد. متأسفانه این روش، عیب گسسته بودن فضای پارامترها را (به تعداد محدودی بردار مرجع) داشته و در نتیجه از نویز چندی‌سازی بردارها تأثیر سوء می‌پذیرد (که موجب کاهش کیفیت می‌شود). روش VC_LMR مشکل بالا را به نحو مناسبی جبران می‌کند اما در نقاط گذرای بین خوشه‌های مربوط به کلاسها، حالت ناپیوستگی دارد. پدیده ناپیوستگی موجب ایجاد صدای کلیک‌دار ناخوشایند در سیگنال سنتز شده، می‌شود. روش GMM_VC روش جامع‌تری نسبت به دو روش قبلی است و دو مشکل مطرح شده را به نحو مناسبی جبران می‌سازد. این روش به‌علت دخالت دادن تمامی خوشه‌ها در تولید بردار برای گوینده جدید، حالت بلورشدگی[1] (کاهش وضوح) در صدای بازسازی شده دارد. در حال حاضر، بازسازی صدا (با تغییر گوینده) توسط روشهای مذکور نسبت به حالت کاملاً طبیعی، فاصله زیادی دارد. در اغلب تحقیقات مطرح شده تاکنون، قابل فهم بودن سیگنال بازسازی شده و شباهت بیشتر به گوینده B ملاک عمل بوده است [۸، ۹]. بنابراین تلاش در جهت بهبود کیفیت و طبیعی‌تر کردن صدای بازسازی شده، ارزش

1. Smoothing

تحقیقاتی دارد. با توجه به اینکه روش GMM_VC توسعه دو روش قبلی بوده و معایب دو روش قبلی را نیز تا حدودی جبران می‌کند، در این تحقیق به‌عنوان روش پایه برای بهبود عملکرد سیستمهای تبدیل گفتار (VC) مورد استفاده قرار گرفته است.

در بیشتر روشهای ذکر شده، تناظر بین مرکز خوشه‌های دو گوینده (μ_i, ν_i) با استفاده از روش DTW تعیین می‌شود. DTW به‌عنوان یک روش پیچش زمانی غیرخطی برای انطباق زمان دو کلمه از یک گوینده، عملکرد خوبی دارد؛ لیکن هنگامی که فضای پارامترهای دو کلمه از یکدیگر فاصله زیادی داشته باشند (در تلفظ کلمه توسط یک مرد و یک زن یا بچه)، رفتار DTW در انطباق زمانی قابل اعتماد نیست. علت آن است که تابع اعوجاج تجمعی DTW خاصیت خود را برای انطباق زمانی با حفظ قابلیت تمایز پذیری وقایع اکوستیکی مشابه و نزدیک به هم (واجهای مشابه) در طول کلمه یا جمله، تا حدودی از دست می‌دهد. این نکته یکی از منابع اصلی تولید خطا در استفاده از روش DTW است که تاکنون کمتر مورد توجه محققان قرار گرفته است. به منظور فائق آمدن بر دو مشکل اساسی روش GMM در تولید صدای طبیعی یعنی: الف) برطرف کردن حالت بلورشدگی یا کاهش وضوح صدا ب) استفاده مؤثر و قابل اعتماد از روش DTW در فرایند متناظر کردن مرکز خوشه‌های دو گوینده، به روش جدیدی را برای آموزش توابع تبدیل و همچنین نگاشت ویژگیهای طیفی/ عروضی ارائه خواهیم کرد.

نحوه آموزش مدل مورد استفاده در روش جدید به شرح زیر است:

الف) برای تهیه دادگان مورد استفاده برای تبدیل گفتار، دو گوینده A و B کلمات (یا جملات) مشخص و مشترکی را با تعدادی تکرار (حداقل ۱۲ بار) در حالتهای بیانی مختلف (آرام، تند، سؤالی، خبری، تعجبی، امری ، عصبانی و...) تلفظ می‌کنند. هر چه تعداد تکرار بیشتر باشد و این تکرار، در طی روزها یا حتی ماههای متفاوت ضبط شود، نتایج بیشتر قابل اعتماد خواهد بود. پس از ضبط جملات، کلیه کلمات یا جملات تلفظ شده دو گوینده به روش با سرپرستی (دستی) زیرنویس واجی[1] می‌شوند.

ب) با استفاده ازمدل HMM آموزش داده شده بر روی واحدهای گفتاری هر گوینده و همچنین الگوریتم ویتربی[2][۱۱]، هر واج به سه حالت[3] متوالی تقسیم می‌شود. انتخاب این سه ناحیه بر اساس تجربیات بازشناسی گفتار انجام می‌شود. بدین ترتیب، واج‌ها به ترتیب به سه ناحیه گذرای ابتدایی، همگون میانی و گذرای انتهایی تقسیم می‌شوند. بنابراین برای هر حالت هر واج، یک مدل GMM تخمین زده می‌شود. با توجه به اینکه صحبت‌کنندگان در این تحقیق فارسی زبان هستند، تعداد واجها برابر ۳۰ بوده و در نتیجه ۹۰ مدل GMM برای هر گوینده طراحی می‌شود (توجه شود که در روش پایه GMM صرفاً یک مدل برای هر گوینده طراحی می‌شود).

ج) درهنگام تعیین تناظر بین خوشه‌های مدل GMM هر حالت دو گوینده، از روش DTW استفاده می‌شود. لیکن قبل از کاربرد روش DTW، یک تبدیل LMR به کلیه بردارهای هر حالت گوینده A اعمال می‌شود، به قسمی که بیشترین شباهت بین دو حالت متناظر دو گوینده A و B در سطح آن حالت ایجاد شود. پس از اعمال تبدیل LMR به روش ذکر شده، از DTW برای یافتن تناظر مرکز خوشه‌های حالتهای متناظر دو گوینده استفاده می‌شود.

د) روش آسان آن است که تعداد خوشه‌های (یا تعداد مخلوطهای[4]) هر مدل GMM را ثابت در نظر بگیریم. یکسان گرفتن تعداد مخلوطها به دلایل زیر روشی بهینه

1. Phoneme Transcription
2. Viterbi Algorithm
3. State
4. Number of Mixture

نیست: ۱- اهمیت واجهای واکدار در تبدیل گفتار به مراتب بیشتر از واجهای بی واک است؛ ۲- تنوع و تغییرات واجهای واکه در متنهای مختلف به مراتب بیشتر از سایر واجهای واکدار است؛ ۳- طول واجهای واکدار بی واکه بسیار کوتاهتر از سایر واجها (بویژه واکه ها) است. بنابراین تعداد مخلوطهای آن بسیار کمتر ازسایر واجها است. بنابراین مدلسازی واجهای واکه در درجه اول اهمیت قرار گرفته و بعد از آن برای واجهای واکدار بی واکه و در نهایت برای واجهای بی واک در درجه سوم قرار می‌گیرد. البته این موضوع در آزمایشهای مکرر شنیداری انجام شده بر روی دادگان مورد استفاده، به اثبات رسیده است. .لذا در طی این تحقیق، تعداد مخلوطهای واجهای بی واک و بی واکه به نسبت ۱، برای واجهای واکه واکدار به نسبت ۲ و برای واکه‌ها به نسبت ۴ در نظر گرفته می‌شوند.

هـ) پس از ضبط کلمات (یا جملات)، توان کلمات (یا جملات) متناظر از دو گوینده را به نحوی نرمالیزه می‌کنیم که توان متوسط (یا توان حداکثر) آنها یکسان باشد. سپس با مدلسازی GMM از نوع اسکالر، توابع تبدیل مربوط به توان و فرکانس گام (برای واجهای واکدار) را در سطح حالات متناظر به‌دست می‌آوریم. از تبدیلهای فوق به‌عنوان توابع تبدیل اطلاعات تحریک دو گوینده A و B استفاده می‌شود.

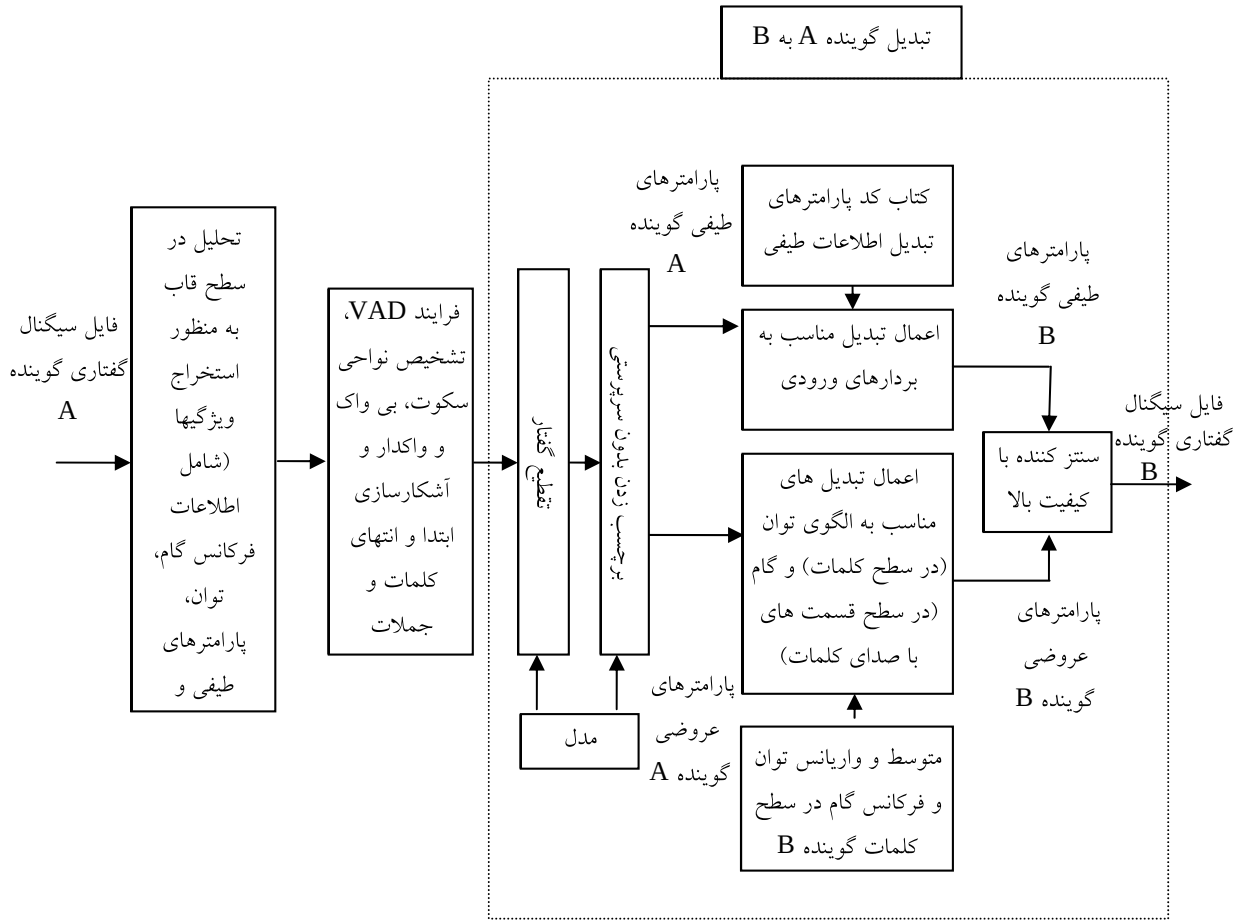
بلوک دیاگرام سیستم تبدیل گفتار جدید در شکل ۶ نشان داده شده است. در این قسمت، نحوه تبدیل گفتار را با استفاده از مدل جدید ارائه شده بررسی خواهیم کرد:

الف) استفاده از VAD برای جداسازی نواحی سکوت از قسمتهای واکدار و بی واک: الگوریتم مورد استفاده برای این کار، MIP_SPED است [۱۴] که الگوریتمی صریح برای تخمین ابتدا و انتهای کلمات گفتاری در محیطهای تمیز و همچنین در محیطهای نویزی است. از جمله ویژگیهای مهم این روش آن است که از تمامی اطلاعات مفید و قابل ارائه سیگنال گفتار شامل پریودیک بودن، بالا بودن سطح انرژی قسمتهای واکدار نسبت به قسمتهای بی واک، اندازه غیر ایستایی قسمتهای گفتار نسبت به نویزهای تداومدار محیطی و غیره، به نحو موثری برای جداسازی گفتار از نویزهای زمینه استفاده می‌کند. یکی از ویژگیهای موثر روش ارائه شده آن است که هم برای نویزهای پریودیک و هم برای نویزهای غیر پریودیک دقت لازم را دارد. از ویژگیهای دیگر این روش، امکان استفاده برای سیستمهای برخط[1] و همچنین برای سیستم دسته‌ای[2] است. از جمله مشکلات عمده روشهای دیگر، تخمین پارامترهای مدل نویز زمینه تداومدار است. در اغلب روشهای کلاسیک، از اطلاعات چندین قاب آغازین ضبط برای مدلسازی نویز استفاده می‌شود. در این صورت، مشخصات نویز زمینه تداومدار با بهترین دقت و مستقل از وقایع آکوستیکی اطراف گفتار تخمین زده می‌شود.

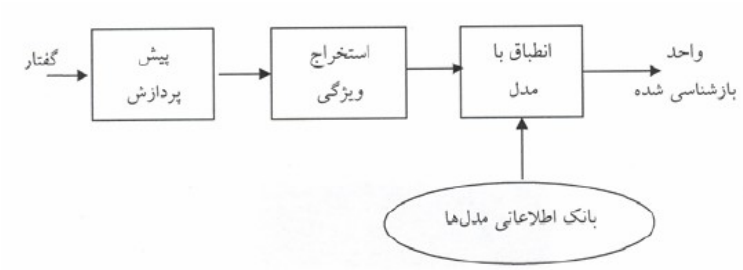
ب) شکل ۷، بلوک دیاگرام کلی سیستم طبقه بندی واحدهای گفتاری را در مرحله بازشناسی نشان می‌دهد. پیش پردازش شامل روشهای بهبود کیفیت سیگنال ورودی، نمونه برداری، تقسیم سیگنال به قابهای کوچک زمانی تقریبا ایستان، جداسازی سیگنال از نویز و مانند آن می‌شود. بدین ترتیب، به ازای هر قاب زمانی از واحد زبانی، یک بردار ویژگی استخراج شده و دنباله بردارهای ویژگی به عنوان شناسه گفتار ورودی به واحدهای مختلف زبانی منطبق می‌شود. ویژگی برجسته این روش، دقت عمل در بازشناسی و در دست داشتن روش یادگیری مطمئن و همگرا برای آموزش پارامترهای مدل از روی دادگان واحدهای بازشناسی است.

ج) جداسازی نواحی واکدار و بی واک و همچنین تعیین فرکانس گام برای نواحی واکدار: روش مورد استفاده در این قسمت، الگوریتم تعیین گام با استفاده از شباهت زمانی بهینه[3] است [۱۵].

1. on-line
2. Batch
3. Optimal Temporal Similarity



شکل ۶ بلوک دیاگرام سیستم جدید تبدیل گفتار



شکل ۷ بلوک دیاگرام سیستم طبقه بندی واحدهای گفتاری

نتایج ناشی از مقایسه این روش با سایر روشهای مطرح، نشاندهنده قدرت بالای الگوریتم در مقابله با نویز سیگنال گفتار و همچنین دقت بالای تخمین نواحی واکدار از نواحی بی واک است[۱۵].

د) اعمال تبدیل بر روی قطعات گفتاری: بدین منظور ابتدا شماره مدل هر حالت را برای هر قاب X در جمله تلفظ شده توسط گوینده A تعیین کرده و سپس از توابع تبدیل متناظر با آن مدل برای تبدیل گفتار استفاده

می‌کنیم. توجه می‌کنید که در روش GMM_VC پایه، صرفاً یک مدل برای هر گوینده وجود دارد (اما در روش جدید ۹۰ مدل)، بنابراین به تخمین شماره مدل نیاز نداریم. برای انجام فرایند تخمین شماره مدل مناسب (برای گوینده A)، احتمال تعلق X به مدل i را برطبق رابطه زیر به‌دست می‌آوریم (برای ۱تا۹۰ =i):

$$P(MD_i \mid X) = \sum_{j=1}^{Mi} \alpha(i,j).N[X;\mu(i,j),\Sigma(i,j)] \quad (۷)$$

M_i تعداد مخلوطهای (خوشه‌های) مدل i ام، $\mu(i,j)$، $\Sigma(i,j)$ و $\alpha(i,j)$ به‌ترتیب بردار متوسط، ماتریس کواریانس و ضریب وزنی خوشه j ام از مدل i است.

مدلی که بیشترین احتمال تعلق به X را داشته باشد، به‌عنوان مدل برنده در نظر گرفته می‌شود. پس از انتخاب مدل، برای تبدیل گفتار به روش GMM_VC پایه عمل می‌شود (با استفاده از توابع متناظر مدلهای دو گوینده).

هـ) تبدیل ویژگیهای عروضی: برای این منظور از مدل خطی استفاده می‌شود.

در این کار، پارامترهای عروضی نسبت به متوسط و انحراف استاندارد ویژگیهای گفتاری گوینده مبدأ نرمالیزه و سپس با مقادیر متناظر از گفتار گوینده مقصد بازسازی می‌شوند.

و) سنتز گفتارخروجی: پس از اعمال تبدیلات طیفی و همچنین تبدیلات عروضی بر روی گفتار گوینده مبدأ، پارامترهای گفتار تبدیل شده به‌دست می‌آید. در این صورت با استفاده از سنتز کننده SM[1] به همراه با روشهای تغییر و اصلاح پارامترهای طیفی/ عروضی و متوالی سازی[2] قاب‌های سنتز شده، قادر به تولید گفتار طبیعی خواهیم بود.

٤- نتایج شبیه‌سازی

روشهای تحلیل و سنتز پارامتری مطرح در زمینه گفتار که تا کنون برای الگوریتمهای تبدیل گفتار استفاده شده‌اند روش LPC [۲-۷][3] و روش HNM [۹][4] است. روش HNM کیفیت بالاتری را در تولید گفتار طبیعی و قابل فهم ارائه کرده است [۹].

در این تحقیق از مدل سینوسی SM با ساختار هارمونیکی در باندهای واکدار استفاده کرده‌ایم [۱۲، ۱۳]. برای تولید ساختار نویزی مرتبط با باندهای بی واک، از فاز تصادفی استفاده شده است . مدل سینوسی با ساختار ذکر شده عملکرد بهتری در مقایسه با مدل HNM (به روش مطرح شده در مراجع ذیربط) دارد. سیگنال سنتز شده با مدل SM، طبیعی بوده و مانند سیگنال گفتار اصلی است. سیگنالهای مورد استفاده، توسط کارت صوتی تجاری و به‌صورت ۱۶ بیتی و با فرکانس نمونه‌برداری ۸ KHZ ضبط شده است. ۲۰۰ جمله برای آموزش سیستم و ۲۰ جمله برای آزمودن و ارزیابی، توسط چهارگوینده ادا و ضبط شده است. جملات آموزشی را به‌صورت دستی برچسب واجی زده‌ایم. جملات آموزشی و آزمون از دو گوینده مرد (با صدای متمایز) و دو گوینده زن (با صدای متمایز) جمع آوری شده است.

برای بررسی و ارزیابی عملکرد روشهای تبدیل گفتار از دو نوع آزمون؛ الف) کمی[5] و ب) کیفی[6] استفاده شد. در آزمون کمی از اعوجاج متوسط (برحسب dB) بین بردار پارامترهای طیفی متناظر دو گوینده قبل و بعد از تبدیل [۹] (در مجموعه بردارهای آموزشی) استفاده کردیم.

در آزمون کیفی از آزمون شنیداری MOS[7] برای مقایسه کیفیت بازسازی جملات اصلی و جملات تبدیل یافته استفاده شده است. برای روش VQ–VC از کتاب کد ۱۰۲۴ کلمه‌ای (برای هر گوینده) و برای روش LMR و GMM به ترتیب از ۲۵۶ و ۱۲۸ خوشه برای هر گوینده استفاده شده است [۹].

1. Sinusoidal Model
2. Frame Concatenation

3. Linear Predictive Coding
4. Harmonic Plus Noise Model
5. Objective Test
6. Subjective Test
7. Mean Opinion Score

جدول ۱ نتایج آزمون کمی اعوجاج متوسط (برحسب dB) بین بردار پارامترهای طیفی متناظر دو گوینده قبل و بعد از تبدیل

		روش VQ_VC	روش LMR_VC	روش GMM_VC	روش جدید
مرد به مرد	قبل از تبدیل	۹/۷۶	۹/۷۶	۹/۷۶	۹/۷۶
	بعد از تبدیل	۴/۲۳	۳/۵۴	۲/۴۸	۱/۱۱
زن به زن	قبل از تبدیل	۸/۸۳	۸/۸۳	۸/۸۳	۸/۸۳
	بعد از تبدیل	۴/۱۲	۳/۱۷	۲/۲۳	۰/۹۳
مرد به زن	قبل از تبدیل	۱۲/۱۷	۱۲/۱۷	۱۲/۱۷	۱۲/۱۷
	بعد از تبدیل	۵/۲۶	۴/۱۱	۲/۹۸	۱/۹۶

جدول ۲ نتایج آزمون شنیداری MOS برای روشهای مختلف تبدیل گفتار

	روش VQ_VC	روش LMR_VC	روش GMM_VC	روش جدید
مرد به مرد	۲/۶۵	۲/۹۴	۳/۱۲	۳/۵۲
زن به زن	۲/۸۱	۳/۰	۳/۲۱	۳/۶۸
مرد به زن	۲/۴۸	۲/۶۹	۲/۹۵	۳/۳۲

تعداد مخلوطهای هر حالت در روش جدید برابر ۶۴، ۳۲ و ۱۶ به ترتیب برای واجهای واکه، بی واکه واکدار و بیواکه بی واک انتخاب کردیم. آزمونها برای سه حالت مختلف: الف) تبدیل گفتاری مرد به مرد ب) تبدیل گفتاری زن به زن ج) تبدیل گفتاری مرد به زن انجام شده است.

نتایج آزمون کمی در جدول ۱ و نتایج آزمون شنیداری MOS در جدول ۲ درج شده است. با مرور نتایج جداول ۱ و ۲ ملاحظه می‌شود که روش جدید هم در آزمونهای کمی و هم در آزمونهای کیفی، عملکرد به مراتب بهتری در مقایسه با روشهای دیگر تبدیل گفتار دارد.

همچنین ملاحظه می‌شود که کیفیت بازسازی روش جدید در کلاس کیفیت CQ[1] (یعنی کیفیت مخابراتی) قرار دارد. نتیجه دیگر اینکه عملکرد روشهای تبدیل گفتار در تبدیل گفتاری زن به زن بهتر از حالات دیگر (مرد به مرد و مرد به زن) عمل می‌کند. در تمامی آزمونهای کیفی، کلیه شنوندگان اذعان داشته‌اند که در تبدیل گفتاری مرد به زن، روش جدید موفق بوده است. بدین معنا که شنوندگان، صدای بازسازی شده را به عنوان صدای زن قبول کرده‌اند. بهایی که برای وصول به این سطح از کیفیت پرداخته ایم، هزینه محاسباتی بیشتر، داده‌های آموزشی بیشتر و استفاده از فرایند پیچیده و خسته کنندهٔ برچسب زنی واجی داده‌های آموزشی بوده است. البته توجه می‌شود که فرایند ضبط داده‌های آموزشی و برچسب زنی واجی آنها، صرفاً یک بار (برای تولید صدای هر گوینده) انجام می‌شود.

۵- جمع‌بندی و نتیجه‌گیری

در طی این تحقیق ضمن بررسی اجمالی سه روش تبدیل گفتار (یا تبدیل گوینده)، روشی جدید، ارائه و پیاده

1. Communication Quality

سازی شده است. روش جدید مبتنی بر مدلسازی آماری GMM است. در روشهای کلاسیک، از مدل GMM برای مدلسازی کل فضای ویژگیهای هر گوینده استفاده می‌شود. تعداد مدلهای GMM در روش جدید سه برابر تعداد واجهای هر زبان است.

در مدلسازی جدید برای زبان فارسی از ۹۰ مدل GMM برای هر گوینده استفاده شد. همچنین تقطیع[1] هر واج به حالت، توسط الگوریتم ویتربی انجام شده است. در مرحله متناظر کردن خوشه‌های هر حالت، قبل از اعمال الگوریتم DTW از تبدیل LMR برای انطباق بیشتر پارامترهای دوحالت متناظر از دو گوینده استفاده شده است. برای تحلیل و سنتز پارامتری گفتار از مدل سینوسی با تهمیدات مناسبی استفاده شده، به قسمی که کیفیت گفتار بازسازی شده کاملاً طبیعی و غیر قابل تمیز از گفتار اصلی است.

نتایج پیاده‌سازی نشان می‌دهد که عملکرد روش جدید به مراتب بهتر از روشهای مطرح فعلی است. بهایی که در روش جدید باید بپردازیم، افزایش بیشتر داده‌های آموزشی و هزینه محاسباتی بالاتر است.

٦- منابع

[1] K. Shikano, K. Lee, R. Reddy, “Speaker Adaptation Through Vector Quantization”; *IEEE Proc. on ICASSP*; 1986; pp. 2643-2626.

[2] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, “Voice Conversion Through Vector Quantization”; *IEEE Proc. On ICASSP*; 1988; pp. 655-658.

[3] M. Abe, K. Shikano, H. Kuwabara, “Cross-Language Voice Conversion”; I*EEE Proc. on ICASSP*; 1990; pp 345-348.

[4] H. Mizuno, M. Abe, “Voice Conversion Based on Piecewise Linear Conversion Rules of Formant Frequency and Spectrum Tilt”; *IEEE Proc. on ICASSP*; 1994; pp. 1469-472.

[5] H. Valbret, E. Moulines, J.P. Tubach, “Voice Transformation Using PSOLA Techniques”; *IEEE Proc. on ICASSP*; 1992; pp. 1145-1148.

[6] W. Verhelst, J. Mertens, “Voice Conversion Using Partitions of Spectral Feature Space”; *IEEE Proc. on ICASSP*; 1996; pp.365-368.

[7] N.Bi. Y. Qi, “Application of Speech Conversion, to A laryngeal Speech Enhancement”; I*EEE Trans. on Speech and Audio Proc*. Vol. 5, No. 2; 1997.

[8] E. Moulines, Y. Sagisaka, “Voice Conversion, State of the Art and Perspectives”; *Speech Communication*., Vol. 16,No. 2; 1995; pp. 125-126.

[9] Y. Stylianou, O. Cappe, E. Moulines, “Continuous Probabilistic Transform for Voice Conversion”, *Speech Communication.*, Vol. 24, No. 2; 1998; pp. 192-200.

[10] D. A. Reynolds, R.C. Rose, ”Robust text independent Speaker identification using Gaussian mixture Speaker models”; I*EEE Trans. On Speech, Audio Processing*, Vol. 3; 1995; pp. 72-83.

[11] L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice Hall; 1993.

[12] R.J. McAulay, T.F. Quatieri, “Speech Analysis/Synthesis Based on a Sinusoidal Representation”; I*EEE Trans. on ASSP*, Vol. ASSP –34; 1986; pp. 744-754.

[13] E.B. Geroge, M.J.T. Smith, “Speech Analysis, Synthesis and Modification Using an Analysis by Synthesis Overlap add Sinusoidal Model”, *IEEE Trans. on Speech and Audio Processing*. Vol. 5, No. 5; 1997; pp. 389-406.

1. Segmentation

[۱۴] صیادیان، ابوالقاسم؛ بدیع، کامبیز؛ " ارائه الگوریتم دقیق و مقاوم، استفاده از نقاط با حداکثر اطلاعات MIP، برای تشخیص ابتدا و انتهای دستورات گفتاری"؛ نشریه علمی پژوهشی امیرکبیر؛ شماره ۱-۵۸؛ بهار ۱۳۸۳؛ صفحات ۳۲۰-۳۳۷.

[15] P. Veperk, M.S. Scordilis, "Analysis, Enhancement and Evaluation of Five Pitch Determination Techniques"; *Speech Communication*, Vol. 37; 2002; pp.249-270.