علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد۲، شماره ۱و۳ (الف)، بهار و پائیز ۱۳۸۳
صفحات ۱۹-۳۰
مقاله عادی

تولید پارامترهای سنتز گفتار فارسی با استفاده از مدلهای مخفی مارکف و درخت تصمیم گیری

محمد مهدی همایون پور　　سید مصطفی موسوی

آزمایشگاه سیستم های هوشمند صوتی-گفتاری
دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

مدل مخفی مارکوف(HMM) یکی از مدلهای آماری مناسب برای مدل کردن دنباله پارامترهای گفتار می باشد که استفاده از آن در سیستم تبدیل متن به گفتار، موفقیت آمیز بوده است. در این مقاله، برای پیاده سازی سیستم سنتز، از مدل های مخفی مارکوف برای مدل کردن پارامترهای مربوط به واحد های گفتاری استفاده شده است. برای تبدیل ضرائب کپسترال به سیگنال صحبت، از فیلتر MLSA استفاده نموده ایم. برای استخراج فرکانس گام، روش اتوکورلیشن اصلاح شده مورد استفاده قرار گرفته است. برای تولید پارامترهای سنتز گفتار توسط HMM ها، از الگوریتمی استفاده نموده ایم که در آن ، برای در نظر گرفتن اطلاعات بافت، علاوه بر ویژگی های ضرائب کپستروم و فرکانس گام، مشتق اول و دوم آنها نیز، مورد استفاده قرار گرفته است. برای بدست آوردن مدل طول زمانی واجها، مشاهدات موجود از هر ترایفون را در پایگاه داده، طبق الگوریتم ویتربی با مدل HMM آن مقایسه نموده و دنباله حالات طی شده را بدست آورده و با میانگین گیری از تعداد دفعات حضور در هر حالت مدل HMM ترایفون، متوسط طول زمانی حضور در هر حالت را برای هر ترایفون بدست آورده ایم. زمانهای میانگین حاصل، مدلهای طول زمانی برای هر ترایفون را تشکیل می دهند. در هنگام سنتز با توجه به مدل طول هر حالت از مدل HMM هر ترایفون، پارامترهای هر کدام از حالتهای HMM، شامل بردار میانگین و بردار واریانس آن حالت تکرار و با استفاده از این پارامترها، دنباله ضرائب کپسترال و گام مورد نیاز برای سنتز گفتار بدست آمده و توسط فیلتر MLSA به گفتار تبدیل شده اند. برای در نظر گرفتن تاثیر پارامترهای مختلف بر نحوه تلفظ آواها، از درخت های تصمیم گیری CART، استفاده شده است. این درخت ها نقش تولید گام و طول زمانی واج ها را بر عهده دارند. برای تولید اتوماتیک فرکانس گام از روش مطرح شده توسط فوجی ساکی استفاده نموده ایم . در این روش ، برای مدل کردن شکل کلی سیگنال گام، یک جزء عمومی و برای مدل کردن تاکید ها، تعدادی اجزای محلی در نظر گرفته می شود. برای ارزیابی سیستم از تست MOS و DRT استفاده شده است. امتیازات بدست آمده برای تست MOS در مورد سنتز با استفاده از مدلهای ترایفون و بدون استفاده از درخت تصمیم گیری برای تعیین طول زمانی ترایفون و گام، برای پارامترهای های قابل فهم بودن، طبیعی بودن و خوشایند بودن برای جملات آموزشی به ترتیب ۳/۸، ۳/۹ و ۳/۵، می باشد. در مورد سنتز با استفاده از درختهای تصمیم گیری برای مدل طول واجها و گام، امتیازات بدست آمده برای جملات آموزشی(یعنی جملاتی که در دادگان مورد استفاده، موجود بوده اند) به ازاء پارامترهای فوق به ترتیب ۴/۲، ۴/۴ و ۴/۱، و برای جملات آزمایشی(جملات خارج از دادگان مورد استفاده)، به ترتیب ۴/۳، ۴/۲ و ۳/۴ می باشد. بار دیگر همین تست را با استفاده از روش فوجی ساکی برای تخمین گام تکرار نموده ایم. در این حالت درخت تصمیم گیری برای تعیین طول زمانی ترایفونها استفاده شده است و برای تخمین گام مورد استفاده قرار نگرفته است که نتیجه آن برای جملات آموزشی به ترتیب ۴/۶، ۴/۳ و ۴/۵، و برای جملات آزمایشی، به ترتیب ۴/۵، ۴/۰ و ۴/۳ بدست آمده است. نتیجه بدست آمده برای تست DRT نیز برای حالتی که از درخت های تصمیم گیری برای تعیین طول زمانی ترایفون و گام استفاده نموده ایم برابر ۸۸٪ می باشد. نتایج تست ها نشاندهنده مناسب بودن روش های بکار رفته در این مقاله می باشد.

کلمات کلیدی : تبدیل متن به گفتار، سنتز گفتار، فیلتر سنتز، درخت تصمیم گیری، مدل مخفی مارکوف، ضرایب کپسترال

۱- مقدمه

تبدیل متن به گفتار در دو مرحله انجام می گیرد. در مرحله اول رشته متنی به دنباله ای از واحدهای صوتی تبدیل می گردد و در مرحله دوم این واحدهای صوتی به رشته گفتاری نظیر آن تبدیل می شوند. مرحله اول با استفاده از پردازش زبان طبیعی و با انجام تحلیل های لغوی و صرفی، نحوی و معنایی در زبان مورد نظر و بسیاری امور دیگر مانند تشخیص انتهای جمله، تبدیل کوته نوشت ها و اعداد بصورت متن و مانند آن انجام می گیرد. در مرحله دوم هدف اینست که رشته های واجی تولید شده در مرحله اول را به گفتار معادل آن تبدیل نماییم. برای انجام اینکار روشهای مختلفی وجود دارد که همانطور که ذکر شد مهمترین این روش ها عبارتند از سنتز کننده های مفصلی، سنتز کننده های فرمنتی، سنتز کننده های پیوندی و سایر سنتزکننده ها. در گذشته سنتز کننده های فرمنتی بیشتر مورد

استفاده قرار می گرفتند، در حالیکه امروزه سنتز کننده های پیوندی موارد استفاده فراوانی یافته اند. سنتز کننده های مفصلی هنوز پیچیدگی های خاص خود را دارند، اما در آینده انتظار می رود که به عنوان مهمترین سنتز کننده، مطرح شوند. مدل مخفی مارکوف یکی از مدلهای آماری مناسب برای مدل کردن دنباله پارامترهای گفتار می باشد که استفاده از آن در سیستم تبدیل متن به گفتار، موفقیت آمیز بوده است[۱-۱۷]. در این مقاله، برای پیاده سازی سیستم سنتز، از HMM ها برای مدل کردن پارامترهای مربوط به واحد های گفتاری استفاده شده است. یکی از مهمترین بخش های یک سیستم تبدیل متن به گفتار، روش مورد استفاده برای آنالیز و سنتز گفتار می باشد، چرا که مشخصات مدل طیف از قبیل پایداری فیلتر سنتز، کیفیت سنتز و حتی ساختار سیستم را تحت تاثیر قرار می دهد. ضرائب کپسترال مبتنی بر معیار مل که وضوح فرکانسی طیف تولید شده توسط آن الهام گرفته از سیستم شنیداری انسان است، برای سنتز گفتار بسیار مناسب می باشد. در روش مورد استفاده در این مقاله، با مینیمم کردن معیار بایاس نشده لگاریتم طیف با توجه به ضرائب کپسترال، این ضرائب استخراج می گردند[۱۸-۱۹]. برای تبدیل ضرائب کپسترال به سیگنال صحبت، از فیلتر MLSA استفاده نموده ایم[۱۸-۱۹]. این فیلتر ضرائب کپسترال و موج تحریک را به عنوان ورودی دریافت می کند و در خروجی سیگنال صحبت را تولید می نماید. برای استخراج فرکانس گام از روش اتوکورلیشن اصلاح شده استفاده نموده ایم [۲۰] در الگوریتم استفاده شده، تاثیر پنجره گذاری و نمونه برداری در محاسبه فرکانس گام، در نظر گرفته شده و خطای حاصل از آنها مینیمم شده است. برای تولید پارامترهای سنتز گفتار توسط HMM ها، از الگوریتمی استفاده نموده ایم که در آن، برای در نظر گرفتن اطلاعات بافت، علاوه بر ویژگی های ضرائب کپستروم و فرکانس گام، مشتق اول و دوم آنها نیز، مورد استفاده قرار گرفته است[۸-۹]. گفتار استفاده شده بعنوان دادگان، حدود ۴۰۰ جمله از یک گوینده ۲۹ ساله مرد با مدرک تحصیلی لیسانس می باشد که از دادگان فارسی FARSDAT بزرگ، استخراج شده است و برای استفاده در این تحقیق، تمامی جملات را بر چسب گذاری واج نموده ایم. واحد گفتاری مورد استفاده ترایفون می باشد که در این مقاله، منظور از آن واجی است که در نامگذاری آن واج ماقبل و مابعد آن نیز لحاظ شده است. با استفاده از ضرائب کپسترال و فرکانس گام، مدلهای مخفی مارکوف آموزش داده شده است. برای بدست آوردن مدل طول زمانی واجها، مشاهدات موجود از هر ترایفون را در پایگاه داده، طبق الگوریتم ویتربی با مدل HMM آن مقایسه نموده و دنباله حالات طی شده را بدست آورده ایم و بدین ترتیب مشخص شده است که در هر تکرار ترایفون چند بار هر یک از حالتهای مدل HMM آن ترایفون، بطور متوالی تکرار شده است. با میانگین گیری از تعداد دفعات حضور در هر حالت مدل HMM ترایفون، متوسط طول زمانی حضور در هر حالت برای هر ترایفون بدست می آید. زمانهای میانگین حاصل، مدلهای طول زمانی برای هر ترایفون را تشکیل می دهند. در هنگام سنتز با توجه به مدل طول هر حالت از مدل HMM هر ترایفون، پارامترهای هر کدام از حالتهای HMM، شامل بردار میانگین و بردار واریانس آن حالت را تکرار می کنیم و با استفاده از این پارامترها، دنباله ضرائب کپسترال و گام مورد نیاز برای سنتز گفتار را بدست آورده و توسط فیلتر MLSA به گفتار تبدیل می نماییم.

۲- دادگان گفتاری

دادگان گفتاری مورد استفاده، دادگان فارسی FARSDAT بزرگ می باشد که شامل گفتار ۱۰۰ گوینده مختلف در شرایط سنی، تحصیلی و اجتماعی مختلف است. گوینده انتخاب شده، یک فرد ۲۹ ساله با مدرک تحصیلی لیسانس می باشد. در این دادگان حدود نیم ساعت از صدای این شخص موجود می باشد که چیزی در حدود ۴۰۰ جمله است. بر چسب گذاری این دادگان، بر چسب کلمه است، بنابراین کل این داده ها را مجددا برچسب گذاری واج نموده ایم. جملات موجود در دادگان فارس دات با فرکانس ۲۲۰۵۰ هرتز نمونه برداری شده اند. برای استفاده در این مقاله، فرکانس نمونه برداری جملات را به ۱۱۰۲۵ هرتز کاهش دادیم. برای استخراج ویژگی ها و گام از فریمهایی به طول ۲۳ میلی ثانیه با ۱۰ میلی ثانیه شیفت، استفاده شده است. برای پنجره گذاری از پنجره هنینگ استفاده نموده ایم. طول هر فریم برحسب تعداد نمونه ها، برابر ۲۵۶ نمونه است.

۳- استخراج گام

محاسبه فرکانس گام در پردازش گفتار، ومخصوصا در سیستم های سنتز گفتار از اهمیت ویژه ای برخوردار است. کیفیت صدای سنتز شده رابطه خیلی نزدیکی با دقت گام استخراج شده دارد. الگوریتم های استخراج به دو دسته الگوریتم های حوزه زمان و الگوریتم های حوزه فرکانس تقسیم می شوند. روش های مختلفی برای استخراج فرکانس گام پیشنهاد شده است. در این مقاله برای استخراج گام از روش اتوکورلیشن استفاده نموده ایم. در روش استفاده شده با در نظر گرفتن تاثیر پنجره گذاری و نمونه برداری سعی شده است که خطای تخمین گام حداقل گردد[۱۳].

۴- آنالیز و سنتز بر اساس ضرائب کپسترال

روش مورد استفاده برای آنالیز و سنتز گفتار یکی از مهمترین بخش های یک سنتز کننده گفتار می باشد. چراکه مشخصات مدل طیف از قبیل پایداری فیلتر سنتز، کیفیت سنتز و حتی ساختار سیستم را تحت تاثیر قرار می دهند. آنالیز و سنتز کپسترال، برای تخمین طیف در سیستم سنتز بر اساس HMM استفاده شده است. وضوح فرکانسی طیف تولید شده بروش آنالیز کپسترال مبتنی بر معیار مل که الهام گرفته از سیستم شنیداری انسان است، برای سنتز گفتار بسیار مناسب می باشد. برای بدست آوردن ضرائب کپستروم روش های مختلفی وجود دارد. برخی از این روشها ضرائب کپستروم را توسط ضرائب پیشگوئی خطی LPC بدست می آورند و بعضی دیگر ضرائب کپستروم را با استفاده از تبدیل فوریه و مبتنی بر معیار مل بدست می آورند. در الگوریتم مورد استفاده در این مقاله برای استخراج این ویژگی ها، با مینیمم کردن معیار بایاس نشده لگاریتم طیف[۱] با توجه به ضرائب کپستروم، این ضرائب را استخراج می نماییم[۱۸-۱۹].

۵- فیلتر MLSA برای سنتز

در آنالیز کپسترال بر اساس معیار مل، تابع انتقال مجرای گفتار H(z) توسط M ضریب کپستروم $c=[c(0),c(1),...,c(M)]^T$ توسط رابطه ذیل مدل می شود:

$$H(Z)=\exp\sum_{m=0}^{m}c(m)\tilde{z}^{-m} \qquad (۱)$$

که $\tilde{z}=[1,\tilde{z}^{-1},\tilde{z}^{-2},...,\tilde{z}^{-M}]^T$ توسط تابع زیر بیان می گردد:

$$\tilde{z}^{-1}=\frac{z^{-1}-a}{1-az^{-1}} \qquad |a|<1 \qquad (۲)$$

H(z) را به شکل زیر بازنویسی می نماییم:

$$H(z)=K.D(z) \qquad (۳)$$

که

$$K=\exp\sum_{m=0}^{M}(-a)^m c(m) \qquad (۴)$$

$$D(z)=\exp\sum_{m=1}^{M}c_1(m)\tilde{z}^{-m} \quad (۵)$$

و

$$a=[1,(-a),(-a)^2,...,(-a)^M]^T \quad (۶)$$

$$c_1=[c_1(0),c_1(1),...,c_1(M)]^T \quad (۷)$$

رابطه بین c و c_1 به شکل زیر بیان می شود:

$$c_1(m)=\begin{cases}c(0)-a^T c & m=0\\ c(m) & 1\le m\le M\end{cases} \quad (۸)$$

برای اینکه گفتار را توسط ضرائب کپستروم تولید نماییم، باید از فیلتر تابع انتقال D(z) استفاده نماییم. هر چند که D(z) یک تابع کسری نمی باشد، می توان آنرا با استفاده از فیلتر MLSA[۲] با دقت خوبی تقریب زد. تابع $\exp(w)$ با استفاده از یک تابع کسری به شکل زیر تقریب زده می شود:

$$\exp(w)=R_L(w)=\frac{1+\sum_{l=1}^{L}A_{L,l}w^l}{1+\sum_{l=1}^{L}A_{L,l}(-w)^l} \quad (۹)$$

با استفاده از این رابطه می توان D(z) را به شکل زیر تقریب زد:

$$D(z)=\exp F(z)\simeq R_L(F(z)) \quad (۱۰)$$

که

$$F(z)=\tilde{z}^T c_1=\sum_{m=0}^{M}c_1(m)\tilde{z}^{-m} \quad (۱۱)$$

که $A_{L,l}(l=1,2,...,L)$ دارای مقادیر ثابت می باشد و مقادیر $c_1(m)$ متغییر هستند. برای اینکه پاسخ ضربه F(z) در زمان صفر برابر صفر شود و حلقه بدون تاخیر در F(z) نداشته باشیم رابطه فوق را به شکل زیر بازنویسی می کنیم:

$$F(z)=\tilde{z}^T c_1=\tilde{z}^T AA^{-1}c_1=\Phi b=\sum_{m=1}^{M}b(m)\Phi_m(z) \quad (۱۲)$$

که

$$A=\begin{bmatrix}1 & a & 0 & \dots & 0\\ 0 & 1 & a & \dots & 0\\ 0 & 0 & 1 & \dots & 0\\ \dots & \dots & \dots & & \dots\\ 0 & \dots & \dots & 0 & 1\end{bmatrix} \quad (۱۳)$$

$$A^{-1}=\begin{bmatrix}1 & (-a) & (-a)^2 & \dots & (-a)^M\\ 0 & 1 & (-a) & \dots. & \dots.\\ 0 & 0 & 1 & & (-a)^2\\ & \dots & \dots & \dots & (-a)\\ 0 & \dots & \dots & 0 & 1\end{bmatrix} \quad (۱۴)$$

بردار Φ به شکل زیر بیان می شود:

$$\Phi=A^T\tilde{z}=[1,\Phi_1(z),\Phi_2(z),....,\Phi_M(z)]^T \quad (۱۵)$$

که

$$\Phi_m(z)=\frac{(1-a^2)z^{-1}}{1-az^{-1}}\tilde{z}^{-(m-1)} \qquad m\ge 1 \quad (۱۶)$$

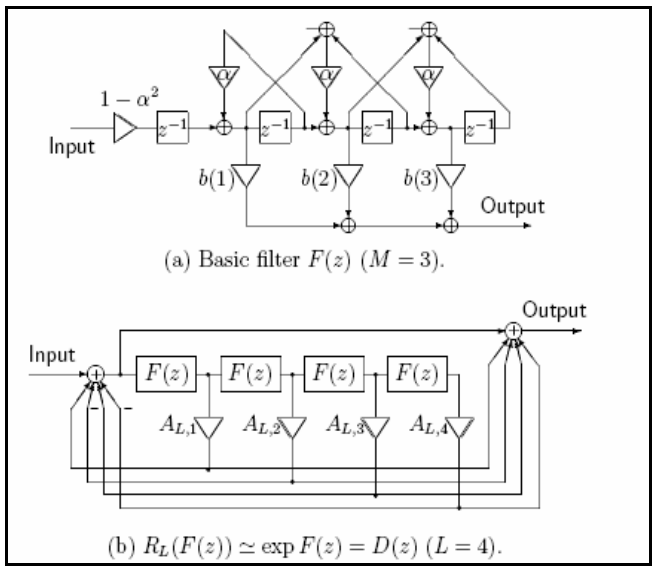
ضرائب b طبق روابط زیر از c_1 بدست می آیند:

$$b=A^T c_1=[0,b(1),b(2),...,b(M)]^T \quad (۱۷)$$

یعنی:

$$b(m)=\begin{cases}c_1(M) & m=M\\ c_1(m)-ab(m+1) & 0\le m\le M-1\end{cases} \quad (۱۸)$$

بلاک دیاگرام فیلتر MLSA بصورت $R_L(F(z))\simeq D(z)$ در شکل ۱ مشاهده می شود.



شکل ۱- فیلتر D(z)

۶- استخراج پارامترهای سنتز گفتار از HMM چند مخلوطی با در نظر گرفتن مقادیر دلتا و دلتا دلتا

فرض کنید که برای یک HMM با مخلوط های گوسی پیوسته، همانند الگوریتم ویتربی مقادیر مخلوط ها را به عنوان زیر حالت های یک حالت، در نظر بگیریم و سعی کنیم مقدار $p[Q,O|l]$ را با توجه به دنباله بردار پارامترهای گفتار یعنی $O=[O_1',O_2',...,O_T']'$ و دنباله زیر حالت یعنی $Q=\{(q_1,i_1),(q_2,i_2),...(q_T,i_T)\}$ به گونه ای که (q,i) مخلوط i ام حالت q را نشان می دهد، ماکزیمم نماییم. علاوه بر اینها مقادیر طول حالت ها را نیز در محاسبات دخالت می دهیم. با در نظر گرفتن:

$$m=\left[m'_{q_1,i_1},m'_{q_2,i_2},...,m_{q_T,i_T}\right]' \quad (۱۹)$$

$$U=diag\left[U_{q_1,i_1},U_{q_2,i_2},...,U_{q_T,i_T}\right] \quad (۲۰)$$

مقدار $p[Q,O|l]$ به شکل زیر محاسبه می گردد[10]:

$$\log P[Q,O|l]=a\sum_{k=1}^{K}\log p_k(d_k)+\sum_{t=1}^{T}\log c_{q_t,i_t}-\frac{1}{2}e(c)-\frac{1}{2}\log|u|-\frac{3MT}{2}\log 2p \quad (۲۱)$$

که

$$e(c)=(Wc-m)'U^{-1}(Wc-m) \quad (۲۲)$$

و

$$W = [w_1, w_2, ..., w_T]' \quad (۲۳)$$

$$W_t = [w_t^{(0)}, w_t^{(1)}, w_t^{(2)}] \quad (۲۴)$$

$$w_t^{(n)} = [\underbrace{0_{M\times M}}_{1st}, ..., 0_{M\times M}, \underbrace{w^{(n)}(-L^{(n)})I_{M\times M}}_{(t-L^{(n)})-th}, \quad (۲۵)$$
$$..., \underbrace{w^{(n)}(0)_{M\times M}}_{t-th}, ..., \underbrace{w^{(n)}(L^{(n)})I_{M\times M}}_{(t+L^{(n)})-th},$$
$$\underbrace{0_{M\times M}, ..., 0_{M\times M}}_{T-th}] \qquad n = 0,1,2,$$

که $0_{M\times M}$ و $I_{M\times M}$ به ترتیب بیانگر ماتریس صفر و همانی $M \times M$ می باشند. فرض می کنیم $c_t = 0_M, t<1, t>T$ که 0_M بردار صفر $M\times 1$ می باشد. حال با قراردادن $\partial \log P[Q,O|l]/\partial c = 0_{TM}$ به مجموعه معادلات زیر می رسیم :

$$\mathrm{Rc} = \mathrm{r} \quad (۲۶)$$

که

$$R = W'U^{-1}W \quad (۲۷)$$

$$r = W'U^{-1}m \quad (۲۸)$$

که حل مستقیم این معادله از درجه $O(T^3M^3)$ است. برای حل این معادله از ایده فیلتر های RLS استفاده می کنیم. برای اینکار (q_t, i_t) را با $(\hat{q}_t, \hat{i}_t)$ جایگزین می نمائیم. یعنی اگر مخلوط i ام از حالت q_t را با یک مقدار جدید جایگزین نماییم. نهایتاً خواهیم داشت:

$$\hat{\mathrm{R}}\mathrm{c} = \hat{\mathrm{r}} \quad (۲۹)$$

که

$$\hat{R} = R + w_t D w_t' \quad (۳۰)$$

$$\hat{r} = r + w_t d \quad (۳۱)$$

$$D = U^{-1}_{\hat{q}_t,\hat{i}_t} - U^{-1}_{q_t,i_t} \quad (۳۲)$$

$$d = U^{-1}_{\hat{q}_t,\hat{i}_t} m_{\hat{q}_t,\hat{i}_t} - U^{-1}_{q_t,i_t} m_{q_t,i_t} \quad (۳۳)$$

مشاهده می شود که رابطه بین R و $\hat{\mathrm{R}}$ همانند رابطه ایست که در ایده فیلترهای RLS مطرح است. بنابراین طبق نظریه الگوریتم RLS، می توان یک راه حل سریع برای حل معادله فوق ارائه نمود. الگوریتم کار را در جدول ۱ مشاهده می کنید. پیچیدگی محاسباتی این الگوریتم از درجه $O(T^3M^3)$ خواهد بود و به فرض اینکه ماتریس U قطری باشد این مقدار به $O(T^3M)$ کاهش خواهد یافت. اگر فرض کنیم که میانگین و واریانس در هر فریم t تحت تاثیر S فریم همسایه قرار گیرد، پیچیدگی محاسباتی الگوریتم به $O(S^2M^3)$ و در حالت قطری بودن ماتریس کواریانس به $O(S^2M)$ کاهش می یابد. تجربه نشان داده است که مقدار $S = 30$ مقدار مناسبی است.

با استفاده از الگوریتم بازگشتی می توانیم دنباله زیر حالاتی را که c بهینه را با توجه به ماکزیمم نمودن $\log P[Q,O|l]$ نسبت به c نتیجه می دهـــد، پیدا کنیم. برای هر دنباله از زیر حالات می توانیم به جای حل مستقیم رابطه (۲۹) از

جدول ۱- خلاصه الگوریتم استخراج پارامترهای سنتز گفتار از HMM چند مخلوطی با در نظر گرفتن مقادیر دلتا و دلتا دلتا

- در هر مرحله مقادیر $\hat{P}$ ،$\hat{c}$ و $\hat{e}$ را با مقادیر P ،c و e مرحله قبل جایگزین
- می کنیم و عملیات زیر را انجام می دهیم:

$$p = Pw_t \quad (T.1)$$

$$v = w_t'p \quad (T.2)$$

$$k = p\{I_{3M} + (U^{-1}_{\hat{q}_t,\hat{i}_t} - U^{-1}_{q_t,i_t})v\}^{-1} \quad (T.3)$$

$$\hat{c} = c + k\{U^{-1}_{\hat{q}_t,\hat{i}_t}(m_{\hat{q}_t,\hat{i}_t} - w_t'c) - U^{-1}_{q_t,i_t}(m_{q_t,i_{tt}} - w_t'c)\} \quad (T.4)$$

$$\hat{e} = e + (m_{\hat{q}_t,\hat{i}_t} - w_t'\hat{c})U^{-1}_{\hat{q}_t,\hat{i}_t}(m_{\hat{q}_t,\hat{i}_t} - w_t'c) - (m_{q_t,i_t} - w_t'\hat{c})U^{-1}_{q_t,i_t}(m_{q_t,i_t} - w_t'c) \quad (T.5)$$

$$\hat{P} = P - k(U^{-1}_{\hat{q}_t,\hat{i}_t} - U^{-1}_{q_t,i_t})p \quad (T.6)$$

رابطه بازگشتی استفاده کنیم. بنابراین پیچیدگی کل محاسبات بسیار کاهش می یابد. برای بدست آوردن مجموعه زیر حالات بهینه می توان از الگوریتم زیر استفاده نمود:

۱) مقدار دهی اولیه

الف- دنباله حالات اولیه Q را مشخص می کنیم.

ب- برای دنباله زیرحالت اولیه مقادیر c، e و P را محاسبه می کنیم.

۲) تکرار مراحل زیر :

الف- برای t=1,2, . . . ,T

(الف-۱) مقادیر (T.1) و (T.2) را محاسبه کن.

(الف-۲) برای هر زیر حالت از فریم t مقادیر (T.3) تا (T.5) را محاسبه کن و $\log P[Q,O|l]$ را با استفاده از رابطه (۲۱) بدست آور.

(الف-۳) بهترین زیر حالتی را که بتواند $\log P[Q,O|l]$ را ماکزیمم کند بدست آور.

ب- بهترین فریمی را که بتواند با جابجا کردن زیر حالات مقدار $\log P[Q,O|l]$ را ماکزیمم کند بدست آور.

ج- اگر بهترین فریم نتواند مقدار $\log P[Q,O|l]$ را ماکزیمم کند، تکرار را ادامه نده.

د- با محاسبه (T.1) تا (T.6) مقدار زیر حالت بهترین فریم را جایگزین کن و مقدار $\hat{c}$، $\hat{e}$و $\hat{P}$ را بدست آور

ه- به ۲-الف برگرد.

برای اینکه دنباله بهینه را با انجام تکرارهای کم، پیدا کنیم، بهتر است که دنباله حالات اولیه، نزدیک به دنباله حالات بهینه باشد. برای انجام اینکار توسط الگوریتم ویتربی دنباله حالات $q = \{q_1, q_2, ..., q_T\}$ را که بتواند رابطه زیر را با توجه به q ماکزیمم نماید، بدست می آوریم.

$$\log P[q|l] = \sum_{k=1}^{K} \log p_{q_k}(d_{q_k}) \quad (۳۴)$$

دنباله زیر حالات $i = \{i_1, i_2, ..., i_T\}$ را باید به گونه ای تعیین کنیم که $\log c_{q_t,i_t} - \frac{1}{2}\log|U_{q_t,i_t}|$، با در نظر گرفتن i_t، ماکزیمم شود. زیر

حالت های مجازی را در نظر بگیرید که که میانگین و کوواریانس آنها به شکل زیر تعریف شده است:

$$\bar{m}_{q_t,i_t} = \left[m'^{(0)}_{q_t,i_t}, 0'_M, 0'_M\right]' \quad (۳۵)$$

$$\bar{U}^{-1}_{q_t,i_t} = \begin{bmatrix} \left(U^{(0)}_{q_t,i_t}\right)^{-1} & 0_{M\times M} & 0_{M\times M} \\ 0_{M\times M} & 0_{M\times M} & 0_{M\times M} \\ 0_{M\times M} & 0_{M\times M} & 0_{M\times M} \end{bmatrix} \quad (۳۶)$$

که $m^{(0)}_{q_t,i_t}$ و $U^{(0)}_{q_t,i_t}$ به ترتیب بردار میانگین $M\times 1$ و ماتریس کوواریانس $M\times M$، از بردار ویژگی c_t در مخلوط i ام از حالت q_t می باشند. با استفاده از روابط فوق می توانیم مقادیر اولیه c، P و e را به شکل زیر پیدا کنیم:

$$\bar{c} = \left[m'^{(0)}_{q_1,i_1}, m'^{(0)}_{q_2,i_2} \ldots, m'^{(0)}_{q_T,i_T}\right] \quad (۳۷)$$

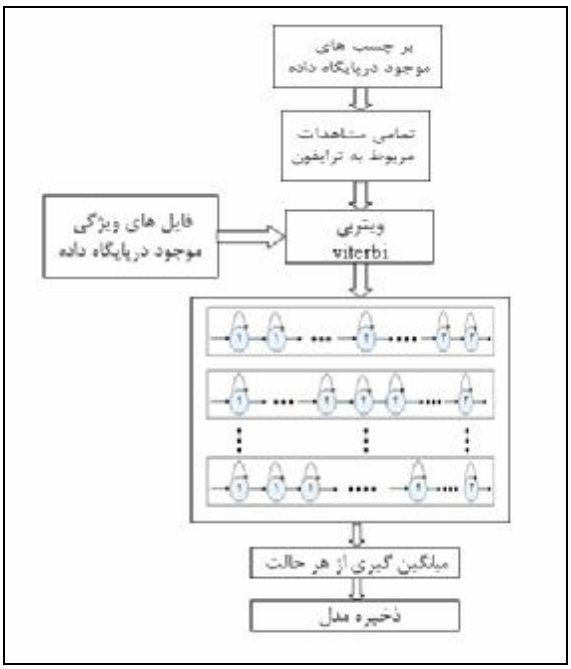
$$\bar{P} = diag\left[U^{(0)}_{q_1,i_1}, U^{(0)}_{q_2,i_2}, \ldots, U^{(0)}_{q_T,i_T}\right] \quad (۳۸)$$

و $\bar{e} = 0$. حال با قرار دادن مقادیر $m^{(0)}_{q_t,i_t}$ و $U^{(0)}_{q_t,i_t}$ بازاء زمانهای $t = 1,2,\ldots,T$ و T بار تکرار الگوریتم، مقادیر c، P و e را برای دنباله زیر حالت های اولیه پیدا می کنیم.

۷- پیاده سازی سنتز کننده با استفاده از مدلهای ترایفون

۷-۱ آموزش مدلهای ترایفون

منظور از ترایفون در این مقاله، واجی است که در نامگذاری آن واج ماقبل و ما بعد آن نیز لحاظ شده باشد. برای این منظور لازمست برای هر واج نام ترایفون آنرا در دادگان مشخص کنیم. برای این منظور، برای هر واج، واج قبل و بعد آنرا در نظر گرفته و برچسب ترایفون آنرا تشکیل می دهیم. سپس با استفاده از تکرارهای هر ترایفون، مدل HMM ویژگی و گام را برای هر ترایفون آموزش می دهیم. برای مقادیر گام و ویژگی، علاوه بر خود این مقادیر، مقادیر دلتا و دلتادلتای آنها نیز در نظر گرفته شده است. مدل طول ترایفون، علاوه بر طول زمانی هر ترایفون، باید شامل دنباله حالتهای پیموده شده برای HMM، نیز باشد. مراحل لازم برای بدست آوردن دنباله حالات و طول ترایفونها در شکل ۲ مشاهده می شود. برای آموزش مدل ابتدا با توجه به برچسب های موجود در پایگاه داده، تمامی تکرارهای ترایفون مورد نظر را جمع آوری می کنیم و سپس دنباله حالات را بازاء هر یک از این مشاهدات بدست می آوریم. حال با داشتن این دنباله حالات، می توان با استفاده از یک الگوریتم مناسب، مدلی برای طول ترایفون بدست آورد. در این مقاله برای بدست آوردن مدل های طول زمانی، از روش میانگین گیری استفاده شده است. برای پیدا نمودن مشاهدات، ابتدا و انتهای هر ترایفون را در فایل برچسب پیدا می کنیم. سپس با توجه به نام ترایفون، مدل HMM آنرا فراخوانی می کنیم. فرض کنید این مدل را با l نمایش دهیم. اگر یک دنباله مشاهده مانند $O = (o_1, o_2, \ldots, o_t)$ داشته باشیم، طبق الگوریتم ویتربی، O را به مدل l اعمال می کنیم و دنباله حالت بهینه $Q^* = (q_1, q_2, \ldots, q_t)$ بدست می آید. بدست آوردن دنباله حالات بهینه را برای کلیه تکرارهای ترایفون ها انجام می دهیم. بدین ترتیب با توجه به دنباله حالات بهینه بازاء هر تکرار از ترایفون، معلوم می شود که در هر تکرار ترایفون چند بار هر یک از حالتهای مدل HMM آن ترایفون بطور متوالی تکرار شده است. تعداد دفعات حضور در هر حالت مدل HMM ترایفون را برای کلیه تکرارهای آن تعیین و میانگین گیری می نماییم. بدین ترتیب اگر اینکار برای کلیه حالات مدل HMM ترایفون انجام شود، متوسط طول زمانی حضور در هر حالت برای هر ترایفون بدست می آید. زمانهای میانگین حاصل، مدلهای طول زمانی برای هر ترایفون را تشکیل می دهند.



شکل ۲- ساختار مورد نیاز برای تولید مدل طول واجها

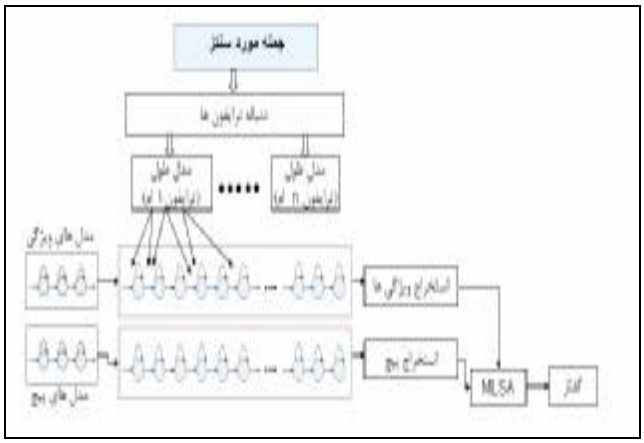
۷-۲ سنتز گفتار با استفاده از مدلهای ترایفون

در هنگام سنتز، با توجه به رشته واج ورودی برای هر واج با توجه به واج قبل و واج بعد از آن، ترایفون مربوطه را مشخص می کنیم. همانطور که در بخش قبل، در مورد آموزش مدل طول ترایفون ها ذکر شد، برای مدل HMM هر ترایفون، تعداد تکرار حالتهای آنرا مشخص نموده و ذخیره کرده ایم. بنابراین در ابتدا، مدل طول مربوط به ترایفون را فراخوانی می کنیم. فرض کنید تعداد تکرار های ذخیره شده برای حالتهای s_1، s_2 و s_3 از HMM به ترتیب، d_1، d_2 و d_3 باشد. حال مدلهای HMM ویژگی و گام را فراخوانی می کنیم. ابتدا d_1 مرتبه، حالت s_1 از HMM را تکرار می کنیم و برای هر تکرار، در فایل خروجی ابتدا بردار میانگین، همراه با مشتقات اول و دوم آن، و سپس بردار واریانس، همراه با مشتقات اول و دوم آن، ذخیره می گردد. برای گام نیز به همین ترتیب، حالت اول از HMM آنرا d_1 مرتبه تکرار می کنیم و در یک فایل خروجی دیگر قرار می دهیم. سپس این عمل را برای حالتهای s_2 و s_3 از HMM ها، به ترتیب d_2 و d_3 مرتبه انجام می دهیم. برای ترایفون بعدی، مراحل فوق را تکرار می نماییم. در نهایت یک دنباله از پارامترها شامل بردارهای میانگین و مقادیر واریانس برای ویژگی ها و گام بدست می آید. با اعمال الگوریتم بیان شده در فصل های قبل به هر کدام از این دنباله پارامترها، مقادیر نهایی ویژگی ها و گام، که باید به فیلتر سنتز اعمال نمود، بدست می آید. حال این مقادیر را مستقیما به فیلتر سنتز اعمال می کنیم و صدای سنتز شده بدست می آید. ساختار کامل سنتز کننده در شکل۳ مشاهده میشود.

۸- سنتز گفتار با استفاده از درخت های تصمیم گیری

از آنجا که علاوه بر خود واج و واجهای قبل و بعد، پارامترهای دیگری نیز بر نحوه تلفظ آواها تاثیر می گذارند ناگزیر هستیم که برای بالا بردن کیفیت سنتز کننده این پارامترها را برای تولید پارامترهای گفتار در نظر بگیریم. از جمله این پارامترها

می توان به تکیه دار بودن یا نبودن هجا، محل واج در هجا، محل هجا در کلمه، تعداد واجهای موجود در هجا و ... اشاره نمود. در اینجا، سعی نموده ایم همه پارامترهایی را که بر روی تلفظ آواها تاثیر می گذارند در نظر بگیریم. در جدول ۲ پارامترهای استفاده شده در این سنتز کننده، مشاهده میشوند. برای استخراج اتو ماتیک پارامترهای ذکر شده در جدول ۲، از ساختار شکل ۴ استفاده می نماییم. در این ساختار برای هر کدام از واحدهای گفتاری یک کلاس در نظر گرفته شده است. همانطور که در بحث مربوط به دادگان گفتیم، جمله های موجود در پایگاه داده طوری برچسب گذاری شده اند که انتهای هر کدام از واحدها با علامت مناسبی مشخص شده است. بنابراین هنگامیکه جمله ورودی دریافت شد، وارد کلاس جمله می شود و در این کلاس، عبارتهای[۳] موجود در آن تشخیص داده میشود و هر کدام از عبارتها را به کلاس عبارت ارسال می نماییم. علاوه براین اطلاعاتی از قبیل تعداد عبارتهای موجود در جمله ذخیره می گردد. در کلاس عبارت نیز به همین ترتیب کلمات موجود در هر عبارت با توجه به برچسب های کلمه مشخص می شوند و هر کدام به یک کلاس کلمه ارسال می گردند. در کلاس کلمه، باید هجا های موجود در هر کلمه را تعیین نماییم. برای تعیین هجاها، با توجه به این مسئله که هجاها در زبان فارسی به سه شکل CV، CVC وCVCC می باشند، به راحتی می توان مرز بین هجاهای موجود در هر کلمه را مشخص نمود. در این کلاس علاوه بر مرز هجا، باید تکیه دار بودن یا نبودن هجا را نیز تعیین نماییم، چرا که تکیه هجا نیز یکی از پارامترهای ذکر شده در جدول ۲ است که یکی از پارامترهای مهم مربوط به بافت است. از آنجا که محل قرار گرفتن تکیه در زبان فارسی هجا است، همانطور که ذکر شد، در هنگام بر چسب گذاری اگر هجایی تکیه دار باشد، علامتی روی واکه مربوط به آن قرار می گیرد. در نهایت در کلاس هجا، واجهای موجود در آنرا تشخیص می دهیم و هر واج را به یک کلاس واج ارسال می نماییم. با توجه به داده هایی که در هر کدام از کلاس های فوق جمع آوری شده است، به راحتی می توان برای هر واج، پارامترهای موجود در جدول ۲ را بدست آورد. به عنوان مثال برای محاسبه پارامتر B3، که تعداد واجهای موجود در هجا را مشخص می کند، کافی است برای هر واج، به کلاس بالاتر از آن یعنی کلاس هجا رجوع نماییم و تعداد واجهای موجود در آنرا که از قبل ذخیره نموده ایم، بدست آوریم.



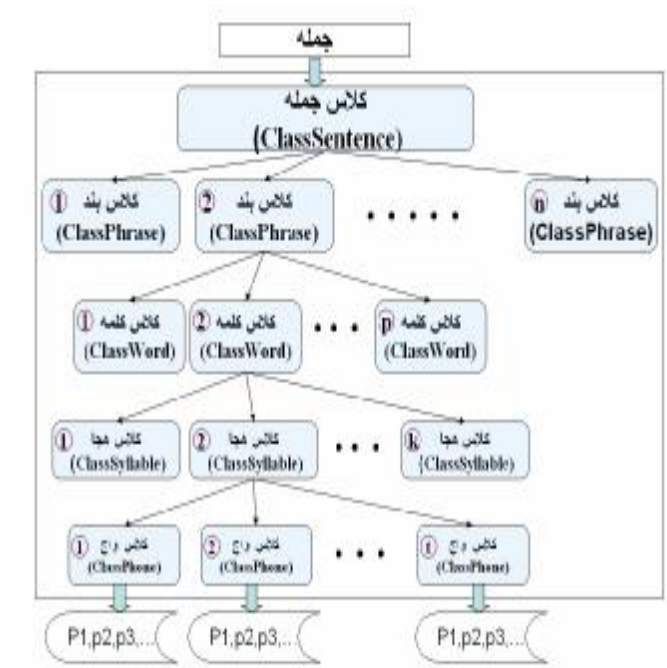
شکل ۳- ساختار استفاده شده برای سنتز گفتار با استفاده از مدل های ترایفون

جدول ۲- پارامتر های استفاده شده برای در نظر گرفتن موارد مربوط به بافت در سنتز کننده

پارامتر	توضیحات
P1	واج قبل از واج قبلی
P2	واج قبلی
P3	واج فعلی
P4	واج بعدی
P5	واج بعد از واج بعدی
P6	محل واج فعلی در هجای فعلی (پیشرو)
P7	محل واج فعلی در هجای فعلی (پسرو)
A1	آیا هجای قبلی تکیه دارد یا خیر (۱:بله ۰:خیر)
A2	تعداد واجها در هجای قبلی
B1	آیا هجای فعلی تکیه دارد یا خیر (۱:بله ۰:خیر)
B2	تعداد واجها در هجای فعلی
B3	محل هجای فعلی در کلمه فعلی (پیشرو)
B4	محل هجای فعلی در کلمه فعلی (پسرو)
B5	محل هجای فعلی در عبارت فعلی (پیشرو)
B6	محل هجای فعلی در عبارت فعلی (پسرو)
B7	تعداد هجاهای تکیه دار قبل از هجای فعلی در عبارت فعلی
B8	تعداد هجاهای تکیه دار بعد از هجای فعلی در عبارت فعلی
B9	تعداد هجاها از هجای تکیه دار قبلی تا هجای فعلی
B10	تعداد هجاها از هجای فعلی تا هجای تکیه دار قبلی
B11	واکه موجود در هجای فعلی
C1	آیا هجای بعدی تکیه دارد یا خیر (۰:بله ۱:خیر)
C2	تعداد واجها در هجای بعدی
D1	تعدادهجاها در کلمه قبلی
E1	تعدادهجاها در کلمه فعلی
E2	محل کلمه فعلی در عبارت فعلی (پیشرو)
E3	محل کلمه فعلی در عبارت فعلی (پسرو)
F1	تعدادهجاها در کلمه بعدی
G1	تعداد هجاها در عبارت قبلی
G2	تعداد کلمه ها در عبارت قبلی
H1	تعداد هجاها در عبارت فعلی
H2	تعداد کلمه ها در عبارت فعلی
H3	محل عبارت فعلی در جمله (پیشرو)
H4	محل عبارت فعلی در جمله پسرو)
I1	تعداد هجاها در عبارت بعدی
I2	تعداد کلمه ها در عبارت بعدی
J1	تعداد هجاها در جمله
J2	تعداد کلمات در جمله
J3	تعداد عبارت ها در جمله

۸-۱ آموزش مدلهای طول واج

مراحل کار برای ایجاد مدلها در شکل ۵ مشاهده می شود. در ابتدا برای هر جمله آموزشی، برای هر واج مقادیر پارامتر ها را بدست می آوریم. سپس هر کدام از واجها را به درخت اعمال می کنیم و نمونه مشاهـــدات مربوط به هـــر برگ از این درخت را جمع آوری می کنیم. با داشتن فایل ویژگی مربوط به هر جمله، برای هر نمونه از هر برگ، الگوریتم ویتربی را اعمال می کنیم و دنباله حالت ها را بدست می آوریم. در نهایت با توجه به دنباله حالت ها، از تعداد دفعات حضور در هر حالت، میانگین گیری و آنرا ذخیره می نماییم.



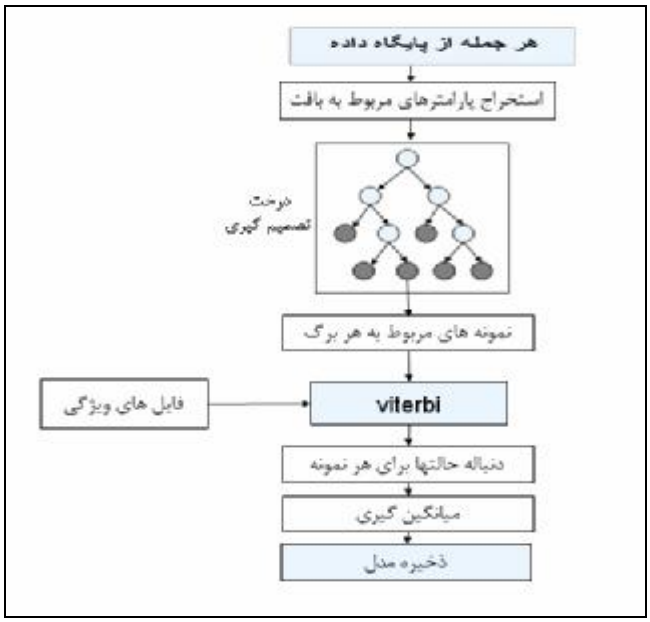
شکل ۴- ساختار استفاده شده برای استخراج پارامترهای مربوط به بافت

۲-۸ سنتز گفتار با داشتن مدل های مربوط به هر برگ از درخت

برای سنتز گفتار، با استفاده از درخت تصمیم گیری برای مشخص کردن طول واجها، از ساختار شکل ۶ استفاده می نماییم. این ساختار همانند ساختار شکل ۲ می باشد با این تفاوت که برای تشخیص طول واج ها به جای مدل های ترایفون از درخت تصمیم گیری استفاده می شود. برای سنتز نیز، ابتدا جمله ای که باید سنتز شود، توسط ساختار شکل۴ تحلیل می شود و پارامتر های جدول ۲ برای هر واج استخراج می گردد. با داشتن پارامتر ها، هر کدام از واج ها را به درخت تصمیم گیری اعمال می کنیم و شماره برگ درخت را برای هر واج بدست می آوریم و با توجه به شماره برگ، مدل طول واج بدست می آید. برای بدست آوردن مدل های ویژگی با توجه به پارامتر های p2 و p4 که به تر تیب واج های قبل و بعد را مشخص می کنند، می توان مدل ترایفون هر واج را بدست آورد. بنابراین مدل های ضرائب، گام و طول برای هر واج بدست می آیند. ادامه کار برای تولید گفتار از این پارامترها، همانند مطالبی است که در مورد ترایفونها ذکر گردید

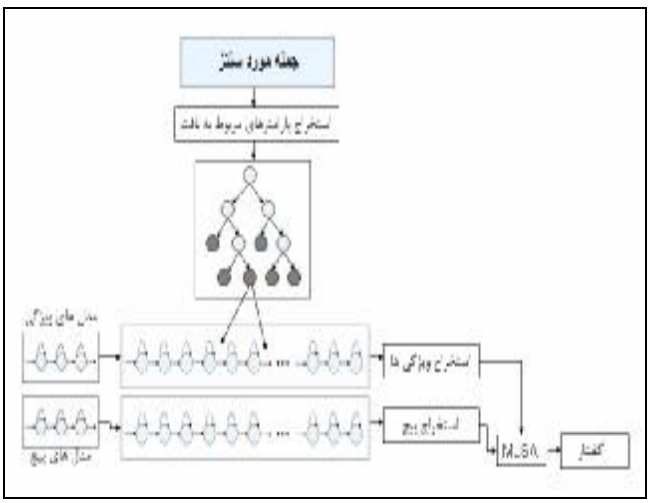
۳-۸ آموزش مدلهای گام

آموزش مدلهای گام، با استفاده از برچسب های موجود در فایل برچسب انجام می گیرد. بنابراین برای آموزش مدلهای گام با استفاده از درخت تصمیم گیری، نیاز به تغییر برچسب های موجود در فایل برچسب داریم. تاکنون مدلهای آموزشی برای گام، مدلهای ترایفون بود و فایل های برچسب نیز برچسب ترایفون داشتند ولی اکنون لازمست که فایل های برچسب را برای آموزش مدلهای HMM گام، تغییر دهیم و نام برچسب ها را با توجه به شماره حالت و شماره برگ درخت، تعیین نماییم. در شکل ۷، نحوه تغییر برچسب ها را مشاهده می کنیم. در این مثال واج "a" را می بینیم که با در نظر گرفتن واج ما قبل و ما بعد از آن بر چسب ترایفون این واج "]a$" می شود. برای تغییر برچسب باید شماره برگ واج را در هر کدا م از درخت های مربوط به حالتهای اول، دوم و سوم، تعیین نماییم. برای اینکار، واج را با توجه به پارامترهای آن که با استفـاده از ساختـار شکل ۲ بدست آمـده، به هر کدام از درخت های مربوط به حالت اول، دوم و سوم از مدل HMM گام اعمال می کنیم و شماره برگ درخت را برای هر کدام از این درخت ها بدست می آوریم. فرض کنید برای درخت های حالت اول، دوم و سوم به ترتیب شماره برگهای ۱۰، ۲۵۸ و ۴۵ بدست آمده باشد. با توجه به این شماره ها، برچسب واج مورد نظر برای



شکل ۵- ساختار مورد استفاده برای آموزش مدلهای طول واج با استفاده از درخت تصمیم گیری

حالت های اول، دوم و سوم را به ترتیب با "P1-10"، "P2-258" و "P3-45" نامگذاری می کنیم.در این نامگذاری "P"، نشاندهنده اینست که این برچسب مربوط به گام می باشد. شماره بعد از آن، شماره حالت مربوط به گام است و پس از آن شماره برگ مربوط به آن حالت را لحاظ می نماییم. حال که نحوه نامگذاری جدید حالتها را تعیین نموده ایم، باید مرز بین حالت های اول، دوم و سوم از واج را مشخص نماییم. برای اینکار دنباله بردارهای ویژگی مربوط به واج را طبق الگوریتم ویتربی با مدل HMM آن مقایسه می کنیم و تعیین می نماییم که هر حالت از HMM، چند بار تکرار می شود.



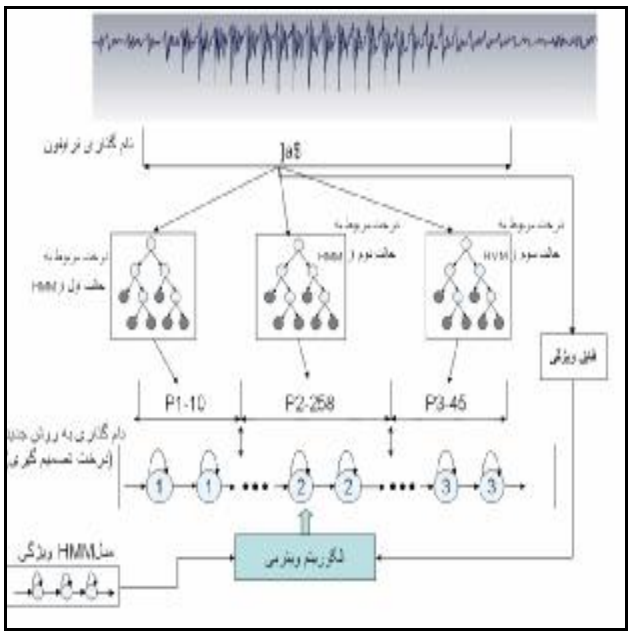
شکل ۶- ساختار مورد استفاده برای سنتز با استفاده از مدلهای طول واج تولید شده با استفاده از درخت تصمیم گیری

پس از اینکه تعداد تکرارهای هر حالت از HMM را بدست آوردیم با توجه به نسبت بین آنها، مرز بین حالتها را تعیین می کنیم. بنابراین به طور کلی، برای تغییر نامگذاری بر چسب ها، فایل های برچسب موجود در دادگان را، یکی یکی باز می کنیم و برای هر واج موجود در فایل، با توجه به مطالب ذکر شده، مرز بین حالت ها و نام هر حالت را پیدا می کنیم و بر چسب های جدید را با توجه به زمان ابتدا و انتهای هر حالت، ذخیره می نماییم. این کار را برای تمامی واجهای موجود در دادگان انجام می دهیم. نمونه ای از فایل بر چسب ایجاد شده به این روش را در

شکل ۸ مشاهده می کنیم. حال که نامگذاری جدید برچسب ها را تعیین نمودیم، باید مدل های HMM مربوط به گام را آموزش دهیم. مدلهای HMM جدید، باید مدلهای یک حالته باشد، چرا که در هنگام برچسب زدن مرز بین حالتها را مشخص نموده ایم و در واقع هر برچسب نشاندهنده یک حالت از مدل HMM است. حال این مدلهای HMM گام را آموزش می دهیم و ذخیره می نماییم تا در هنگام سنتز مورد استفاده قرار گیرند.

۸-۴ سنتز گفتار با استفاده از درخت تصمیم گیری برای مدل های گام و طول واج

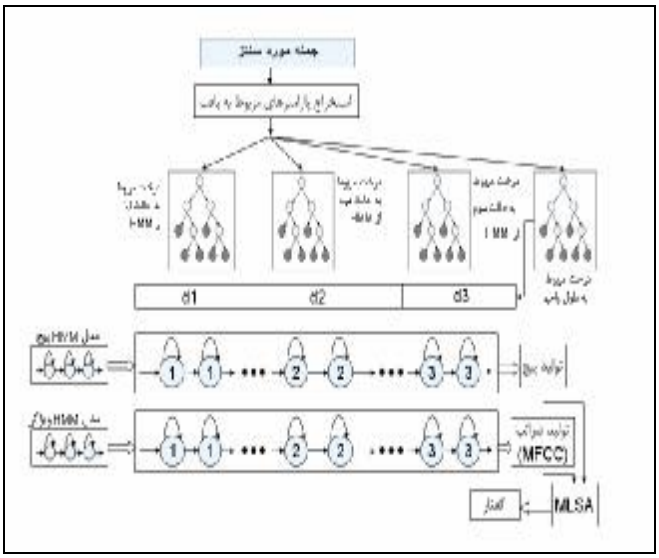
ساختار مورد استفاده برای سنتز گفتار با استفاده از درخت تصمیم گیری برای مدل های گام و طول واج را در شکل ۹ مشاهده می نماییم. در ابتدا دنباله واج ها را به ساختار شکل ۴ اعمال می کنیم و برای هر واج مقادیر پارامترهای جدول ۲ را بدست می آوریم. سپس برای هر واج مورد نظر با پیمایش در درخت مربوط به طول واج، شماره برگ هر درخت را بدست می آوریم. با توجه به شماره برگ درخت، مدل طول واج را فراخوانی می کنیم و تعیین می کنیم که هر کدام از حالت های HMM، چندبار باید تکرار شوند. سپس واج مورد نظر را به هر کدام از درخت های مربوط به حالت اول، دوم و سوم از HMM اعمال می کنیم و شماره برگ مربوط به هر درخت را بدست می آوریم.



شکل ۷- نامگذاری واجها با توجه به درخت تصمیم گیری برای گام

```
41556816    41726412    p1-2
41726412    41896008    p2-62
41896008    41980812    p3-27
41980812    42234688    p1-36
42234688    42403940    p2-222
42403940    42573196    p3-13
42573188    43044720    p1-1
43044720    43233332    p2-1
43233332    43516256    p3-2
43516256    43790372    p1-31
43790372    44521360    p2-459
44521360    44612732    p3-35
44612736    44796476    p1-25
44796476    45072088    p2-438
45072088    45255828    p3-34
```

شکل ۸- نمونه ای از فایل برچسب برای آموزش مدلهای گام با استفاده از درخت تصمیم گیری



شکل ۹- ساختار مورد استفاده برای سنتز گفتار با استفاده از درخت تصمیم گیری برای گام و طول واج

با توجه به شماره برگ درخت ها، نام مدل گام مربوط به هر حالت را بدست می آوریم و مدل HMM آنرا فراخوانی می کنیم. همچنین مدل ترایفون مربوط به ویژگی MFCC را نیز فراخوانی می نماییم. حال با توجه به تعداد تکرار های هر حالت از HMM، که با استفاده از مدل طول بدست آورده ایم هر حالت از مدل HMM ویژگی های MFCC و گام را به تعداد مورد نظر تکرار می کنیم و ذخیره می نماییم. با انجام این کار، دنباله ای از پارامترها شامل بردارهای میانگین و ماتریس های کوواریانس برای ویژگی های MFCC و گام بدست می آید. سپس الگوریتم بیان شده در بخش های قبل را روی هر کدام از دنباله پارامترها اعمال می کنیم و مقادیر نهایی ضرائب کپسترال و گام را بدست می آوریم. سپس این ضرائب را مستقیما به فیلتر MLSA، اعمال می کنیم و گفتار حاصل بدست می آید.

۹- ارزیابی

برای ارزیابی سیستم های سنتز گفتار عمدتا پارامترهای قابل فهم بودن، طبیعی بودن و خوشایند بودن صدای سنتز شده، مورد ارزیابی قرار می گیرد. قابل فهم بودن، میزان وضوح وقابل درک بودن صدا را نشان می دهد. طبیعی بودن، مشخص میکند که صدای سنتز شده چقدر به صدای طبیعی انسان نزدیک است و خوشایند بودن نیز مشخص می کند که صدای تولید شده، تا چه حدی برای شنونده خوشایند است.

برای ارزیابی صدای سنتز شده، روشهای مختلفی پیشنهاد شده است که تست MOS [۴] وDRT[۵]، از جمله مهمترین این روشها به شمار می روند[21] و [22] در تست MOS، چند جمله مختلف با مشخصات متفاوت از لحاظ گوناگونی واجها، برای تعدادی شنونده پخش می شود. هر شنونده یک ارزیابی از پارامترهای قابل فهم بودن، طبیعی بودن و خوشایند بودن ارائه می نماید. این ارزیابی به این صورت است که به ازاء هر کدام از پارامترهای فوق یک نمره به هر جمله داده می شود. نمرات مورد استفاده و مفهوم آنها، در جدول ۳ مشاهده می شود.

جدول ۳- نمرات و مفهوم آنها در تست MOS

نمره	مفهوم
۱	بد
۲	ضعیف
۳	نسبتا خوب
۴	خوب
۵	فوق العاده

سپس بازاء هر کدام از پارامترهای فوق، از نمرات میانگین گیری می کنیم و نمره نهایی را مشخص می نماییم. در این مقاله تست MOS را با استفاده از ده جمله انجام داده ایم که پنج جمله از این جملات، در دادگان موجود می باشند (جملات آموزشی) و پنج جمله دیگر، در دادگان موجود نمی باشند(جملات آزمایشی). برای تشکیل جدولهای ۴ و ۶ از ۱۰ شنونده و برای تشکیل جدول ۵ از ۲۰ شنونده استفاده نموده ایم. برای انجام تست MOS، در ابتدا جملات را طبق مطالب ذکر شده در بخش ۷ سنتز نموده ایم. یعنی برای استخراج ویژگی ها، گام و طول زمانی واج ها صرفا از مدل های ترایفون استفاده نموده ایم. نتایج حاصل از این تست در جدول۴ مشاهده می شود.

سپس همین جملات را با استفاده از الگوریتم بیان شده در بخش ۸ سنتز نموده ایم، یعنی برای تولید طول زمانی واج ها و کنتور گام از درخت های تصمیم گیری استفاده نموده ایم. نتایج حاصل از این تست در جدول ۵ مشاهده می شود.

با مقایسه دو جدول فوق متوجه می شویم که استفاده از درخت های تصمیم گیری تاثیر به سزائی در بالا رفتن کیفیت صدای سنتز شده داشته است به طوری که هر سه پارامتر مربوط به تست MOS افزایش قابل ملاحظه ای داشته اند.

در یک اقدام دیگر، کنتور گام را برای همان جملات، با استفاده از روش فوجی ساکی ارائه شده در مرجع [23] بدست آوردیم. در این روش، پیچ به شکل ترکیبی از یک قسمت عمومی[۴] و تعدادی اجزای محلی[۵] در نظر گرفته می شود. قسمت عمومی برای مدل کردن ریتم کلی گفتار به کار می رود. مثلا در جملات خبری، مقدار پیچ با نزدیک شدن به انتهای جمله کاهش می یابد و ریتم خبری بودن به جمله می دهد. اجزاء محلی برای مدل کردن تکیه[۶] به کار می روند. اگر فقط از جزء عمومی استفاده نماییم، به دلیل یکنواخت بودن پیچ، جمله ریتم یکنواختی به خود میگیرد و حالت طبیعی خود را از دست می دهد. با استفاده از اجزاء محلی، تغییرات مناسب زیر و بمی باعث حالت دادن به ریتم جمله می شود و به آن حالت طبیعی تری می دهد. بدین ترتیب برای تولید کنتور گام از روش فوجی ساکی و برای تولید طول زمانی واجها و بردارهای ویژگی از همان روش های قبل استفاده نمودیم.

جدول ۴- نتایج حاصل از تستMOS روی جملات سنتز شده (با استفاده از مدلهای ترایفون برای طول واج، گام و ویژگیها)

جمله	قابل فهم بودن	طبیعی بودن	خوشایند بودن
تئاتر امکانی برای ارائه حقیقت زندگی و یا حکمت زندگی است.	۳/۱	۳/۵	۳/۴
در دو سال گذشته تلاش موفقی صورت گرفت.	۴/۱	۳/۹	۳/۹
داستان شنل را قبل از دیدن نمایش نامه خواندم اما نمایش حادثه قریبی بود.	۳/۹	۳/۶	۳/۴
در مورد داوری هایی هم که در باره هنر نمایش صورت می گیرد.	۴/۲	۳/۸	۳/۶
نمایش دندون طلا که تا به امروز بیشترین استقبال را داشته است.	۳/۹	۴/۵	۳/۲
میانگین	۳/۸	۳/۹	۳/۵

نتایج حاصل از این تست را در جدول ۶ مشاهده می نماییم. همانطور که مشاهده می شود، در این حالت پارامترهای قابل فهم بودن و خوشایند بودن امتیازات بهتری را نسبت به جدول قبل کسب نموده اند که علت آن اینست که چون گام تولید شده سطح هموارتری دارد بنابراین کیفیت صدا تا حدودی بهتر می شود اما هموار بودن کنتور گام باعث می شود که صدای حاصل از این روش کاملا طبیعی نباشد و تا حدودی حالت رباتی به خود گیرد و با توجه به جدول نیز مشاهده می شود که این روش در پارامتر طبیعی بودن امتیاز کمتری بدست آورده است.

با مقایسه نتایج جدول ۶ با سنتز کننده پیاده سازی شده توسط آیت [21] و [24] متوجه می شویم که این سنتزکننده از لحاظ هر سه پارامتر قابل فهم بودن، طبیعی بودن و خوشایند بودن امتیازات بهتری را کسب نموده است.

سنتز کننده طراحی شده توسط آیت به روش هارمونیک نویزی می باشد. نتایج بدست آمده برای این سنتز کننده به ازای پارامترهای فوق به ترتیب ۴/۱۸، ۳/۴۲ و ۳ میباشد.

یکی دیگر از روشهای تست وارزیابی سیستم های سنتز روش DRT می باشد. در تست DRT، تعدادی زوج کلمه تک هجا به شکل /CVC/ که تنها در واج اول با یکدیگر تفاوت دارند انتخاب می کنیم. این واجها باید در یک صفت از صفات voicing، nasality، sustension، sibilation، graveness، compactness و vowel-like بودن متفاوت باشند. سپس کلمات را توسط سنتز کننده سنتز می کنیم و از تعدادی شنونده می خواهیم که به کلمات سنتز شده گوش فرا دهند و برای هر زوج کلمه مشخص کنند که آیا کلمات با یکدیگر متفاوت هستند یا خیر. در نهایت نتیجه این تست توسط رابطه زیر تعیین می شود:

$$DRT(\%) = \frac{N_{correct} - N_{incorrect}}{N_{test}} \times 100 \qquad (۳۹)$$

که در آن $N_{correct}$ تعداد پاسخ های درست، $N_{incorrect}$ تعداد پاسخ های نادرست، N_{test} تعداد کل کلمات می باشد.

جدول ۵- نتایج حاصل از تستMOS روی جملات سنتز شده (تولید گام و طول زمانی واج ها با استفاده از درخت تصمیم گیری)

جمله	قابل فهم بودن	طبیعی بودن	خوشایند بودن
تئاتر امکانی برای ارائه حقیقت زندگی و یا حکمت زندگی است.	۳/۹	۴/۲	۴/۲
در دو سال گذشته تلاش موفقی صورت گرفت.	۴/۸	۴/۵	۴/۳
داستان شنل را قبل از دیدن نمایش نامه خواندم اما نمایش حادثه قریبی بود.	۳/۳	۴/۳	۳/۸
در مورد داوری هایی هم که در باره هنر نمایش صورت می گیرد.	۴/۳	۴/۳	۴/۱
نمایش دندون طلا که تا به امروز بیشترین استقبال را داشته است.	۴/۹	۴/۸	۳/۹
میانگین	۴/۲	۴/۴	۴/۱
ابوذر برخاست و به نماز ایستاد و چهار رکعت نماز خواند.	۴/۳	۳/۸	۳/۳
بازی از لحاظ جسمی باعث هماهنگی عضلات بدن می شود.	۴/۲	۴/۶	۳/۴
در دنیای امروز آموزش خانواده ضرورتی انکار ناپذیر محسوب می گردد.	۴/۵	۴	۳/۳
در این کتاب زندگی وی مورد بررسی قرار گرفته است.	۴/۹	۴/۸	۴
در فصل بعدی کتاب او احترام خود را نثار دو نوع از انسانها می کند.	۳/۸	۴	۳
میانگین	۴/۳	۴/۲	۳/۴

درارزیابی پیاده سازی شده در این مقاله زوجهای انتخاب شده عبارتند از: (نار، مار)، (وی، ری)، (چاه، کاه)، (شاه، چاه)، (حال، خال)، (جاه، چاه)، (یال، وال)، (دود، بود)، (گوی، روی)، (شور، سور)، (غاز، باز)، (ناز، راز)، (پس، کس)، (راد، یاد)، (مَه،

وَه)، (داد، یاد)، (ره، وه)، (بام، شام)، (در، تر)، (تو، بو)[21]. تعداد ده شنونده به این زوج کلمات گوش داده و صحت تلفظ کلمات را مشخص نموده اند. کلمات فوق با استفاده از سیستم سنتز پیشنهادی که در آن تعیین طول زمانی تراپفونها و گام با استفاده از درخت تصمیم گیری انجام شده است سنتز، و برای شنوندگان پخش شده است. با توجه به نظرات شنوندگان، تعداد کلمات نادرست ۲۴ و تعداد کلمات صحیح ۳۷۶ عدد بدست آمده است. با جایگذاری در رابطه فوق نتیجه تست ۸۸ درصد بدست می آید. با در نظر گرفتن این نکته که در تستهای انجام شده بطور کامل شرایط تست رسمی DRT رعایت نشده است، لیکن یک سنتز کننده خوب سنتز کننده ای است که تست DRT آن بین ۸۵ تا ۹۰ درصد باشد[22] .

جدول ۶- نتایج حاصل از تستMOS روی جملات سنتز شده (تولید طول زمانی واج ها با استفاده از درخت تصمیم گیری و تولید گام به روش اتوماتیک)

جمله	قابل فهم بودن	طبیعی بودن	خوشایند بودن
تئاتر امکانی برای ارائه حقیقت زندگی و یا حکمت زندگی است.	۴/۴	۴/۳	۴/۸
در دو سال گذشته تلاش موفقی صورت گرفت.	۴/۹	۴/۲	۴/۶
داستان شنل را قبل از دیدن نمایش نامه خواندم اما نمایش حادثه قریبی بود.	۴/۳	۴/۴	۴/۴
در مورد داوری هایی هم که در باره هنر نمایش صورت می گیرد.	۴/۴	۴/۲	۴/۶
نمایش دندون طلا که تا به امروز بیشترین استقبال را داشته است.	۴/۸	۴/۴	۴/۲
میانگین	۴/۶	۴/۳	۴/۵
ابوذر برخاست و به نماز ایستاد و چهار رکعت نماز خواند.	۴/۴	۴	۴/۵
بازی از لحاظ جسمی باعث هماهنگی عضلات بدن می شود.	۴/۴	۴/۱	۴/۴
در دنیای امروز آموزش خانواده ضرورتی انکار ناپذیر محسوب می گردد.	۴/۸	۴/۲	۴/۳
در این کتاب زندگی وی مورد بررسی قرار گرفته است.	۴/۸	۴/۳	۴/۴
در فصل بعدی کتاب او احترام خود را نثار دو نوع از انسانها می کند.	۴/۱	۳/۶	۴/۲
میانگین	۴/۵	۴/۰	۴/۴

۱۰- نتیجه گیری

در این مقاله برای مدل کردن پارامترهای مورد استفاده در سنتز کننده از مدل مخفی مارکف و برای بدست آوردن ویژگی ها، از آنالیز کپسترال مبتنی بر معیار مل که الهام گرفته از سیستم شنیداری گوش انسان است استفاده شده است. روش استفاده شده در این مقاله، که ضرائب کپسترال را با استفاده از مینیمم کردن معیار بایاس نشده لگاریتم طیف با توجه به ضرائب کپسترال بدست می آورد، روش مناسبی برای استخراج ضرائب کپسترال می باشد و به راحتی می توان با استفاده از فیلتر MLSA، آنرا به گفتار تبدیل نمود.

روش استخراج گام با استفاده از اتوکورلیشن اصلاح شده نیز نتایج خوبی را از خود نشان داده است. با در نظر گرفتن خطای مربوط به نمونه برداری و پنجره گذاری، نتایج حاصل ازاین روش بسیار مناسب است. در الگوریتم مورد استفاده برای تولید پارامترهای سنتزگفتار از مدلهای HMM، علاوه بر خود ویژگی ها مشتقات اول و دوم ویژگیها نیز مورد استفاده قرار گرفته است. استفاده از مشتقات ویژگیها باعث می شود که تاثیر بافت در تولید پارامترها لحاظ شود و در نتیجه نتایج مطلوبتری حاصل گردد.

اگر در تولید پارامترها از مدلهای HMM، مشتقات پارامترها لحاظ نشود، پارمترهای تولید شده همان بردارهای میانگین HMMها خواهند بود و بردارهای واریانس مورد استفاده قرار نمی گیرند.

در این مقاله، واحد مورد استفاده برای آموزش مدلهای HMM تراپفون می باشد که منظور از آن واجی است که در نامگذاری آن، نام واج ما قبل و ما بعد آن نیز لحاظ شده است. علاوه بر خود واج و واج های مجاور، پارامترهای دیگری نیز روی تلفظ واج ها تاثیر می گذارند. ازجمله این پارامترها می توان به محل قرار گرفتن واج در هجا،کلمه، عبارت و جمله، محل قرار گرفتن تکیه ها و مواردی از این قبیل اشاره نمود. این پارامترها بر روی گام و طول زمانی واجها تاثیر بسزایی دارند. بنابراین برای تولید مدلهای گام و طول زمانی واجها، باید این پارامترها را به نحوی، مورد استفاده قرار داد.

در این مقاله برای در نظر گرفتن این پارامترها از درخت های تصمیم گیری استفاده نموده ایم.

برای مدل طول زمانی واج ها از یک درخت و برای هر حالت از مدل HMM گام، یک درخت مورد استفاده قرار گرفته است. بنابراین چهار درخت در این سیستم استفاده شده است. در مرحله آموزش، در هر برگ از درخت تصمیم گیری یک مدل HMM تشکیل می گردد و در مرحله سنتز مورد استفاده قرار می گیرد. استفاده از درخت های تصمیم گیری برای در نظر گرفتن پارامترهای ذکر شده روش مناسبی می باشد و نتایج بدست آمده در این مقاله گواه این مدعاست چرا که طول زمانی واجها و گام که با استفاده از درختهای تصمیم گیری تولید شده اند، طبق نتایج ذکر شده در مرحله ارزیابی، نزدیک به مقادیر طبیعی می باشند. در این مقاله، روش اتوماتیک تولید گام نیز مورد بررسی قرار گرفته است. در روش مورد استفاده برای مدل کردن شکل کلی سیگنال گام از یک جزء عمومی و برای مدل کردن تکیه ها از اجزاء محلی استفاده می شود.

نتایج بدست آمده از این روش نیز بسیار جالب توجه است و گفتار تولید شده، تا حدود خیلی زیادی نزدیک به گفتار طبیعی می باشد. ارزیابی سیستم سنتز گفتار توسط تست MOS و DRT، انجام گرفته است. امتیازات بدست آمده از تست MOS برای جملات آموزشی، برای پارامترهای قابل فهم بودن، طبیعی بودن و خوشایند بودن به ترتیب ۴/۲، ۴/۴ و ۴/۱، و برای جملات آزمایشی به ترتیب ۴/۳، ۴/۲ و ۳/۴ می باشد. نتیجه بدست آمده برای تست DRT نیز برای حالتی که از درخت های تصمیم گیری برای تعیین طول زمانی تراپفون و گام استفاده نموده ایم برابر ۸۸٪ می باشد.

قدردانی

این تحقیق مورد حمایت مالی مرکز تحقیقات مخابرات ایران قرار گرفته است.

مراجع

[1] T. Masuko, *HMM-Based Speech Synthesis and Its Aplications*, PhD. Thesis, November 2002.

[2] R. E. Donovan, *Trainable Speech Synthesis*, PhD. Thesis, Cambridge University Engineering Department, Cambridge, UK, 1996.

[3] R. E. Donovan and E. M. Eide, "The IBM Trainable Speech Synthesis System," *ICSLP*, 1998.

[4] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Meredith, and M. Plumpe, "Recent Improvements on Microsoft's Trainable Text-to-Speech System: Whistler," *ICASSP*, 1997.

[5] X. Huang, A. Acero, J. Adcock, H. Hon, J. Goldsmith, and J. Liu, "Whistler: A Trainable Text-to-Speech System", *ICASSP*, 1996.

[6] M. lamura, T. Masuko, K. Tokudat, and T. Kobay, "Adaptation of Pitch and Spectrum for HMM-based Speech Synthesis Using MLLR," *ICASSP*, Vol. II, pp. 805-808, 2001.

[7] H. Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe, "Automatic Generation of Synthesis Units for Trainable Text-to-Speech Systems," *ICASSP*, 1998.

[8] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech Synthesis using HMMs with Dynamic Features," *ICASSP*, pp. 389-392, 1996.

[9] K. Tokuda, T. Kobayashi, and S. Imai, "Speech Parameter Generation from HMM Using Dynamic Features," *ICASSP*, pp. 660-663, 1995.

[10] K. Tokuda, T. Masuko, T. Yamada, T. Kobayashi, and S. Imai, "An Algorithm for Speech Parameter Generation from Continuous Mixture HMMs with Dynamic Features", *EUROSPEECH*, pp.757-760, 1995.

[11] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice Characteristics Conversion for HMM-based Speech Synthesis System", *ICASSP*, pp. 1611-1614, 1997.

[12] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech Parameter Generation Algorithms for HMM-based Speech Synthesis," *ICASSP*, vol. 3, pp. 1315–1318, 2000.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous Modeling Of Spectrum, Pitch and Duration In HMM-Based Speech Synthesis", *EUROSPEECH*, pp. 2347-2350, 1999.

[14] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Duration Modeling in HMM-based Speech Synthesis System," *ICSLP*, vol.2, pp. 29-32, 1998.

[15] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based Speech Synthesis System Applied to English," *IEEE Speech Synthesis Workshop*, 2002.

[16] M. Plumpe, A. Acero, H. Hon, and X. Huang, "HMM-Based Smoothing for Concatenative Speech Synthesis," *ICASSP*, 1998.

[17] G. Corrigan, N. Massey, and O. Karaali, "Generating Segment Durations in A Text-To-Speech System: A Hybrid Rule-Based/Neural Network Approach," *EUROSPEECH*, pp. 2675-2678, 1997.

[18] T. Fukuda, K. Tokuda, T. Kobayashi, and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," *ICASSP*, 1992.

[19] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive Cepstral Analysis of Speech," *IEEE Transaction on Speech and Audio Processings*, vol. SA-3, no. 6, pp. 481-489, 1995.

[20] P. Boersma, "Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound," *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 1993.

[[۲۱] س. س. آیت، **طراحی و پیاده سازی سیستم تولید گفتار فارسی با تاکید بر بهبود هرچه بیشتر کیفیت گفتار تولید شده**، پایان نامه کارشناسی ارشد مهندسی کامپیوتر، دانشگاه صنعتی امیر کبیر، ۱۳۷۹.

[22] S. Lemmetty, *Review of Speech Synthesis Technology*, Thesis in Master of science, Helsinki University of Technology, 1999.

[23] E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale, *Improvements in Speech Synthesis*, John Wiley & Sons, 2002.

[۲۴] م. م. همایون پور و س.س آیت، "سنتز گفتار بروش مدل هارمونیک-نویزی،" **کنفرانس سالانه انجمن کامپیوتر ایران**، ۱۳۸۰.

---

[1] Unbiased Estimation of Log Spectrum
[2] Mel Log Spectrum Approximation
[3] Phrase
[4] Global component
[5] Local components
[6] Accent

سید مصطفی موسوی تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی کامپیوتر به ترتیب در سالهای ۱۳۸۰ و ۱۳۸۲ در دانشگاه شهید بهشتی و دانشگاه صنعتی امیر کبیر به پایان رسانید. زمینه های تحقیقاتی مورد علاقه ایشان عبارتند از: پردازش سیگنال، پردازش گفتار، تحلیل و طراحی نرم افزار.

آدرس پست الکترونیکی ایشان عبارت است از:
Smmousavi_ir@yahoo.com

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 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 بر تدریس، راهنمایی پروژه های کارشناسی، کارشناسی ارشد و دکتری در زمینه های مهندسی کامپیوتر و فناوری اطلاعات و نیز هدایت تعداد زیادی پروژه های صنعتی و ملی را برعهده داشته است. نامبرده عضو انجمن های علمی کامپیوتر، ارتباطات و فناوری اطلاعات و رمز می باشد و مسئولیت های اجرایی متعدد از جمله ریاست و معاونت های آموزشی و پژوهشی دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیر کبیر و شرکت در برگزاری چندین کنفرانس و مسابقه علمی را بر عهده داشته و موفق به انتشار بیش از ۸۰ مقاله علمی- پژوهشی در مجلات و کنفرانس های علمی داخل و خارج از کشور گردیده است.

آدرس پست الکترونیکی ایشان عبارت است از:
homayoun@ce.aut.ac.ir
و آدرس سایت اینترنتی:
www.autice.org