تولید تمام واکه‌های زبان فارسی با در اختیار داشتن تنها یک کلمه از گوینده هدف

محمدجواد جنتی، ابوالقاسم صیادیان*

دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر

چکیده

از جدی‌ترین تنگناها در تبدیل گفتار، نیاز به تعداد بالایی از جملات گوینده هدف برای آموزش سامانه تبدیل گفتار است. شیوه‌های تبدیل موازی گفتار نیازمند حداقل ۵۰ الی ۲۰۰ و شیوه‌های تبدیل غیرموازی نیازمند چندین برابر این تعداد جمله از گوینده هدف هستند. این تحقیق شیوه‌ای را معرفی می‌کند که با کمک آن با داشتن تنها یک کلمه از گوینده هدف، تمام واکه‌های زبان فارسی از آن گوینده تولید می‌شوند. چون در هر کلمه حداقل یک هجا و در هر هجا یک واکه وجود دارد. هر کلمه از گوینده هدف حداقل حاوی یک هجا است. اندیشه اصلی تحقیق طراحی ترادیس‌هایی (تبدیلاتی) مستقیم بین یک واکه و سایر واکه‌ها و تولید مستقیم تمام واکه‌ها از روی آن است. بر این اساس با تولید یک مجموعه ترادیس‌ها (تبدیلات) که به صورت برون‌خط آموزش می‌بینند، می‌توان از روی یک واکه مابقی واکه‌ها را تولید کرد. با توجه به اینکه در زبان فارسی ۶ واکه وجود دارد، در عمل به طراحی ۳۰ ترادیس (تبدیل) نیاز است. این ترادیس‌ها (تبدیلات) با استفاده از توابع گوسی شرطی طراحی و پیاده‌سازی و با استفاده از ۱۰ گوینده آموزشی به‌صورت باسرپرستی آموزش داده شدند. براساس معیار فاصله دُش‌پیچی (اعوجاج)، میانگین فاصله واکه‌های ابداعی از این شیوه و واکه‌های واقعی برابر با ۰٫۴۴۵۹ است.

کلید واژه‌ها: تبدیل واکه، تبدیل گفتار، آموزش غیرموازی.

۱. مقدمه

پس از ایما و اشاره اولین راه ارتباطی کشف شده توسط بشر گفتار است. این شیوه ارتباطی علاوه بر انتقال مستقیم اطلاعات می‌تواند به طور ضمنی ناقل اطلاعاتی چون احساسات، طرز بیان خاص، شخصیت و سایر ویژگی‌های گوینده باشد. این اطلاعات جانبی بسیار ارزشمند هستند تا جایی که در برخی موارد حتی می‌توان با استفاده از آن به صحت و سقم اطلاعات اصلی نیز پی‌برد. تحلیل و بررسی علامت (سیگنال) گفتار مانند دیگر علامت‌های (سیگنال‌های) طبیعیِ انسان، موضوعی جذاب برای پژوهشگران است. این موضوع در ذیلِ عنوانِ کلی پردازش علامت و در شاخه پردازش گفتار دسته‌بندی می‌شود.

یکی از شاخه‌های پرکاربرد پردازش گفتار، تبدیل[1] گفتار است. به فرآیند تولید علامت گفتارِ یک گوینده (گوینده هدف)[2] از روی علامت گفتار یک گوینده دیگر (گوینده مبدا)[3] تبدیل گفتار می‌گویند. هدف تبدیل گفتار، تغییر اطلاعاتِ غیرِ زبان‌شناختی[4] جملات تلفظ شده، مانند هویت گوینده، در عینِ بی‌تغییر ماندن اطلاعات زبان‌شناختی است [۱]. به عبارت دیگر یک سامانه تبدیل گفتار باید بتواند بر پایه آموزش کافی و مناسب، نگاشت[5] بهینه‌ای بین گوینده مبدا و هدف ایجاد کند. در حین این تبدیل باید اطلاعات زبان‌شناختی که همان محتوای[6] گفتار است، حفظ شود و تا حد ممکن گفتار تولیدی با کیفیت باشد. کسانی که با گفتار گوینده هدف آشنایی دارند باید شباهت گفتار تبدیلی با گفتار واقعی گوینده هدف را تایید کنند.

* نویسنده پاسخگو: sayadian@aut.ac.ir

[1] Conversion

[2] Target
[3] Source
[4] Linguistic
[5] Map
[6] Context

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

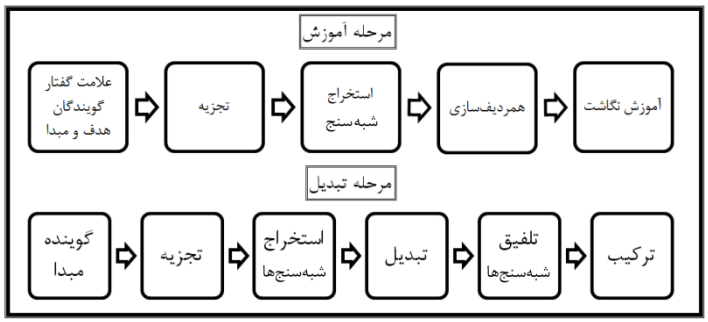
۱-۱. کاربردهای تبدیل گفتار

در مقالات و کتب مرتبط با موضوع پردازش گفتار کاربردهای بسیاری برای تبدیل گفتار برشمرده شده که مواردی از آن‌ها هنوز بالقوه مانده‌اند.

یکی از کاربردهای تبدیل گفتار تولید صدای افرادی که فوت کرده‌اند و یا افرادی که مشهور می‌باشند. به این شکل که در صورت وجود مقدار کافی از صدای ضبط شده از این افراد می‌توان با استفاده از روش‌های تبدیل گفتار به تولید صدای آن‌ها پرداخت. تبدیل گفتار می‌تواند در دوبله و صداگذاری مورد استفاده قرار گیرد [۲] و به اصطلاح دوبله ماشینی انجام شود. یکی از مهم‌ترین و در حقیقت اصلی‌ترین کاربردهای تبدیل گفتار، تغییر شخصیت گوینده در عین حفظ اطلاعات زبان‌شناختی است [۳]. از تبدیل گفتار می‌توان برای شخصیت بخشی به متون نیز استفاده کرد [۳]. به عنوان مثال متونی که در یک فیلم از زبان راوی داستان و نه شخصیت‌های اصلی آن خوانده می‌شوند می‌توانند توسط سامانه‌های تبدیل متن به گفتار و با کمک تبدیل گفتار تولید شوند. به این صورت کیفیت سامانه‌های تبدیل گفتار نیز افزایش می‌یابد [۴]. از دیگر کاربردهای تبدیل گفتار، آهنگین[1] کردن گفتار عادی و تبدیل آن به آواز[2] است [۵]. با کمک تبدیل گفتار می‌توان با تولید صدای شخصیت‌های بازی‌ها در حین بازی و براساس نظر کاربر به هوشمندتر شدن آن‌ها کمک کرد. تبدیل گفتار در درمان برخی بیماران نیز می‌تواند کاربرد داشته باشد. به عنوان مثال در تمرین برای تصحیح گفتار افرادی که مشکل تکلم دارند یا در آموزش تکلم صحیح به کودکان می‌تواند بکار گرفته شود. هم‌چنین می‌تواند در آموزش تلفظ صحیح کلمات به کسانی که می‌خواهند به صورت صحیح به یک زبان جدید صحبت کنند استفاده شود [۶]. از جالب‌ترین کاربردهای تبدیل گفتار می‌توان به ترجمه گفتاری از زبانی به زبان دیگر با حفظ پیام[3] گوینده اول اشاره کرد [۷]. از کاربردهای تخصصی تبدیل گفتار افزایش حجم داده‌های موجود در پایگاه داده‌های آموزشی است. این عمل براساس درونیابی بین داده‌های موجود در پایگاه داده صورت می‌گیرد [۸]. یکی دیگر از کاربردهای تخصصی تبدیل گفتار مقابله با اثر پدیده لومبارد در محیط نوفه‌ای (نویزی) است. به این صورت که برای مقابله با تغییر صدا در محیط نوفه‌ای گوینده را بهنجار (نرمالیزه) می‌کنند [۹]. از تبدیل گفتار می‌توان در افزایش کارایی سامانه‌های تلفن همراه نیز بهره برد. با استفاده از این ابزار می‌توان به افزایش بهره‌وری استفاده از پهنای باند دست یافت [۱۰].

۱-۲. طرح کلی تبدیل گفتار

تبدیل گفتار دارای دو مرحله اصلی است. مرحله اول مرحله آموزش است. در این مرحله براساس روش‌های مختلف و معمولاً به صورت برون‌خط سامانه تبدیل گفتار آموزش داده می‌شود. مرحله دوم مرحله تبدیل گفتار است. در این مرحله گفتاری از یک گوینده مبدا به گفتاری از گوینده هدف تبدیل می‌شود. در شکل ۱ طرح کلی این دو مرحله آمده است.



شکل ۱ مراحل تبدیل گفتار در یک سامانه تبدیل گفتار.

[1] Intonation
[2] Song
[3] Massage

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

بخش اصلی که عمل‌کرد سامانه تبدیل گفتار را از دیگر پردازش‌های انجام شده روی علامت گفتار متمایز می‌کند، جعبه تبدیل است. مهم‌ترین مراحل انجام شده در این قسمت تبدیل پوش طیف[1] و تبدیلات عروضی[2]است.

مهم‌ترین کار در تبدیل گفتار ایجاد ارتباط بین خصوصیات گوینده هدف و گوینده مبدا است. در این راستا شبه‌سنج‌های (پارامترهای) مختلفی مورد بررسی قرار گرفته‌اند که در [۱۱] به ترتیب بسامد گام، محل فُرمنت‌ها[3] و شکل پوش طیف به عنوان با اهمیت‌ترین شبه‌سنج‌ها در تعیین هویت گوینده برشمرده شده‌اند. البته در مواردی دیگر هم برخی این ترتیب را به شکل دیگری ارائه داده‌اند.

به‌عنوان مثال در یک مطالعه [۱۲] تأثیر پوش طیف نسبت به بسامد گام مهم‌تر ارزیابی می‌شود و به بررسی علامت (سیگنال) در حوزه زمان نیز پرداخته شده است. تحقیقی دیگر بسامد گام مرتبه دوم را با اهمیت دانسته [۱۳] و دیگری طیف متوسط گفتار را بسیار مهم می‌داند [۱۴]. به این سبب می‌توان نتیجه گرفت که با یک یا دو شبه‌سنج (پارامتر) نمی‌توان خصوصیات یک فرد که به منزله هویت اوست را نمایندگی کرد و البته اهمیت هر کدام از این شبه‌سنج‌ها (پارامترها) بسته به زبان و افراد مختلف می‌تواند متفاوت باشد. این مسأله در مطالعه دیگری [۱۵] نیز مورد بررسی و تأکید قرار گرفته است. البته بیش‌ترین جذابیت در شبه‌سنج‌ها (پارامترها) را در حال حاضر شبه‌سنج‌هایی چون طیف‌های محلی، بسامد گام و کانتور[4] بسامد گام به خود اختصاص داده‌اند.

۱-۳. آموزش

سامانه‌های تبدیل گفتار براساس نوع آموزشی که می‌بینند، تقسیم‌بندی می‌شوند. دو روش آموزش برای این سامانه‌ها مطرح است. روش اول که روش موازی[5] است به روشی گفته می‌شود که از جملات یکسان از گوینده مبدا و مقصد استفاده می‌کند. به همین دلیل به آن روش موازی گویند. اگر به هر دلیل امکان آموزش براساس جملات یکسان وجود نداشته باشد، از روش‌های غیرموازی[6] استفاده می‌شود. در ادامه این دو روش بررسی و مقایسه می‌شود.

۱-۳-۱. آموزش موازی

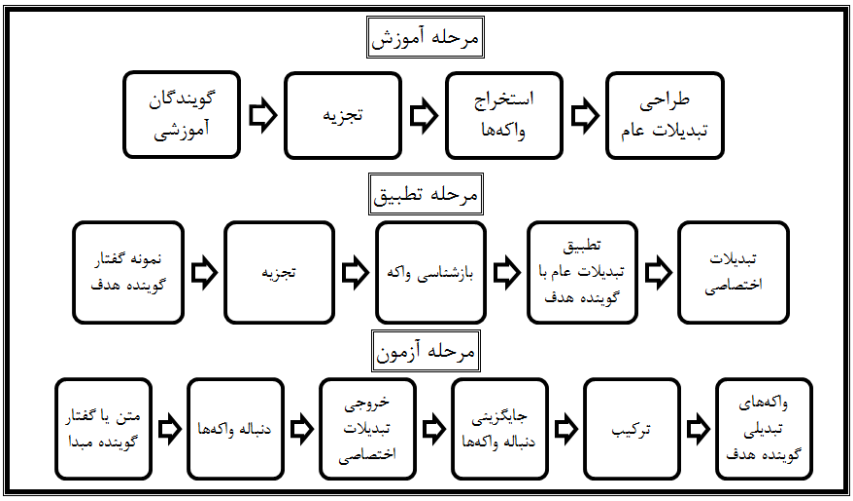
در این روش آموزش تعداد زیادی جملات مشابه از زوج گویندگان (بین ۵۰ تا ۱۰۰۰ جمله) به عنوان داده آموزشی مورد استفاده قرار می‌گیرد. همان‌طور که اشاره شد در آموزش موازی جملات آموزشی میان دو گوینده تا حد زیادی مشابه‌اند. به این معنا که علاوه بر متن حتی ممکن است در گویش و نوع بیان نیز مشابه باشند. این روش در مرحله آموزش ساده است اما در مرحله جمع‌آوری داده مشکل و گاهی غیر ممکن است. به عنوان مثال ممکن است گوینده هدف جمله دلخواه ما را نگوید یا اصلاً فوت شده باشد و نتوان هیچ جمله جدیدی از او داشت. در نتیجه ضعف بزرگ این روش نیاز به داشتن جملات خاصی از گویندگان در مرحله آموزش است [۱۶]. البته روش‌های موازی تبدیل گفتار براساس معیارهای کمی و کیفی موجود تا حد مناسبی توانستند کیفیت را در تبدیل گفتار حفظ کنند. اما مشکل اصلی این روش‌ها نیاز به جملات مشابه بین زوج گوینده‌ها در آموزش و بین گوینده مبدا و هدف است. تهیه چنین پایگاه داده‌ها (دادگان) و نمونه‌هایی مشکل و در برخی اوقات غیر ممکن است. برای حل این مساله روش‌های غیرموازی تبدیل گفتار ابداع شد و توسعه یافت.

۱-۳-۲. آموزش غیرموازی

در روش‌های غیرموازی از جملات متفاوتی از گویندگان در مرحله آموزش استفاده می‌شود و نیازی به تشابه جملات نیست. با توجه به اینکه در بسیاری از موارد امکان جمع‌آوری پایگاه داده‌های موازی برای مرحله آموزش وجود ندارد، روش‌های آموزش غیرموازی از جذابیت خاصی در بین پژوهشگران برخوردار هستند. در این صورت باید تبدیلات طوری طراحی شوند که وابسته به تشابه متن نباشند و مستقل از آن بتوانند دنباله مورد نظر از واج‌ها[7] را براساس خصوصیات گوینده هدف تولید کنند.

[1] Spectrum Envelope
[2] Prosody Transformations
[3] Formants
[4] Contour
[5] Parallel voice conversion
[6] Nonparallel training
[7] Syllable

[DOR: 20.1001.1.23455748.1393.2.2.6.1]



شکل ۲ مراحل تولید واکه گوینده هدف با طراحی ترادیس‌ها (تبدیلات).

در حقیقت آموزش غیرموازی به کاربرد نزدیک‌تر است و در عمل سامانه‌های مناسب‌تری برای بکارگیری در مواضع واقعی را دارد [۱۷]. این روش‌ها نسبت به روش‌های موازی نیاز به پنج تا ده برابر داده‌های آموزشی بیش‌تر دارند. این مسأله به عنوان نقطه ضعف روش‌های غیرموازی مطرح است.

در این تحقیق برای کاهش جملات مورد نیاز برای تبدیل واکه‌ها[1] روش جدیدی پیشنهاد می‌شود که در بخش بعد به آن پرداخته می‌شود.

۲. روش پیشنهادی تحقیق

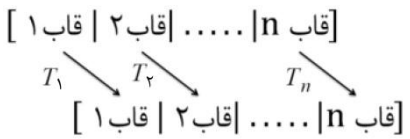
در علامت (سیگنال) گفتار هر گوینده، دو نوع اطلاعات می‌تواند وجود داشته باشد. دسته اول اطلاعات مربوط به دنباله هجاهای تشکیل‌دهنده گفتار شنیده شده بوده و دسته دوم اطلاعات مربوط به دستگاه تولید گفتار شخص گوینده می‌باشد. اطلاعات مربوط به دنباله[2] واج‌ها آن دسته اطلاعاتی است که به نحوه ادای یک واج در زبان مورد استفاده مربوط است و در گوینده‌های مختلف می‌تواند مشابه باشد. دسته دوم اطلاعات مرتبط با دستگاه تکلم گوینده است و در حقیقت هویت گفتاری گوینده را در بر دارد. در تبدیل گفتار، هدف اصلی تولید یا به عبارتی تقلید گفتار گوینده هدف در عین بی‌تغییر ماندن اطلاعات زبان‌شناختی گفتار است. در نتیجه تجزیه و تحلیل و در نهایت تبدیل دسته دوم اطلاعات مورد نظر است. اندیشه تحقیق از آن‌جا نشأت می‌گیرد که بیش‌ترین ویژگی‌های اختصاصی در گوینده به صورت انباشته شده در واکه‌ها موجود هستند. در نتیجه اگر بتوان به روشی واکه‌های گفتار گوینده هدف را تولید کرد، گامی ارزشمند در تولید گفتار وی به حساب خواهد آمد. همان‌طور که در قسمت‌های قبل اشاره شد راه حل مرسوم در تولید گفتار تهیه بانک اطلاعات نسبتاً بزرگی از گویندگان و آموزش سامانه تبدیل گفتار با آن بانک و تولید گفتار گوینده هدف به کمک آن است. در این تحقیق به عنوان یک روش نوین ترادیس‌های (تبدیلات)[3] مستقیمی بین واکه‌ها به صورت برون‌خط طراحی و تولید می‌شوند. با در اختیار داشتن این ترادیس‌ها (تبدیلات) می‌توان صرفاً با داشتن یک واکه از گوینده هدف سایر واکه‌های آن گوینده را تولید کرد و در نتیجه بخش مناسبی از هویت گوینده هدف را به گفتار خروجی اعمال کرد. چون در هر واج یک واکه و در هر کلمه حداقل یک واج موجود است پس با داشتن صرفاً یک کلمه از گوینده هدف یک واکه از آن گوینده در اختیار است. لذا با طراحی یک مجموعه تبدیلات مابین واکه‌ها می‌توان خصوصیات گوینده هدف را برای گفتار ترکیب شده نهایی تا حد مناسبی برآورده کرد. بر این اساس سعی می‌شود با طراحی این ترادیس‌ها تا جای ممکن خصوصیات گوینده هدف در گفتار ترکیب شده حفظ و اعمال شود. با توجه به این که در این روش

[1] Vowel
[2] Sequences
[3] Transformation

واکه‌های گوینده هدف به طور مستقیم تولید می‌شوند، لذا یک روش نوین برای بهبود تبدیل گفتار برشمرده می‌شود. در شکل ۲ نمودار مراحل تبدیل گفتار با طراحی تبدیلات از یک واکه به سایر واکه‌ها دیده می‌شود.

همان‌طور که در شکل ۲ نشان داده شده است این روش دارای سه مرحله آموزش، تطبیق و آزمون است. مرحله آموزش به صورت برون‌خط و با استفاده از تعدادی گوینده آموزشی انجام می‌شود. در مرحله آموزش، علامت (سیگنال) گفتار هر کدام از گویندگان آموزشی به صورت علامت (سیگنال) صوتی ۱۶ کیلوهرتز مونو ۱۶ بیتی[1] ذخیره می‌شود. این علامت‌ها با استفاده از روش استریت[2] که یک سامانه تجزیه و ترکیب با کیفیت محسوب می‌شود، تجزیه می‌شوند. از گفتار هر کدام از گویندگان آموزشی تمام واکه‌های زبان استخراج و در مجموعه‌های جداگانه ذخیره می‌شوند. برای این منظور جملاتی استفاده می‌شوند که حاوی تمام واکه‌های زبان به تعداد کافی باشند. سپس این جملات به صورت باسرپرستی[3] تقطیع شده و واکه‌ها جدا می‌شوند. تمام واکه‌های یکسان از تمام گوینده‌ها در یک مجموعه جمع می‌شوند. چون تنها تمام تلفظ‌های[4] مختلف یک واکه در یک مجموعه قرار می‌گیرند، تعداد مجموعه‌ها برابر با تعداد واکه‌های زبان خواهد بود. پس در نهایت هر مجموعه حاوی نمونه‌های ذخیره شده از یک واکه بوده و تعداد اعضای آن برابر با تعداد کل تلفظ‌های آن واکه توسط تمام گوینده‌ها و تعداد مجموعه‌ها نیز به تعداد واکه‌های زبان خواهد بود. لذا اگر تعداد واکه‌های زبان l بوده و K گوینده موجود باشند و هر کدام از گویندگان هر واکه را D بار تلفظ کنند، داده‌های آموزشی l مجموعه با $K \times D$ عضو خواهند بود. حال براساس این داده‌ها ترادیس‌ها (تبدیلات) از هر واکه به سایر واکه‌ها تولید خواهند شد. ترادیس‌ها $S_{r,i}$ یک واکه را به عنوان ورودی و واکه دیگر را به عنوان خروجی خواهند داشت که r اندیس مربوط به واکه ورودی و i در آن اندیس واکه خروجی است. چون تا این مرحله گوینده هدف تأثیری در طراحی ترادیس‌ها نداشته است.

این ترادیس‌ها در حقیقت نماینده میانگین زبانی[5] نمونه‌های گرفته شده از گویندگان آموزشی است. لذا به این ترادیس‌ها، ترادیس‌های عام[6] گفته می‌شود. این ترادیس‌ها بر هیچ یک از گویندگان تطبیق[7] داده نشده‌اند و صرفاً نمایشگر میانگین[8] آن‌ها هستند. نکته مهمی که باید به آن اشاره کرد ماتریسی بودن ترادیس‌ها است. به این معنا که هر ترادیس در واقع خود یک ماتریس ترادیس است. هر قاب[9] از واکه ورودی ترادیس متناسب با خود را خواهد داشت. به عبارت دیگر ترادیسی که اولین قاب واکه را به اولین قاب واکه هدف ترادیس می‌کند متفاوت از ترادیسی است که روی قاب بعدی اعمال می‌شود. لذا هر ترادیس، خود یک گروه ترادیس یا به عبارتی یک ماتریس ترادیس است. این مسأله در شکل ۴ آمده است.



شکل ۳ نمایش زیر ترادیس‌های (زیر تبدیلات) یک تبدیل واکه به واکه دیگر.

همان‌طور که در شکل فوق دیده می‌شود، ترادیس واکه به واکه جدید در حقیقت n زیر ترادیس (زیر تبدیل)[10] است. که n تعداد قاب‌های ورودی و خروجی است، لذا برای هر ترادیس S داریم:

$$S = \{T_1, T_2, \dots, T_n\} \quad (1)$$

در عمل می‌توان یک نمایش ماتریسی نیز از ترادیس‌های این مرحله داشت که در شکل ۴ است. همان‌طور که در شکل ۴ دیده می‌شود از روی هر کدام از واکه‌ها که در علامت (سیگنال) ورودی وجود داشته باشد به وسیله یک مجموعه ترادیس، مابقی واکه‌های مورد نیاز برای ترکیب گفتار ساخته می‌شوند. در حقیقت باید در مرحله آموزش تمام درایه‌های ماتریسی شکل که هر کدام خود یک ماتریس نگاشت غیرخطی هست، محاسبه شوند تا در زمان

[1] 16 bit mono
[2] STRAIGHT; Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
[3] Supervised
[4] Pronunciation

[5] Language mean
[6] General Transformations
[7] Modification
[8] Mean
[9] Frame
[10] Sub Transform

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

ورود علامت جدید با استفاده از این ترادیس‌ها (تبدیلات)، تمام واکه‌های مورد نیاز ساخته شوند:

$$\begin{array}{c} \\ V_1 \\ V_2 \\ \vdots \\ V_6 \end{array} \begin{array}{c} \begin{array}{cccc} V_1 & V_2 & \cdots & V_6 \end{array} \\ \begin{bmatrix} 1 & S_{1,2} & \cdots & S_{1,6} \\ S_{2,1} & 1 & & \vdots \\ \vdots & & \ddots & \\ S_{6,1} & \cdots & & 1 \end{bmatrix} \end{array}$$

شکل ۴ ماتریس ترادیس‌های مستقیم.

همان‌طور که قبلاً نیز اشاره شد، نکته مهم در مورد ماتریس فوق آن است که هر کدام از اعضای ماتریس در حقیقت خود یک ماتریس است (شکل ۲). این ماتریس ترادیس‌های قاب به قاب از ورودی به خروجی را در بر دارد. به عنوان مثال برای $S_{r,i}$ که یک عضو از ماتریس فوق است، داریم:

$$S_{r,i} = \begin{bmatrix} T_{1,1} v_r v_j & \cdots & T_{1,M} v_r v_j \\ \vdots & \ddots & \vdots \\ T_{M,1} v_r v_j & \cdots & T_{M,M} v_r v_j \end{bmatrix} \quad (۲)$$

که در آن M تعداد قاب‌های واکه ورودی است. از هر قاب از واکه ورودی با یک ترادیس خاص، یک قاب از واکه ترادیس خروجی ساخته می‌شود. پس از تولید خروجی با درون‌یابی طول آن به اندازه دلخواه ترادیس می‌شود.

در مرحله تطبیق، ترادیس‌های عام بر گوینده هدف تطبیق داده می‌شوند و یا به عبارتی این ترادیس‌ها برای گوینده هدف اختصاصی می‌شوند. این مرحله به صورت برخط و با در اختیار داشتن گفتار گوینده هدف صرفاً به اندازه یک کلمه انجام می‌شود. برای این کار، پس از تجزیه گفتار گوینده، هدف واکه‌های موجود در آن نمونه گفتار به صورت باسرپرستی استخراج می‌شوند. سپس با استفاده از این واکه یا واکه‌های استخراج شده، ترادیس‌ها برای گوینده هدف اختصاصی می‌شوند. با توجه به اینکه کدام واکه توسط گوینده هدف تلفظ شده باشد، دسته ترادیس‌های مربوط به آن واکه انتخاب می‌شوند. به این صورت که اگر واکه v که از گفتار گوینده هدف بدست آمده واکه h-ام زبان باشد، $S_{h,i}$- ها به عنوان ترادیس‌های نامزد برای تطبیق با گوینده هدف جدا می‌شوند. سپس از روی آن‌ها ترادیس‌های اختصاصی که با p_i نمایش داده می‌شوند، تولید می‌گردند. در حقیقت p_i همان ترادیس $S_{h,i}$ است که به گوینده هدف تطبیق داده شده و برای آن اختصاصی شده است. p- ها نیز مانند S- ها در واقع یک مجموعه ترادیس یا یک ماتریس ترادیس‌ها هستند.

در مرحله آزمون، ابتدا یک متن یا جمله از گوینده مبدا به عنوان ورودی به سامانه وارد می‌شود. سپس دنباله واکه‌های متناسب با این متن تعیین می‌شود. این دنباله $O = \{o_1, o_2, \dots, o_m\}$ نام دارد که در آن هر کدام از o_i- ها یک واکه و m تعداد واکه‌های گفتار ورودی از گوینده مبدا است. در حقیقت O حاوی عنوان واکه‌ها (نه خود آن‌ها) است و o_i- ها اعدادی بین ۱ تا ۶ هستند. این دنباله نشان‌دهنده این است که گوینده کدام واکه‌های زبان را به دنبال هم در گفتارش استفاده کرده است. حال باید از روی این دنباله، دنباله گفتاری اختصاصی گوینده هدف تولید شود. به این دنباله $K = \{k_1, k_2, \dots, k_m\}$ می‌گویند که در آن k_i واکه‌های تولید شده معادل با o_i است. برای تولید k_i ابتدا بررسی می‌شود که واکه ذخیره شده از گوینده هدف چه واکه‌ای است. سپس آن واکه به عنوان ورودی به ترادیس اختصاصی o_i-ام داده می‌شود. خروجی این ترادیس k_i است. توجه شود o_i صرفاً عنوان واکه‌ها را در بر دارد، اما k_i علامت (سیگنال) گفتاری واکه مستخرج از ترادیس اختصاصی گوینده هدف است. به عبارتی داریم:

$$k_i = P_{o_i}(v) \quad (۳)$$

که در آن P_{o_i} ترادیس اختصاصی واکه Oi و v واکه بریده شده از گفتار گوینده هدف است. بنابراین با جای‌گذاری در رابطه مربوط به K داریم:

$$K = \{P_{o_1}(v), P_{o_2}(v), \dots, P_{o_m}(v)\} \quad (۴)$$

پس هر واکه در دنباله به وسیله ترادیس مربوط به خودش، از واکه استخراج شده از گوینده هدف تولید می‌شود. به این ترتیب دنباله واکه‌های گوینده هدف، از واکه مستخرج از گفتارش با کمک ترادیس‌های اختصاصی متناسب با دنباله ورودی تولید می‌شوند. در نهایت این دنباله به جعبه ترکیب[1] فرستاده شده علامت (سیگنال) خروجی تولید می‌شود.

[1] Synthesis Box

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

۲-۱. طراحی ترادیس‌ها (تبدیلات)

با توجه به بکر بودن اندیشه تحقیق هیچ‌گونه سابقه‌ای از نحوه تولید ترادیس‌ها (تبدیلات) در دست نبود. لذا باید مبناهای مختلف برای تولید ترادیس‌های (تبدیلات) آزمایش و در نهایت بهترین مبنا برای الگوسازی انتخاب می‌شد. پس از بررسی مبناهایی چون روش‌های شبکه عصبی، الگوی مخفی مارکوف و الگوی مخلوط گوسی، توابع گوسی شرطی به عنوان مبنا برای طراحی ترادیس‌ها مورد استفاده قرار گرفت. علت این امر آن بود که شبکه عصبی به علت عدم توانایی در دنبال کردن توالی زمانی قاب‌ها، عملکرد مناسبی از خود نشان نداد. علاوه بر عدم توانایی کافی در دنبال کردن توالی زمانی، کم بودن داده‌های آموزشی به نسبت درجه آزادی یا به عبارتی تعداد شبه‌سنجها (پارامترها) نیز دلیل دیگری بر ضعف عملکرد این سامانه‌های بر این مبنا است. تعداد ضرایبی که نماینده هر قاب هستند، ۲۴ عدد در نظر گرفته می‌شوند. لذا باید در هر قاب ۲۴ ضریب توسط شبکه تخمین زده شوند که برای این کار حداقل ۵ برابر این تعداد ضرایب نمونه آموزشی نیاز است. با توجه به کمبود داده‌های آموزشی عملاً شبکه برای ترادیس این ضرایب به‌طور مناسب آموزش نمی‌بیند. با توجه به موارد فوق شبکه عصبی در کاربرد مورد نیاز این تحقیق کارایی لازم را نداشته و در عمل کنار گذاشته شد.

الگوی گوسی در مرحله آموزش، یعنی تولید الگوی میانگین زبان، دارای عملکرد قابل قبولی است. اما برخلاف مرحله آموزش، در مرحله تطبیق، دیگر الگوی گوسی توانایی تطبیق سامانه را با گوینده هدف با تعداد اندک نمونه از آن را ندارد. الگوی گوسی صرفاً یک میانگین از داده‌های موجود ارائه می‌دهد. لذا با تعداد محدودی داده از گوینده هدف امکان تطبیق سامانه با گوینده هدف وجود ندارد. زیرا با در دست بودن داده‌های اندک شبه‌سنجهای (پارامترهای) الگو را نمی‌توان به نحوی اصلاح کرد که بر گوینده هدف تطبیق یابند.

در بکارگیری الگوی مخفی مارکوف هم مشکلاتی وجود دارند. معمولاً استفاده از این الگوها در ترادیس کلمات و جملات و واحدهایی بزرگ‌تر از هجا کاربرد دارد. با توجه به ورودی ترادیس‌ها که تنها یک واکه است، این الگوها قادر به نمایندگی کردن ترادیس‌ها نیستند. الگوی مخلوط گوسی نیز دارای ضعف مشابهی است. معمولاً در استفاده از مخلوط گوسی به هر هجا یک الگوی گوسی نسبت داده می‌شود و بر این اساس ترادیس‌ها میان کلمات و جملات طراحی می‌شوند. در نتیجه در این‌جا نیز واحد فعالیت در سطح هجا خواهد بود. با توجه به کوچک‌تر از یک هجا بودن واحد مورد استفاده در تحقیق، عملاً این الگو نیز حذف می‌شود.

عملکرد مناسب تابع گوسی در مرحله آموزش، این تابع را به عنوان کاندید مناسبی برای الگوسازی معرفی می‌کند. البته باید عدم توانایی کافی در مرحله تطبیق نیز با تمهیداتی بر طرف شود. این نقطه ضعف با استفاده از توابع گوسی شرطی مرتفع می‌شود. تابع گوسی شرطی می‌تواند تابع تولیدی را بر نمونه‌ای دلخواه تطبیق داده و خروجی متناسب را تولید کند. در همین راستا و در ادامه به معرفی الگوی گوسی شرطی مورد استفاده پرداخته می‌شود.

۲-۲. الگوی ریاضی ترادیس‌ها (تبدیلات)

با توجه به مطالب قسمت قبل، توابع گوسی شرطی برای استفاده به عنوان مبنای طراحی ترادیس‌ها (تبدیلات) در این تحقیق مورد استفاده قرار گرفتند. فرض کنید X یک بردار گوسی D بعدی با توزیع گوسی $N(X|\mu,\Sigma)$ باشد که X به دو زیر بردار[1] X_a و X_b افراز شود. بدون هیچ تخصیصی می‌توان X_a را M عضو اول X و X_b را $D-M$ باقی‌مانده آن در نظر گرفت. چنان‌که داریم:

$$X=\begin{pmatrix}X_a\\X_b\end{pmatrix}$$

بر همین اساس بردار میانگین و ماتریس کوواریانس مربوطه نیز به صورت زیر خواهند بود:

$$\mu=\begin{pmatrix}\mu_a\\\mu_b\end{pmatrix},\qquad \Sigma=\begin{pmatrix}\Sigma_{aa}\,\Sigma_{ab}\\\Sigma_{ba}\,\Sigma_{bb}\end{pmatrix}$$

با توجه به تقارن ماتریس کوواریانس یعنی $\Sigma=\Sigma^T$ ماتریس‌های Σ_{aa} و Σ_{bb} نیز متقارن شده و هم‌چنین $\Sigma_{ba}={\Sigma_{ab}}^T$ می‌شود. اگر $\Lambda=\Sigma^{-1}$ باشد و داشته باشیم:

$$\Lambda=\begin{pmatrix}\Lambda_{aa}\,\Lambda_{ab}\\\Lambda_{ba}\,\Lambda_{bb}\end{pmatrix} \qquad (۵)$$

[1] Sub Vector

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

قطعاً Λ نیز متقارن خواهد بود، زیرا معکوس یک ماتریس متقارن، متقارن است. لذا به مانند Σ-ها ماتریس‌های Λ_{aa} و Λ_{bb} نیز متقارن شده و $\Lambda_{ba} = {\Lambda_{ab}}^T$ می‌شود. باید احتمال شرطی یا به عبارتی $p(X_a|X_b)$ را محاسبه کرد. برای X داریم:

$$N(X|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}}|\Sigma|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)\right] \quad (۶)$$

با توجه به رابطه ۵، نمای عبارت ۶ را می‌توان به شکل زیر بازنویسی کرد:

$$-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) = \quad (۷)$$
$$-\frac{1}{2}(X_a-\mu_a)^T\Lambda_{aa}(X_a-\mu_a) - \frac{1}{2}(X_a-\mu_a)^T\Lambda_{ab}(X_b-\mu_b)$$
$$-\frac{1}{2}(X_b-\mu_b)^T\Lambda_{ba}(X_a-\mu_a) - \frac{1}{2}(X_b-\mu_b)^T\Lambda_{bb}(X_b-\mu_b)$$

در عبارت بالا نسبت X_a به صورت مربعی[1] است و در نتیجه $p(X_a|X_b)$ گوسی خواهد شد. به این سبب اگر میانگین و واریانس آن مشخص شوند، توزیع آن کاملاً مشخص خواهد شد. برای محاسبه میانگین و واریانس توزیع، با توجه به این‌که Σ متقارن است، سمت چپ عبارت اخیر به‌صورت زیر بازنویسی می‌شود:

$$-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) = \quad (۸)$$
$$-\frac{1}{2}X^T\Sigma^{-1}X + X^T\Sigma^{-1}\mu + C$$

که در آن C به معنای جمله‌هایی از عبارت است که مستقل از X هستند. بخش اول عبارت سمت راست نسبت به X درجه دوم و بخش دوم عبارت سمت چپ نسبت به X درجه اول است. سمت چپ عبارات را متحد قرار می‌دهیم. باتوجه به اینکه متغیر مورد نظر ما X_a است مابقی شبه‌سنج‌ها (پارامترها) و متغیرها ثابت در نظر گرفته می‌شوند در نتیجه برای عبارات درجه دوم داریم:

$$-\frac{1}{2}X^T\Sigma^{-1}X = -\frac{1}{2}{X_a}^T\Lambda_{aa}X_a \quad (۹)$$

که در نتیجه:

$$\Sigma_{a|b} = {\Lambda_{aa}}^{-1} \quad (۱۰)$$

هم‌چنین برای عبارات درجه یک داریم:

$$X^T\Sigma^{-1}\mu = {X_a}^T\{\Lambda_{aa}\mu_a - \Lambda_{ab}(X_b-\mu_b)\} \quad (۱۱)$$

در نتیجه با توجه به اینکه ${\Lambda_{ab}} = {\Lambda_{ab}}^T$ داریم:

$${\Sigma_{a|b}}^{-1}\mu_{a|b} = \{\Lambda_{aa}\mu_a - \Lambda_{ab}(X_b-\mu_b)\} \quad (۱۲)$$

و در نتیجه:

$$\mu_{a|b} = \Sigma_{a|b}\{\Lambda_{aa}\mu_a - \Lambda_{ab}(X_b-\mu_b)\} = \mu_a - {\Lambda_{aa}}^{-1}\Lambda_{ab}(X_b-\mu_b) \quad (۱۳)$$

با توجه به عبارات فوق داریم:

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(X_b-\mu_b),$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ab} \quad (۱۴)$$

در کاربرد این تحقیق اگر واکه ورودی را V_a و مجموعه آموزشی متناسب با آن را a و واکه خروجی را V_b و مجموعه آموزشی متناسب با آن را b بنامیم و $V_b = X_a$ و $V_a = X_b$ باشد، داریم[2]:

$$P(V_b|V_a) = N(V_b|\mu_{b|a},\Sigma_{b|a}) \quad (۱۵)$$

که در آن:

$$\mu_{b|a} = \mu_b + \Sigma_{ba}\Sigma_{aa}^{-1}(V_a-\mu_a),$$
$$\Sigma_{b|a} = \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}$$

بر این اساس V_b طوری انتخاب می‌شود که تابع احتمال را حداکثر کند. یعنی:

$$max[P(V_b|V_a)] = \max_{V_b}[N(V_b|\mu_{b|a},\Sigma_{b|a})] \quad (۱۶)$$

به این ترتیب V_b که خروجی ترادیس (تبدیل) است، براساس ترادیس گوسی شرطی و با کمک رابطه ۱۶ مشخص می‌شود.

۳. نتایج و بحث

۳-۱. پیاده‌سازی ترادیس‌ها (تبدیلات)

با توجه به مطالبی که در قسمت‌های قبل آمد ۱۰ گوینده برای مرحله آموزش ترادیس‌ها (تبدیلات) وارد مطالعه شدند. براساس ۶ واکه زبان فارسی، از هر گوینده یک نمونه تلفظ شده از هر واکه به صورت علامت (سیگنال) ۱۶ کیلوهرتز مونو ۱۶ بیتی ذخیره شد. به عبارت دقیق‌تر از گفتار هر یک از این گوینده‌ها شش واکه {َ ِ ُ آ ای او} به طور دستی بریده شد. با توجه به اینکه طول نمونه‌ها

[1] Quadratic

[2] منظور از V_a و V_b شبه‌سنج‌هایی است که از طیف استخراج شده‌اند که در این مطالعه ضرایب مل کپستروم می‌باشند.

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

متفاوت است، با درون‌یابی، تعداد قاب‌های همه ورودی‌ها به مقدار یکسان تغییر یافت. این مقدار یکسان برابر با میانگین طول نمونه‌ها به‌علاوه انحراف معیار آن در نظر گرفته شد که برای نمونه‌های ثبت شده برای این تحقیق برابر ۲۰ است. طول هر قاب ۵ میلی‌ثانیه در نظر گرفته شد که در نتیجه طول نمونه‌ها ۱۰۰ میلی‌ثانیه بدست آمد. بر این اساس برای هر واکه یک مجموعه متشکل از ۱۰ تلفظ از آن واکه توسط ۱۰ گوینده متفاوت ذخیره گردید. لذا در کل، ۶۰ نمونه از گویندگان ذخیره شد. این نمونه‌ها براساس نوع واکه به شش مجموعه تقسیم شدند. هر مجموعه دارای ۱۰ نمونه شد. به این ترتیب در هر مجموعه ۱۰ تلفظ از یک واکه قرار گرفت. به عبارتی مجموعه اول حاوی ۱۰ تلفظ از { ً } و به ترتیب هر مجموعه شامل ۱۰ تلفظ از یکی دیگر از واکه‌ها بود. پس شش مجموعه ۱۰ عضوی با اعضایی به طول ۱۰۰ میلی‌ثانیه بدست آمد. به این شکل هر عضو از یک مجموعه با دیگر عضوهای آن مجموعه از نظر معادل زبان‌شناختی مشابه بود. از سوی دیگر هر گوینده یکی از اعضای هر مجموعه را تلفظ کرده بود. لذا هر یک از اعضای مجموعه با یکی از اعضای هر یک از مجموعه‌های دیگر از نظر داشتن گوینده یکسان نیز متناسب شد. زیرا یک گوینده آن‌ها را تلفظ کرده بود. به عنوان مثال از یک سو واکه { ً } تلفظ شده توسط گوینده اول، با نه تلفظ این واکه توسط دیگر گویندگان به عنوان هم مجموعه‌ای‌اش متناسب است و از سوی دیگر و به علت تلفظ شدن توسط یک گوینده به شکلی دیگر با واکه { ِ } تلفظ شده توسط گوینده اول که در مجموعه دیگر قرار دارد، متناسب است. لذا داده‌های شش مجموعه هم به صورت طولی[1] (با هم‌مجموعه‌ای‌ها) و هم به صورت عرضی (با تلفظ دیگر واکه‌ها توسط همان گوینده) با یکدیگر در ارتباط هستند.

با توجه به مطالب فوق باید ترادیسی (تبدیلی) طراحی شود که یک واکه را به عنوان ورودی دریافت و واکه دیگر به عنوان خروجی تحویل دهد. این ترادیس باید تا جای ممکن خصوصیات گوینده مبدا را به واکه خروجی منتقل کند یا به عبارتی به گوینده هدف تطبیق یابد. با توجه به مطالب فوق باید بر هریک از شش مجموعه یک تابع گوسی جداگانه منطبق شود تا تناسب طولی داده‌ها رعایت شود. برای رعایت تناسب عرضی نیز این توابع گوسی باید به هم مشروط شوند. الگوی گوسی شرطی مطرح شده در بخش قبل توانایی تطبیق به چنین شرایطی را دارد، لذا برای تولید ترادیس مورد استفاده قرار گرفت. در عمل براساس رابطه ۱۶ و با استفاده از بانک داده‌های ذخیره شده با توصیفات فوق، ترادیس‌ها (تبدیلات) طراحی شدند و آموزش دیدند.

۳-۲. آزمون

برای سنجش توانمندی ترادیس‌های تولید شده به روش پیشنهادی در مقاله، آزمونی به شرح زیر طراحی شد. در این آزمون از بانک داده‌ها توضیح داده شده در قسمت قبل استفاده گردید. این بانک داده‌ها در آزمایشگاه تحقیقاتی پردازش اطلاعات واقع در دانشکده برق دانشگاه صنعتی امیرکبیر ایران ثبت و ضبط شد. داده‌ها با یک دستگاه گوشی مجهز به میکروفون[2] ثبت و به وسیله یک دستگاه رایانه قابل حمل ذخیره شدند. هدف از آزمون سنجش میزان موفقیت ترادیس‌ها (تبدیلات) در تبدیل یک واکه به واکه جدید است. میزان شباهت یک واکه تولیدی به واکه واقعی را می‌توان براساس معیارهای کمی، کیفی و یا کمی و کیفی به صورت توأم سنجید. در این آزمون از یک معیار کمی یعنی فاصله[3] دُش‌پیچی (اعوجاج)[4] و یک معیار کیفی یعنی ترسیم طیف‌نگاشت (اسپکتوگرام)[5] بهره گرفته شده است. معیار فاصله دُش‌پیچی (اعوجاج) به معنای میزان عدم شباهت ویژگی‌های مستخرج از پوش طیف علامت (سیگنال) تولیدی و واقعی است. در حقیقت هر چه فاصله دُش‌پیچی (اعوجاج) بیش‌تر باشد، دو علامت (سیگنال) عدم شباهت بیش‌تری دارند یا به عبارتی هرچه این فاصله کم‌تر باشد، ترادیس بهتر عمل کرده است. روش محاسبه فاصله دُش‌پیچی (اعوجاج) بین دو قاب متناظر از دو علامت در رابطه ۱۷ آمده است. این عبارت فاصله دُش‌پیچی (اعوجاج) بین دو قاب را می‌سنجد.

$$\mathrm{DD}\,(V_a, V_b) = \frac{\mathrm{norm}(V_b - V_a)}{\mathrm{norm}(V_b)} \qquad (۱۷)$$

[1] Longitudinal

[2] A4TECH HS-50

[3] Distance

[4] Distorsion

[5] Spectrogram

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

در عبارت فوق DD فاصله دُش‌پیچی V_b بردار ضرایب پوش طیف واکه تولید شده و V_a بردار ضرایب پوش طیف واکه واقعی در یک قاب هستند. با توجه به اینکه در این آزمون از ضرایب امافسی‌سی[1] برای بازنمایی پوش طیف استفاده شد، بردارهای V_a و V_b به ترتیب حاوی ضرایب امافسی‌سی یک قاب از واکه واقعی و قاب متناظر آن در واکه تولیدی هستند. لذا فاصله کلی بین یک واکه تولیدی و واکه واقعی برابر است با میانگین فواصل قاب‌های متناظر آن‌ها که در رابطه ۱۸ آمده است.

$$TD = \sum_{1}^{n} DD(V_a, V_b)/n \qquad (۱۸)$$

n در عبارت فوق برابر با تعداد قاب‌های واکه است و TD فاصله کلی دُش‌پیچی (دُش‌پیچی تام) می‌باشد.

در عمل برای اجرای آزمون از روش بیرون نگه داشتن یک عضو[2] و نگه داشتن بقیه استفاده گردیده است. به این ترتیب که هر بار یک گوینده برای آزمون خارج می‌شد و تحت آموزش قرار نمی‌گرفت، سامانه با ۹ گوینده باقی‌مانده آموزش داده می‌شد. سپس علامت (سیگنال) گوینده آزمون (بیرون نگه‌داشته شده) به عنوان ورودی به ترادیس داده می‌شد و شبه‌سنج‌های (پارامترهای) ترادیس با استفاده از علامت‌های (سیگنال‌های) نُه گوینده باقی‌مانده تنظیم می‌شدند. سپس خروجی ترادیس به عنوان واکه تولیدی ثبت و فاصله مطرح شده در رابطه ۸ مابین این واکه تولیدی و واکه واقعی متناظر محاسبه و ذخیره می‌شد. لذا هر آزمایش به ازای ۱۰ گوینده، ۱۰ بار تکرار می‌شد و میانگین فاصله این ۱۰ آزمایش به عنوان نتیجه آزمون ثبت می‌گردید. همان‌طور که اشاره شد برای نمایش پوش طیف از ضرایب امافسی‌سی به تعداد ۲۴ عدد به علاوه یک ضریب انرژی برای هر قاب استفاده شد که در حقیقت ترادیس‌ها (تبدیلات) بر روی این ضرایب اعمال گشتند. برای ترادیس کانتور گام نیز از ترادیس لگاریتمی استفاده شد.

جدول ۱ حاوی نتایج بدست آمده از آزمون به تفکیک ورودی و خروجی است. ستون سمت راست نمایان‌گر واکه ورودی و سطر بالایی نمایان‌گر واکه خروجی هستند. در حقیقت واکه مندرج در ستون به عنوان ورودی به ترادیس مربوطه داده شده و خروجی آن که واکه مندرج در سطر است، تولید و ثبت می‌شود. سپس با روش آزمون ذکر شده، میانگین فواصل دُش‌پیچی (اعوجاج) کلی در ۱۰ مرحله آزمون محاسبه می‌شود که در جدول زیر نمایش داده شده است. به عنوان مثال میانگین فاصله دُش‌پیچی (اعوجاج) در تولید واکه {ِ} از روی واکه {َ} با استفاده از ۱۰ گوینده و با استفاده از ترادیس طراحی شده برابر ۰٫۵۴ است. این میزان در تولید {ُ} از روی همان واکه برابر ۰٫۳۷ است. این بدان معناست که ترادیس طراحی شده برای تولید {ُ} از {َ} عملکرد بهتری نسبت به ترادیس طراحی شده برای تولید {ِ} از {َ} دارد. میانگین مقادیر

جدول ۱ میزان فاصله دُش‌پیچی برای ترادیس‌ها. ستون سمت راست واکه‌های ورودی و سطر بالایی واکه‌های خروجی.

ورودی / خروجی	{َ}	{ِ}	{ُ}	{آ}	{ای}	{او}
{َ}	۰	۰٫۵۴	۰٫۳۷	۰٫۴۳	۰٫۳۹	۰٫۳۲
{ِ}	۰٫۶۲	۰	۰٫۳۶	۰٫۴۳	۰٫۳۵	۰٫۲۸
{ُ}	۰٫۶۳	۰٫۵۱	۰	۰٫۴۰	۰٫۳۷	۰٫۲۹
{آ}	۰٫۶۰	۰٫۵۳	۰٫۳۳	۰	۰٫۳۶	۰٫۲۹
{ای}	۰٫۷۴	۰٫۵۸	۰٫۳۹	۰٫۴۴	۰	۰٫۳۳
{او}	۰٫۶۶	۰٫۵۳	۰٫۳۵	۰٫۴۳	۰٫۳۵	۰

جدول ۱ یا به عبارتی میزان کلی فاصله دُش‌پیچی (اعوجاج) برای کل ترادیس‌ها ۰٫۴۴۵۹ است.

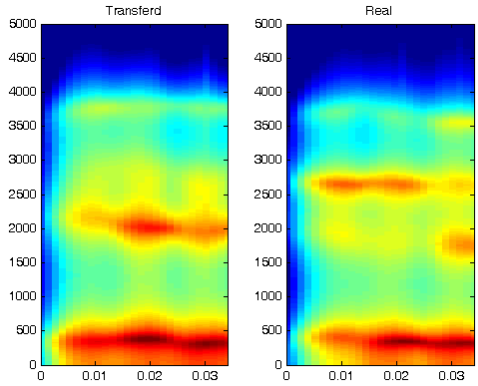
علاوه بر معیار کمّی فاصله دُش‌پیچی (اعوجاج)، معیار کیفی ترسیم طیف‌نگاشت (اسپکتوگرام) نیز مورد استفاده قرار گرفت. برای استفاده از این معیار، با استفاده از روش‌های ترسیمی طیف‌نگاشت (اسپکتوگرام) واکه تولیدی و واقعی برای مقایسه دیداری در کنار هم رسم می‌شوند. در شکل ۵ یک نمونه از مقایسه طیف‌نگاشت (اسپکتوگرام) علامت (سیگنال) واقعی و تولید شده واکه {ِ} را از واکه {َ} از گوینده سوم نمایش داده شده است. در این شکل طیف‌نگاشت (اسپکتوگرام) واکه تولیدی سمت چپ و طیف‌نگاشت (اسپکتوگرام) مربوط به واکه واقعی در سمت راست آمده‌اند. میزان فاصله دُش‌پیچی

[1] MFCC; Mel Frequency Cepstrum Coefficients
[2] One Leave Out

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

(اعوجاج) مربوط به علامت‌های (سیگنال‌های) مربوط به این دو طیف‌نگاشت (اسپکتوگرام) برابر ۰/۳۳۴۷ می‌باشد.

با توجه به نتایج فوق دیده می‌شود که صرفاً با استفاده یک کلمه به عنوان داده آموزشی از گوینده هدف نتایج قابل قبولی بدست آمده‌اند. از جدیدترین تحقیقات در موضوع تبدیل گفتار که با هدف کاهش میزان داده‌های آموزشی است، به مرجع [۱۸] و [۱۹] می‌توان اشاره کرد. مرجع [۱۸] یکی از دستاوردهایش را استفاده از داده محدود عنوان کرده و تعداد جملات آموزشی مورد استفاده‌اش را ۵ الی ۱۰۰ جمله اعلام کرده است. هم‌چنین مرجع [۱۹] ادعا کرده که جملات آموزشی را به حداقل ۱۰ جمله محدود کرده است. لذا تحقیق حاضر که جملات آموزشی را به کمتر از یک جمله - یک کلمه تک هجایی- محدود کرده توفیق بیشتری در زمینه کاهش جملات آموزشی داشته است.



شکل ۵ طیف‌نگاشت (اسپکتوگرام) واکه {ِ} واقعی و تولیدشده از واکه {ُ} مربوط به گوینده سوم.

۴. نتیجه‌گیری

در این تحقیق یک روش نوین برای تولید تمام واکه‌های گوینده هدف صرفاً با داشتن یک کلمه از آن گوینده ارائه شد. در مقایسه با سایر روش‌های موجود که حداقل به چند جمله آموزشی برای تبدیل گفتار نیاز دارند، این روش با در اختیار داشتن حتی بخشی از یک جمله یا به عبارتی یک کلمه تک هجایی[1] به عنوان علامت (سیگنال) آموزشی از گوینده هدف می‌تواند تبدیل گفتار را آغاز کند. در این روش با طراحی یک مجموعه ترادیس‌ها (تبدیلات) بر مبنای توابع گوسی شرطی، به صورت مستقیم، واکه‌های گوینده هدف از واکه مستخرج از نمونه گفتاری آن گوینده تولید می‌شوند. با توجه به این‌که هر هجا دارای یک واکه است، حداقل داده‌های آموزشی مورد نیاز برای ترادیس گفتار با روش پیشنهادی این مقاله به یک کلمه تقلیل می‌یابد. از سوی دیگر برای سنجش میزان عملکرد این ترادیس‌ها یک بانک داده‌ها از ۱۰ گوینده ذخیره و ثبت شد. با طراحی یک آزمون و با استفاده از این بانک داده‌ها به سنجش عملکرد ترادیس‌های طراحی شده پرداخته شد. در این آزمون از معیارهای سنجش عملکرد کمی و کیفی استفاده شد. شبیه‌سازی‌ها نشان دادند که نتایج ترادیس‌ها هم با توجه به معیارهای کمّی و هم با توجه به معیارهای کیفی قابل قبول هستند. با توجه به کاهش چشمگیر تعداد جملات آموزشی مورد نیاز از گوینده هدف به یک کلمه، روش ارائه شده در تحقیق حاضر یا به عبارتی تبدیل مستقیم واکه‌ها می‌تواند به عنوان سر منشأ یک تحول در عرصه تبدیل گفتار نقش‌آفرینی کند.

[1] One Syllable

۵. فهرست منابع

[1] E. Moulines, Y. Sagisaka, C. Sorin, "Voice conversion: state of the art and perspectives," Speech Communication, vol. 16, no. 2, pp. 125-126, 1995.

[2] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, S. Narayanan, "Text-independent voice conversion based on unit selection," Acoustics, Speech and Signal Processing, ICASSP 2006 Proceedings, 2006 IEEE International Conference on, May 2006.

[3] K. Lee, "Statistical approach for voice personality transformation," IEEE Transactions on Audio, Speech, Language Processing, vol. 15, no. 2, February 2007

[4] X. Huang, A. Acero, H. Hon, "Spoken Language Processing: A Guide to Theory, Algorithm and System Development," Prentice Hall, 2001.

[5] K. Saino, H. Zen, Y. Nankaku, A. Lee, K. Tokuda, "An HMM-based singing voice synthesis system," ICSLP, INTERSPEECH, Pittsburgh, Pennsylvania, ICSLP, pp. 17-21, 2006.

[6] K. Nakamura, T. Toda, H. Saruatri, K. Shikano, "Evaluation of extremely small sound source signals used in speaking-aid system with statistical voice conversion," IEICE Transaction Information and Systems, vol. E93-D, no.7, 2010.

[7] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, T. Dutoit, "Cross-Language Voice Conversion Based on Eigenvoices," Interspeech Brighton UK, 2009.

[8] E. Eide, M. Picheny, "Towards pooled-speaker concatenative text-to-speech," Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toulouse, France, 2006.

[9] H. Boril, P. Fousek, D. Sundermann, P. Cerva and J. Zdansky, "Lombard Speech Recognition: A Comparative Study," in In Proc. of the Czech-German Workshop, Prague, Czech Republic, 2006.

[10] W. Fujitsuru, H. Sekimoto, T. Toda, H. Saruwatari, K. Shikano, "Bandwidth extension of cellular phone speech based on maximum likelihood estimation with GMM," 2008 RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP'08), Gold Coast, Australia, March 2008.

[11] H. Matsumoto, S. Hiki, T. Sone, T. Nimura, "Multidimensional representation of personal quality of vowels and its acoustical correlates," IEEE Transactions on Audio and Electroacoustics, vol. 21, no. 5, pp. 428-436, 1973.

[12] K. Itoh, S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," IEICE Transactions, vol. J56-A, pp. 101-108, 1982.

[13] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," Speech Communication, vol. 5, no. 2, pp. 183-197, 1986.

[14] H. Sato, "Acoustic cues of female voice quality," Electronics and Communication in Japan, vol. 57-A, pp. 29-38, 1974.

[15] H. Kuwabara, Y. Sagisaka, "Acoustic characteristics of speaker individuality: control and conversion," Speech Communication, vol. 16, no. 2, pp. 165-173, 1995.

[16] D. Sündermann, H. Höge, A. Bonafonte, H. Duxans, "Residual prediction," In Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, pp. 512-516, 2005.

[17] A. Mouchtaris, J. Van der Spiegel, P. Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," Acoustics, Speech, and Signal Processing, Proceedings, (ICASSP '04). IEEE International Conference, May 2004.

[18] N. Xu, Y. Tang, J. Bao, A. Jiang, X. Liu, Z. Yang, "Voice conversion based on Gaussian processes by coherent and asymmetric training with limited training data," Speech Communication, vol. 58 , pp. 124–138, 2014.

[19] M. Ghorbandoost, A. Sayadiyan, M. Ahangar, H. Sheikhzadeh, A. Sabzi, J. Amini, "Voice conversion based on feature combination with limited training data," Speech Communication," vol. 67, no. 3, pp. 113-128 2015.

Journal of Acoustical Engineering Society of Iran, Vol. 2, No. 2, 2015

Construction of all persian language vowels only with one word of target speaker

M.J. Jannati, A. Sayadyan*

Dept. of Electrical Eng., Amirkabir Univ. of Tech.

Abstract

One of the most serious bottlenecks of voice conversion is the need of high number of training sentences from target speaker. Parallel voice conversion methods need 50 to 200 training sentences and nonparallel conversion methods need several times of it. In this research, a new method for construction of all Persian language vowels only with one training word of target speaker is introduced. In each word there is at least one syllable and in each syllable there is one vowel. Basic idea of research is designing direct transformations between each vowel and other vowels. Based on this idea, by offline training of sets of transformations, it is possible to construct all vowels from one arbitrary vowel. In Persian language there are 6 vowels. Because of it, 30 of such mentioned transformations must be designed. These transformations were designed with Gaussian conditional functions and trained by 10 speakers in supervised manner. With distortion distance criterion, average distance between real and artificial vowels was 0.4459.

Keywords: Vowel Conversion, Voice Conversion, Nonparallel Training

pp. 41-52 (In Persian)

[DOR: 20.1001.1.23455748.1393.2.2.6.1]

* Corresponding author E-mail: sayadian@aut.ac.ir