سنتز صحبت طبیعی فارسی با استفاده از روش اتصال قصعات صحبت[1]

دکتر محمود تابنده مسلم نوده‌محمدمهدی کلانترهرمزی
دانشگاه صنعتی شریف ـ دانشکدهٔ مهندسی برق

چکیده: در این مقاله روش اتصال قطعات صحبت برای سنتز صحبت مورد بررسی قرار می‌گیرد و ضمن مقایسهٔ مختصر روش‌های مختلف، یک روش مناسب برای سنتز صحبت فارسی پیشنهاد می‌شود. در این پژوهش از دایفون‌ها[2] به‌عنوان قطعات مناسب استفاده می‌شود. در ادامه راه‌حل‌هایی که برای بهبود صحبت تولیدشده استفاده شده است و همچنین اندکی در مورد اعمال اطلاعات آهنگین[3] به خروجی توضیح داده می‌شود.

کلمات کلیدی: ۱- سنتز صحبت ۲- روش اتصال قطعات ۳- دایفون ۴- پریودگام

۱- مقدمه

تولید مصنوعی صحبت دارای استفاده‌های گوناگونی است و چنانچه بتوان صحبت قابل فهم و طبیعی با ماشین تولید نمود، می‌تواند در موارد زیر مورد استفاده قرار گیرد: صنعت مالتی‌مدیا جهت پرهیز از ذخیرهٔ سیگنال‌های حجیم صوتی، سیستم‌های پاسخگو، واسط کاربر برنامه‌های جدید کامپیوتری، تبدیل پیغام‌های پست الکترونیکی[4] به سیگنال صحبت، خواندن کتاب برای نابینایان و استفاده‌های متنوع دیگر.

به‌طور عمده دو روش برای سنتز صحبت مورد استفاده قرار می‌گیرد:

الف ـ سنتز صحبت با استفاده از فرمنت‌ها[5]: در این روش مجرای صوتی توسط فیلتر با دو منبع تولید سیگنال مدل می‌شود. واج‌های واک‌دار با توجه به فورمنت‌های آن واج توسط یک سیگنال پریودیک ساخته می‌شوند. برای

---

[1] - Concatenative Speech Synthesis
[2] - Diphone
[3] - Prosody
[4] - Email
[5] - Formant

واج‌های بی‌واک، منبع تولید نویز در نظر گرفته می‌شود. گرچه با استفاده از این روش می‌توان صحبت قابل فهمی تولید نمود، اما صدای تولیدشده رباتیک و غیرطبیعی است.

ب ـ روش اتصال قطعات صحبت: در این روش صحبت طبیعی به قطعات مناسب تقطیع و در یک بانک ذخیره می‌شوند. در هنگام بازسازی سیگنال، این قطعات به‌طرز مناسب کنار هم قرار می‌گیرند و صحبت تولید می‌شود. با استفاده از این روش می‌توان صحبت قابل فهم و طبیعی تولید نمود، به‌گونه‌ای که قابل استفاده در کاربردهایی نظیر مالتی‌مدیا که احتیاج به کیفیت بالا دارند، باشد. یکی از مهم‌ترین مراحل استفاده از این روش، انتخاب نوع قطعهٔ انتخاب‌شده و نحوهٔ ترکیب است.

در این پژوهش از روش دوم یعنی روش اتصال قطعات صحبت برای تولید صحبت استفاده شد که در ادامه توضیحات بیشتری در مورد آن ارائه می‌گردد.[۳]

۲- انتخاب قطعهٔ مناسب

قطعهٔ انتخاب‌شده برای ذخیره و بازسازی بسته به نوع ماربرد و کیفیت موردنظر می‌تواند دارای طول‌های متفاوت باشد.

الف ـ کلمه: در این روش از کنار هم قراردادن کلمات از پیش ضبط‌شده جملات ساخته می‌شود. همانطور که انتظار می‌رود قواعد ترکیب در این روش ساده‌تر است و صحبت تولیدشده تا حد زیادی قابل فهم خواهد بود اما با استفاده از این روش نمی‌توان جملات نامحدودی تولید نمود. همچنینی حافظهٔ مورد استفاده نسبت به روش‌های دیگر زیاد است. در کاربردهایی که تعداد کلمات محدود باشد، این روش می‌تواند مورد استفاده قرار گیرد.

ب ـ هجا: در این روش هجاها ذخیره می‌شوند. هجا یک رشتهٔ آوایی پیوسته است و در زبان فارسی از یک واکه و یک تا ۳ همخوان تشکیل شده است. در فارسی هجا به ۳ صورت CV (C: همخوان و V: واکه) مانند "با"، CVC مانند "سیب" و CVCC مانند "گفت" می‌باشد. از آنجا که واجگان فارسی از ۶ واکه و ۲۳ همخوان تشکیل شده است، کل هجاهای ممکن ۷۶۳۱۴ خواهد بود ولی در عمل به دلیل محدودیت‌های هم‌نشینی کمتر از یک‌دهم رقم فوق می تواند وجود داشته باشد.[۱]

چنانچه ملاحظه می‌شود، یافتن و ضبط همهٔ هجاها بسیار مشکل می‌باشد. همچنین حجم لازم جهت ذخیرهٔ این مقدار هجا بسیار زیاد خواهد بود.

ج ـ واج: واج کوچکترین جزء جداپذیر صحبت در هر زبان می‌باشد. گرچه با استفاده از واج به‌عنوان قطعهٔ صحبت، حجم بانک داده بسیار کم می‌شود، اما در عوض باید قواعد پیچیده‌ای در ترکیب رعایت شود.

د ـ دایفون: دایفون عبارت است از قطعه‌ای از صحبت که از مرکز پایدار یک واج شروع و در مرکز پایدار واج بعدی خاتمه می‌یابد و شامل ناحیهٔ گذر بین دو واج می‌باشد.[۴] در این روش علاوه بر حل مسألهٔ ناحیهٔ مرزی واج‌ها، تأثیر واج قبل و بعد نیز در هر واج در نظر گرفته می‌شود.

با توجه به توضیحات بالا، در این پروژه از قطعهٔ دایفون برای تولید صحبت استفاده شد.

۳- استخراج دایفون از صحبت فارسی

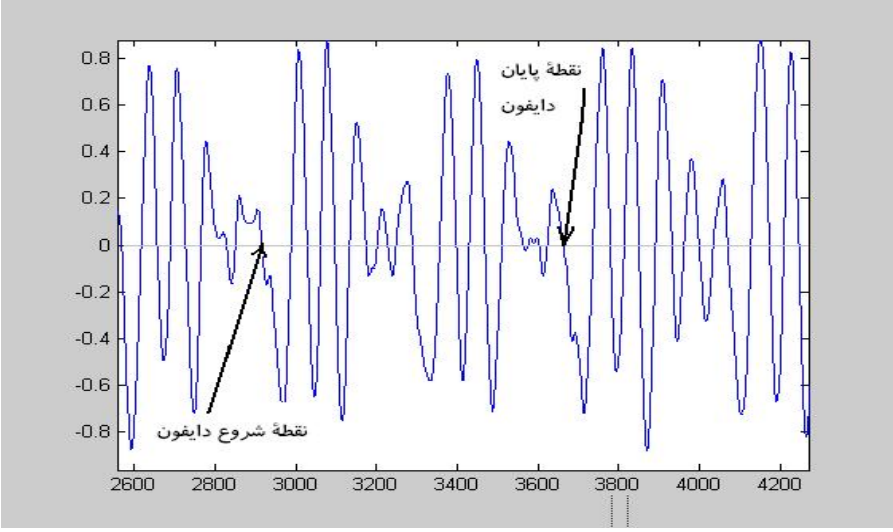
می‌توان محاسبه کرد که تعداد دایفون‌ها در زبان فارسی ۱۰۰۲ عدد است.[۱] اگر بازهٔ زمانی هر دایفون به‌طور متوسط ۱۵۰ میلی‌ثانیه باشد، کل فضای لازم جهت ذخیرهٔ دایفون‌ها برابر مقدار زیر خواهد بود:

$$H = 1002 * 0.15 * B * S$$

چنانچه دایفون‌ها با کیفیت CD (B = ۲ و S = ۴۴۱۰۰) ذخیره شوند، H برابر ۱۳ MB خواهد بود.

دایفون‌ها را نمی‌توان به‌صورت جداگانه ضبط نمود و باید آنها را از صحبت طبیعی استخراج نمود. گرچه تحقیقاتی در مورد "جداکردن خودکار"[1] دایفون‌ها در زبان‌های دیگر انجام شده است[۵]، اما در زبان فارسی هنوز فعالیت ثمربخشی صورت نپذیرفته است. بنابراین در این پروژه تقطیع دایفون به‌صورت دستی انجام شد. برای این منظور ابتدا باید صحبتی تهیه شود که حاوی تمامی دایفون ها باشد و در هنگام ضبط نیز باید سعی کرد تا محیط ضبط یکنواخت و عاری از نویز باشد.

همانگونه که گفته شد، سیگنال صحبت از آواهای بی‌واک و واک‌دار تشکیل شده است. آواهای واک‌دار شبه متناوبند و آواهای بی‌واک به‌شکل نویز هستند. برای واج‌های واک‌دار محل شروع و خاتمهٔ دایفون مهم است. محل شروع دایفون چنانچه یک آوای واک‌دار باشد، از محل شروع ناحیهٔ کمینهٔ سیگنال شبه پریودیک و محل خاتمهٔ دایفون نیز محل شروع ناحیهٔ کمینه در نقطهٔ صفر خواهد بود. برای آواهای بی‌واک که شکل موج آنها نویزی است، محل شروع یا خاتمه هر جایی می‌تواند باشد. شکل ۱ نمونه‌هایی از ابتدا و انتهای یک دایفون را نشان می دهد.



شکل ۱- نحوهٔ انتخاب ابتدا و انتهای دایفون در سیگنال

همچنین در تقطیع دایفون باید دقت نمود محل شروع دایفون در ناحیهٔ درنگ (و تقریباً وسط ناحیهٔ درنگ) یک واج و محل خاتمهٔ آن نیز در ناحیهٔ درنگ واج مجاور باشد.

در اینجا با ذکر یک مثال نحوهٔ تقطیع و سنتز دایفون‌ها را نشان می‌دهیم. در جملهٔ "با صابون صورتم را شستم"، دایفون‌ها را می‌توان به‌صورت زیر جدا کرد:

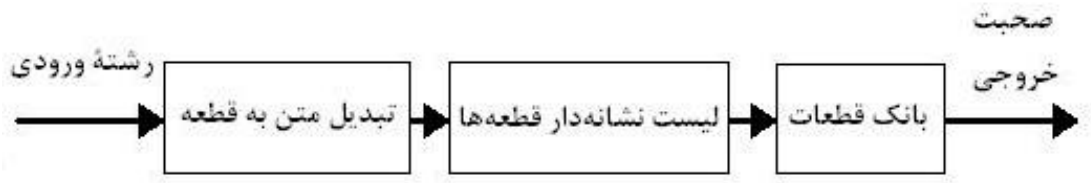
-b + baa + aa- + saa + aa-b + boo + oon + n-s + soo + oo-r + ra + a-t + ta + a-m + m-r + raa + aa-sh + sho + os + s-t + ta + am + m-

در عمل ترکیب باید مکث‌ها نیز به‌نحو مناسب گنجانده شوند.

٤- تبدیل متن به قطعات

در اغلب پروژه‌های سنتز صحبت، از سیستم تبدیل متن به صحبت با دیاگرام شکل ۲ استفاده می‌شود.

[1] - Automatic Extraction



شکل ۲- بلوک‌دیاگرام سیستم

گرچه سنتز صحبت در زبان فارسی با توجه به ویژگی‌های آن نسبت به زبان‌های دیگر ساده‌تر است، اما تبدیل متن به دایفون یا واج بسیار مشکل می‌باشد. زیرا در فارسی واکه‌های "ـَ"، "ـِ" و "ـُ" در متن‌های معمولی کمتر دیده می‌شوند. همچنین سکون و علامت تشدید در متن وجود ندارد. برای حل این مسأله راه‌حل‌های متعددی ارائه شده است که هرکدام دارای مزایا و معایبی است. یکی از این راه‌حل‌ها استفاده از بانکی است که مشابه کلمات متن را به‌همراه اعراب و حرکات لازم در بر داشته باشد.

با استفاده از این روش می‌توان هجاها را از هم جدا نموده و از روی هجاها دایفون‌های مناسب استخراج شوند. مثلاً چنانچه داشته باشیم "مُحَمَّدْ"، رشتهٔ واجگانی به این صورت داریم: "CVCVCCVC". در این رشته از انتها هجاها را جدا می‌کنیم. برای جداکردن هجا باید در نظر داشته باشیم که واکه مرکز هجاست و نمی‌تواند در اول هجا قرار گیرد و قبل از واکه تنها یک همخوان می‌تواند در هجا وجود داشته باشد. بنابراین هجای "CVC" از انتها جدا می‌شود و به همین ترتیب الگوریتم می‌تواند تکرار شود تا همهٔ هجاها به‌دست آیند.

۵- تولید صحبت

تا این مرحله قطعات متن و متناظر با آنها قطعات صحبت با نام‌گذاری مشخص تهیه شد. آنچه برای تولید صوت نهایی لازم است، استفاده از یک محیط برنامه‌نویسی و یک الگوریتم ساده برای اتصال قطعات است. در این پروژه از محیط برنامه‌نویسی visual C ++ و امکانات این محیط برای فایل‌های صوتی استفاده شد.

٦- بهبود خروجی

خروجی چنین سیستمی تا حد مناسبی قابل فهم و طبیعی است اما هنوز با صدای طبیعی انسان فاصله دارد. از جمله علل این موضوع این است که قطعات صحبت از محل‌های مختلف استخراج شده‌اند و دو دایفون در محل‌های مختلف لزوماً دارای ویژگی‌های سیگنالی دقیقاً یکسانی نیستند.

از لحاظ سیگنالی ۳ عامل اصلی وجود دارد که موجب می‌شود صحبت ماشینی با صحبت طبیعی متفاوت باشد که البته برای برطرف‌کردن هر مورد الگوریتم‌ها و روش‌هایی نیز وجود دارد:

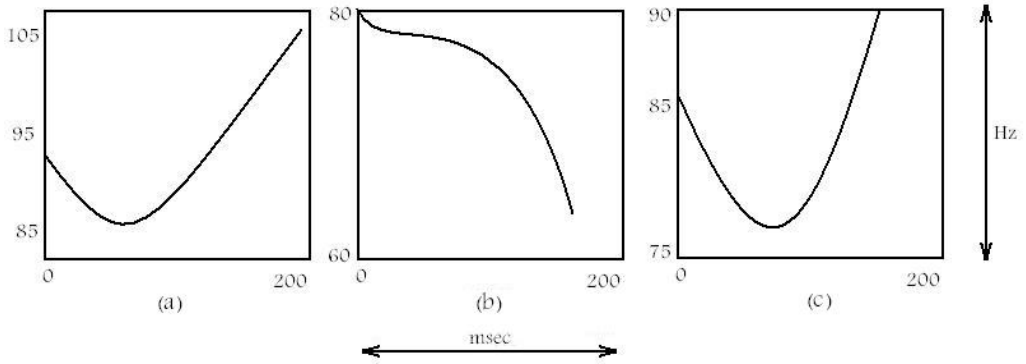
الف ـ اختلاف دامنه: از آنجا که در صحبت طبیعی، میزان دامنهٔ سیگنال و یا بلندی و کوتاهی صوت در مفهوم آن نقش دارد و قطعات مورد استفاده در سیستم هم از صحبت طبیعی استخراج شده‌اند، ممکن است دو دایفون با دامنه‌های غیر مساوی در کنار هم قرار گیرند که این به‌نوبهٔ خود از روان‌بودن خروجی می‌کاهد. نرمالیزه‌کردن دایفون‌ها پیش از اتصال چنانکه در بخش ۷ توضیح داده شده است، می‌تواند این مشکل را تا حدودی برطرف کند.

ب ـ اختلاف فاز: همانگونه در بخش تقطیع دایفون‌ها گفته شد، هنگام جداکردن دایفون‌ها به‌صورت دستی باید به محل تقطیع خیلی دقت کرد. از جمله مواردی که معمولاً به‌هنگام اتصال دایفون‌ها پیش می‌آید و موجب کاهش کیفیت می‌شود این است که دو دایفون دارای فاز یکسان در محل اتصال نیستند. الگوریتم‌هایی وجود دارد که پس از انجام عمل اتصال، چنین اختلاف فازهایی را به حداقل می‌رساند.

ج ـ اختلاف گام[1]: پریود گام یکی از مشخصه‌های اصلی آواهای واکدار است. در طول صحبت شخص و حتی یک جمله پریود گام شخص تغییر می‌کند. این تغییرات در طول یک کلمه کم و به‌آرامی ولی در جمله گاه زیاد و محسوس است. در این مورد در بخش ۷ بیشتر توضیح داده می‌شود. از آنجا که تقطیع دایفون‌ها از جملات مختلف یا محل‌های مختلف در یک جمله صورت می‌گیرد، واج محل اتصال دو دایفون در دو طرف اتصال دارای پریود گام متفاوت است. یعنی ممکن است در طول یک واج تغییرات گام بسیار شدیدتر از حالت طبیعی باشد. برای این مشکل هم الگوریتم‌هایی موجود است که می تواند پریود گام را نرمالیزه کند. PSOLA[2] از جملهٔ این الگوریتم‌ها است.

۷- افزودن آهنگ کلام به خروجی

آهنگ کلام در مفهوم عام ناظر بر تغییرات "زیروبمی" یا گام در گفتار است و چگونگی تولید گفتار از سوی گویندگان و دریافت آن از سوی شنوندگان نشان می دهد که آنان ناخودآگاه و به شیوه‌ای نظام‌مند از زیروبمی در انتقال درک و معنا استفاده می‌کنند. در تقسیم‌بندی‌های زبان‌شناسی، زبان فارسی یک "زبان آهنگی"[3] محسوب می‌شود یعنی زبانی که تغییر سطح و جهت زیروبمی، معنای واژگانی را تغییر نمی‌دهد بلکه فقط در معنای بافتی "پاره‌گفتارها[4]" تأثیر می‌گذارد.[۲] برای مثال یک جملهٔ مشخص می‌تواند در موقعیت‌های مختلف، با توجه به "منحنی زیروبمی"[5] مربوط به آن دارای مفاهیم مختلف امری، سئوالی، تعجبی، خبری و ... داشته باشد. در شکل ۳ منحنی زیروبمی مربوط به چند نوع از جمله‌ها نشان داده شده است.



شکل ۳- منحنی تغییرات گام در جملات (a) امری، (b) پرسشی و (c) خبری

همانطور که ملاحظه می‌کنید یکی از جنبه‌های اصلی صحبت طبیعی افزودن این آهنگ به خروجی است که باید با توجه به محتوا و محل جمله مشخص گردد. برای این کار نیز روش‌های متعددی وجود دارد که از جمله مشهورترین این روش‌ها استفاده از الگوریتم‌های PSOLA و سنتز مجدد[6] است که البته استفاده از این روش‌ها برای رسیدن به هدف مذکور محاسبات پیچیده‌ای را طلب می‌کند.

۸- جمع‌بندی و نتیجه‌گیری

سنتز صحبت فارسی می‌تواند از روش اتصال قطعات صحبت با کیفیتی مناسب انجام پذیرد. انتخاب قطعه باید به‌گونه‌ای باشد که بتواند سیگنال صحبت مشابه سیگنال طبیعی تولید کند. یکی از قطعات مناسب دایفون می‌باشد. حدود هزار دایفون فارسی مورد نیاز است. خروجی چنین سیستمی قابل فهم است اما هنوز دارای مشکلاتی است که می‌توان با

---

[1] - Pitch
[2] - Pitch synchronous Overlap Add
[3] - Intonational Language
[4] - Utterances
[5] - Pitch Contours
[6] - Re Synthesis

روش‌های مختلف پردازش سیگنال صحبت آن را بهبود داد. افزودن آهنگ صحبت به خروجی نهایی می‌تواند تأثیر به‌سزایی در طبیعی‌تر شدن خروجی بگذارد. از جمله کارهایی که میتواند کیفیت چنین سیستمی را بسیار بالا ببرد، استفاده از روش‌های تقطیع خودکار است که قطعاً دقتی بسیار بیشتر از تقطیع دستی خواهد داشت.

منابع

[۱] یدالله ثمره، آواشناسی زبان فارسی، مرکز نشر دانشگاهی، تهران ۱۳۶۴

[۲] محرم اسلامی، واج‌شناسی: تحلیل نظام آهنگ فارسی، انتشارات سمت، چاپ اول، تهران، پاییز ۱۳۸۴

[3] M. Sahandi, Daniel S.G.Vine., Concatenative Speech Synthesis and Animation, The First Annual C.S.I Computer Conference (CSICC 95)

[4] Y.A. EL-Iman, An Unrestricted Vocabulary Arabic Speech Synthesis System, IEEE Trans. on Speech, December 1989

[5] IMIX Programme Preparation Committee, Interactive Multimodal Information Extraction, October 2002