مجلهٔ فنی و مهندسی مدرس
شمارهٔ ۳٤، زمستان ۱۳۸۷
«ویژه‌نامهٔ مهندسی برق»

بازشناسی گفتار احساسی و شناسایی حالت گفتار در زبان فارسی

داود غرویان[۱*]، سید محمد احدی[۲]

۱- دانشیار مهندسی برق، دانشگاه صنعت آب و برق ایران

۲- استادیار مهندسی برق، دانشگاه صنعتی امیرکبیر

***تهران، صندوق پستی ۱۷۱۹-۱٦۷٦٥**

gharavian@pwut.ac.ir

(دریافت مقاله: بهمن ۱۳۸٤، پذیرش مقاله: مهر ۱۳۸۷)

چکیده- حالت گفتار سبب افزودن اطلاعات اضافی نسبت به اطلاعات نوشتاری می‌شود. از طرف دیگر، وجود حالت در گفتار سبب بروز مشکل در فرایند بازشناسی گفتار می‌شود. در تحقیقات قبلی نشان داده شد که حالت گفتار سبب تغییرات اساسی در پارامترهای گفتاری می‌شود.

برای بهبود نتایج بازشناسی گفتار با حالت، ابتدا باید تأثیر آن را بر پارامترهای گفتاری به‌دست آورد و در مرحلهٔ بعدی، از پارامترهای مناسبی برای بهبود نتایج بازشناسی استفاده کرد. در این تحقیق با توجه به نتایج به‌دست آمده در زمینهٔ تأثیر حالتهای گفتاری خشم و اندوه بر پارامترهای گفتاری نظیر فرمنت‌ها و فرکانس گام در زبان فارسی، بهبود نتایج بازشناسی گفتار با حالت، با مدل‌های عادی مد نظر است. با توجه به تغییرات منظم پارامترهایی نظیر فرکانس گام، فرمنت‌ها و شیب آنها با حالت گفتار، این پارامترها به بردار ویژگی سیستم بازشناسی اضافه می‌شوند. این کار سبب بهبود نتایج بازشناسی می‌شود. میزان این بهبود بستگی به نوع پارامتر، تعداد مخلوط‌ها و حالت گفتار دارد. با توجه به اهمیت شناسایی حالت گفتار و همچنین نقش آن در بهبود کیفیت سیستم بازشناسی گفتار، با استفاده از فرمنت‌ها و فرکانس گام به‌عنوان ویژگیهای ورودی و به‌کارگیری روشهای درخت تصمیم‌گیری و GMM، کار شناسایی حالت گفتار نیز انجام شده است.

کلید واژگان: نوا، فرکانس گام، حالت گفتار، بازشناسی گفتار، شناسایی حالت گفتار.

۱- مقدمه

یکی از ویژگی‌های مهم گفتار، انتقال حالات درونی فرد به شنونده می‌باشد. وقتی گفتاری توسط گوینده بیان می‌شود، این گفتار حاوی حالت شخص نیز هست. شناخت حالت گفتار، اطلاعات بیشتری را علاوه بر معنای لغوی گفتار برای شنونده مشخص می‌کند. لذا انتظار شنونده از گوینده، تنها آنچه گفته می‌شود نیست، بلکه چگونگی بیان آن نیز مهم است.

حالت گفتار اگر چه برای فهم گفتار توسط شنونده لازم است، اما در سیستم‌های بازشناسی، به‌دلیل تغییرات گسترده

در پارامترهای گفتار، می‌تواند سبب مشکلات زیادی نیز بشود. لذا ایجاد سیستم بازشناسی گفتار با حالت، و به‌تبع آن شناسایی حالت گفتار حائز اهمیت است.

حالت گفتار یکی از خواص کلان‌نوایی[1] است که برای درک آن باید بررسی‌ها در طول بیش از یک فریم گفتار صورت گیرد. به‌عبارت دیگر، حالت گفتار بر روی چند فریم تأثیرگذار است [۱، ۲].

کاربرد خواص نوایی عموماً در شناسایی انواع مختلف جمله، آموزش HMM[2] با گفتار آهنگین، پیدا کردن آهنگ در جمله و پیدا کردن مرزها است (HMM ابزاری آماری است که کاربردهای متفاوتی در پردازش گفتار دارد). اگر چه ممکن است بدون استفاده از خواص نوایی نیز بتوان چنین کارهایی را انجام داد، اما مشخص است که خواص نوایی حاوی اطلاعات جدیدی از گفتار است و در جملات آهنگین، عدم استفاده از آنها سبب کاهش دقت خواهد شد [۳].

شناسایی حالت گفتار نقش مهمی در افزایش کارایی بازشناسی دارد. با شناسایی نوع آهنگ در گفتار، می‌توان مدل زبانی متناسب با نوع آهنگ جمله را به‌کار برد. نکتهٔ قابل توجه این است که بر اساس آهنگ جمله مجموعهٔ کلماتی که بیشتر مورد استفاده قرار می‌گیرند، متفاوت هستند [٤].

شبکهٔ عصبی با داده‌های ورودی از چند فریم مجاور می‌تواند برای شناسایی حالت گفتار به‌کار رود. شبکه عصبی مورد نظر شامل چند دسته ورودی است که این ورودیها پارامترهای آکوستیکی و نوایی از چند فریم است. در خروجی شبکهٔ عصبی نیز به تعداد حالتهای گفتار، خروجی وجود دارد [۵] – [۸]. همچنین می‌توان از درخت تصمیم‌گیری[3] برای پیدا کردن حالت گفتار در جمله استفاده کرد [۹] – [۱۱]. یکی دیگر از روشهای شناسایی حالت گفتار استفاده از HMM است [۱۲] – [۱۵]. همچنین در برخی از تحقیقات از روشهای کلاس‌بندی مانند KNN[4] استفاده شده است [۱۵، ١٦].

در تحقیقی در زبان چینی نشان داده شده که پارامترهای نوایی در چهار حالت گفتاری ترس، عصبانیت، لذت و غم و ناراحتی و حالت عادی با هم متفاوت بوده و لذا این پارامترها می‌توانند برای شناسایی این حالات به‌کار روند [۱۷، ۱۸]. در تحقیقی دیگر با استفاده از بردار ویژگی متشکل از ضرایب MFCC[5] و فرکانس گام، فرکانس‌های فرمنت و انرژی همراه با پارامترهای سرعت و شتاب به بازشناسی حالت گفتاری با HMM پرداخته شده است (پارامترهای سرعت و شتاب به‌نوعی مشتق اول و دوم ضرایب MFCC است) [۱۹]. استفاده از ابزار SVM[6] همراه با ویژگی‌هایی در سطح واج، هجا و کلمه یکی دیگر از ابزارهای شناسایی حالت گفتار است [۲۰، ۲۱]متداول‌ترین پارامترهای ورودی سیستم‌های بازشناسی حالت گفتار، معمولاً ضرایب MFCC، فرکانس گام و انرژی است [۲۲] – [۲۸]. به‌عنوان مثال در تحقیقی با استفاده از همین ویژگیها در زبان سوئدی و با ابزار GMM به شناسایی جملات با حالت مثبت، منفی و عادی پرداخته شده است [۲۹]. در تحقیقی دیگر همچنین از GMM برای شناسایی حالت خوشحالی در زبان هلندی استفاده شده است [۲۷]. در تحقیقی دیگر در زبان اسپانیایی از GMM برای شناسایی حالات گفتاری خشم، تعجب، ترس، اضطراب، لذت و غم استفاده شده است [۳۰]. همچنین در تحقیقی دیگر برای شناسایی حالت گفتار در جملات مختلف از ابزار K-NN[7] استفاده شده است [۲۳].

بررسی‌ها نشان داد که در مورد بازشناسی گفتار با حالت کار زیادی انجام نشده است. اما می‌توان به‌عنوان

1. Macro-Prosodic
2. Hidden Markov Model
3. Decision Tre

4. K-Nearest Neighbour
5. Mel Frequency Cepstral Coefficients
6. Support Vector Machines
7. K-Nearest Neighborhood

نمونه به تحقیقی در زبان انگلیسی با استفاده از دادگان BNC[1] اشاره کرد. در این تحقیق با بررسی جملات دارای حالت، اطلاعات اضافی مورد نیاز به دادگان افزوده شده است. همچنین از مدل زبان متناسب با گفتار با حالت استفاده شده است. مجموعهٔ این تدابیر سبب افزایش نسبی دقت بازشناسی به میزان ۲۰٪ شده است [۳۱]. همچنین برخی از نتایج به‌دست آمده در این تحقیق برای زبان فارسی در مورد دو حالت گفتاری خشم و اندوه نیز در [۳۲] ارائه شده است.در تحقیقات قبلی، بررسی اثر حالت گفتار بر پارامترهای گفتاری نظیر فرمنت‌ها، فرکانس گام و طول زمانی[2] در زبان فارسی انجام شده است. طول زمانی، فاصلهٔ زمانی است که سیستم برچسب زنی زمانی برای هر واکدار به‌دست آورده است [۳۳]. در ادامهٔ این تحقیق مختصری دربارهٔ نحوهٔ تأثیرگذاری حالت گفتار بر پارامترهای فوق، توضیح ارائه خواهد شد. از جملهٔ این پارامترها می‌توان به فرکانس گام، انرژی و طول زمانی اشاره کرد [۱، ۳٤، ۳۵].

همانطور که گفته شد، در زمینهٔ بازشناسی گفتار با حالت، تحقیقات گسترده‌ای صورت نگرفته است. لذا با هدف گسترش اطلاعات در این زمینه، خصوصاً در زبان فارسی، در این تحقیق به بررسی تأثیر پارامترهایی نظیر فرمنت‌ها و فرکانس گام در بهبود نتایج بازشناسی گفتار با حالت، پرداخته می‌شود. لازم است ذکر شود که با توجه به عدم وجود تحقیقات مشابه، هدف این تحقیق بررسی انواع پارامترها و یافتن مناسب‌ترین آنها است. لذا لزوماً تمامی نتایجی که در ادامه ارائه می‌شوند، قابل توجه نیست. همچنین روش‌هایی برای شناسایی حالت گفتار بررسی می‌شود.

۲- دادگان و ابزار مورد استفاده

دادگان اصلی مورد استفاده فارس دات است [۳٦]. دادگان فوق با استفاده از گفتار پیوسته در زبان فارسی ایجاد شده است. این دادگان حاوی ٦۰۰۰ جمله از ۳۰۰ گوینده زن و مرد است که با گویش‌های مختلف رایج در ایران بیان شده است. این جملات در واقع ۳۹۰ عبارت مختلف است که توسط گوینده‌های متفاوت بیان شده‌اند. از این میان از ۱۸۰۰ جمله برای آموزش سیستم بازشناسی استفاده شده است. این جملات، با لهجهٔ رسمی ایرانی (تهرانی) بیان شده است. از دادگان فوق برای آموزش مدل‌های بازشناسی استفاده شده است. دادگان فوق را D1TR می‌نامیم.

با توجه به در دسترس نبودن دادگان احساسی در زبان فارسی، در این تحقیق از یک گویندهٔ مرد برای ایجاد این دادگان استفاده شده است. تعداد ٦۳ جملهٔ مناسب از جملات دادگان فارس دات با حالت اندوه، سه بار توسط گویندهٔ فوق ادا شده است. همچنین ۵۳ جمله با شرایط فوق سه بار با حالت خشم بیان شده است. بعضی از این جملات، با جملات دارای حالت احساسی اندوه متفاوت است. این دو دادگان را به‌ترتیب D2EG و D2EA می‌نامیم.

همچنین گویندهٔ فوق سه بار تمامی جملات دادگان فارس دات را به‌صورت عادی بیان کرده است. از این قسمت برای مقایسهٔ نحوهٔ تغییرات پارامترهای گفتاری نسبت به‌حالت عادی استفاده شده است. این دادگان را D2N می‌نامیم. قسمتی از دادگان D2N که از جملات آن برای ایجاد دادگان با حالت غم و اندوه استفاده شد، D2NG و قسمتی را که برای ایجاد دادگان با حالت عصبانیت استفاده شد D2NA می‌نامیم. از دادگان‌های D2NG و D2NA برای مقایسه نتایج بازشناسی با دادگان‌های D2EG و D2EA استفاده شده است.

1. British National Corpus
2. Duration

ابزار مورد استفاده در برچسب‌زنی زمانی در این تحقیق، HMM است. برای راه‌اندازی HMM از نرم‌افزار HTK [۳۷] استفاده شده است. همچنین استخراج فرمنت‌ها با استفاده از برنامه‌های موجود به روش پیش‌بینی خطی طیف [۳۸]، صورت گرفته است. استخراج فرکانس گام در این تحقیق با روش ارائه شده توسط Medan et.al [۳۹] انجام شده است. برای اعمال این روش از ابزار PDA[1] از نرم‌افزار Speech ToolBox [٤٠] استفاده شده است.

۳- بازشناسی گفتار با حالت

همانطور که گفته شد، جمع‌آوری دادگان گفتار با حالت مشکل است و به‌تبع آن ایجاد سیستم بازشناسی که برای آموزش مدل‌های آن از گفتار با حالت استفاده شده، کار ساده‌ای نخواهد بود. لذا در این تحقیق، بازشناسی گفتار با حالت نیز با استفاده از مدل‌های عادی انجام شده است (مدل‌هایی که برای آموزش آنها از گفتار عادی و بدون هر گونه حالت استفاده شده است).

سیستم بازشناسی مورد نظر در این تحقیق HMM است. این سیستم بازشناسی گفتار، مبتنی بر بازشناسی گفتار پیوسته با مدل‌هایی بر مبنای واج‌های پایه است. فریم‌های گفتاری با طول ۲۵msec و با همپوشانی ۱۰msec در نظر گرفته شده است. بردار ویژگی HMM شامل ۱۲ ضریب کپسترال و انرژی، همراه با پارامترهای سرعت و شتاب برای ۱۳ ضریب ذکر شده است. در تحقیق قبلی نشان داده شد که حالت گفتار بر پارامترهایی نظیر فرمنت‌ها، فرکانس گام و طول زمانی تأثیر دارد [۳۳]. در این تحقیق به بررسی تأثیر حالت‌های خشم و اندوه بر پارامترهای گفتاری نظیر طول زمانی، فرمنت‌ها، فرکانس گام و شیب آنها پرداخته شد. برای محاسبهٔ شیب فرکانس گام مانند محاسبهٔ پارامترهای سرعت و شتاب ضرایب کپسترال از رابطهٔ ۱ استفاده شده است.

$$d_t = \frac{\sum_{n=1}^{M} n \times (x_{t+n} - x_{t-n})}{2\sum_{n=1}^{M} n^2} \qquad (1)$$

در رابطهٔ فوق M طول پنجره و x پارامتر مورد نظر است که باید شیب آن محاسبه شود. M در این تحقیق برابر ۲ در نظر گرفته شده است.

از میان موارد ذکر شده در این تحقیق می‌توان به موارد زیر اشاره کرد:

- در حالت اندوه و خشم، طول زمانی واکه‌ها نسبت به حالت عادی افزایش می‌یابد. بررسیها نشان داد که علی‌رغم تصور قبلی در حالت عصبانیت طول واکه‌ها کاهش نمی‌یابد، بلکه گوینده با برجسته کردن واکه‌ها از نظر طول و انرژی، حالت خود را بیان می‌کند. یکی از ویژگیهای بارز گفتار در حالت خشم و عصبانیت، آن است که گوینده فاصلهٔ سکوت بین کلمات را کاهش می‌دهد و لذا در کل، گفتار سریعتر به‌نظر می‌رسد.
- در حالت گفتاری غم و اندوه، گوینده نرخ ادای گفتار کمتری دارد.

در حالت عصبانیت افزایش طول واکه‌های ضعیف (شامل /æ/ ، /ε/ و /o/) بیشتر است.

- تغییرات طول زمانی برای سایر واج‌ها نیز مشاهده می‌شود، اما مانند تغییرات طول زمانی در واکه‌ها، این تغییرات کاملاً منظم نیست.
- بررسی نتایج همچنین مشخص کرد که همخوان‌های انفجاری نظیر /b/ ، /d/ و /p/ معمولاً در حالت غم و اندوه کاهش طول پیدا می‌کنند. همچنین غلتان /r/ تغییرات طولی زیادی ندارد.

1. Pitch Detection Algorithm

• معمولاً طول همخوان‌های واکدار نیز در حالت عصبانیت افزایش می‌یابد.

• حالت خشم تأثیر چشمگیری بر افزایش میانگین فرکانس‌های فرمنت، به‌ویژه فرمنت‌های اول و دوم واکه‌ها دارد. به‌عنوان مثال متوسط فرمنت اول برای همهٔ واکه‌ها به‌شدت زیاد می‌شود. فرمنت دوم نیز معمولاً اضافه می‌شود. فرمنت سوم نسبت به سایر فرمنت‌ها تغییرات کمتری دارد.

• در حالت اندوه نیز فرکانس‌های فرمنت تغییر خواهند کرد. در این میان مقدار میانگین فرمنت اول و سوم افزایش و برای فرمنت دوم کاهش می‌یابد.

• نتایج بررسیها نشان داد که فرمنت اول برای تمامی همخوان‌های واکدار در حالت عصبانیت افزایش قابل‌توجهی نسبت به‌حالت عادی پیدا کرده است. فرمنت دوم نیز برای اکثر این همخوان‌ها بر اثر ایجاد حالت عصبانیت کاهش داشته است.

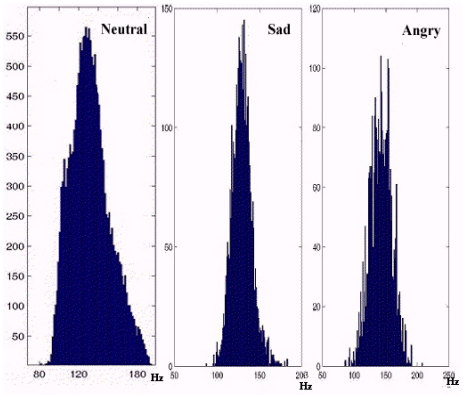
• در مورد حالت گفتاری ناراحتی، میانگین فرمنت اول معمولاً افزایش یافته است. میانگین فرمنت دوم نیز معمولاً در این حالت گفتاری کاهش داشته است. فرمنت سوم هم در بیشتر موارد افزایش داشته است.

• بررسیها نشان داد که حالت خشم سبب افزایش فرکانس گام برای واکه‌های قوی و ضعیف، و نزدیکتر شدن مقادیر آنها به هم خواهد شد. در حالت اندوه نیز فرکانس گام زیاد شده و به مقداری ثابت نزدیک می‌شود. به‌عبارت دیگر، واریانس تغییرات آن کاهش می‌یابد.

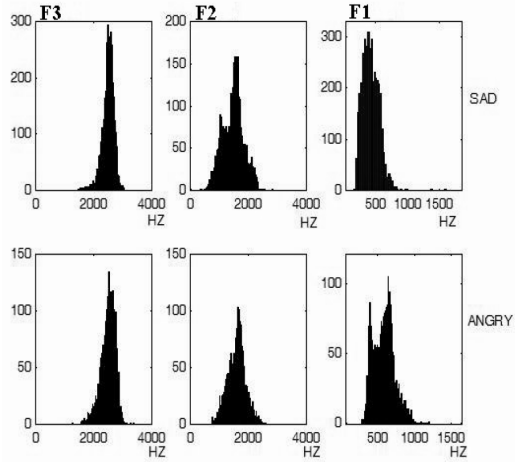
بیشتر نتایج به‌دست آمده در زبان فارسی، در سایر تحقیق‌های انجام شده مورد تأیید قرار گرفته است [۱، ۲، ٤۱، ٤۲].

همچنین در شکلهای ۱ و ۲ هیستوگرام فرکانس پایه و فرکانس‌های فرمنت برای دو حالت گفتاری مشاهده می‌شود. این هیستوگرام‌ها بیانگر نحوهٔ تأثیر حالت گفتار بر پارامترهای موردنظر است.

همچنین در تحقیقی دیگر در زبان فارسی با ۵ گوینده نتایج به‌دست آمده در این تحقیق در مورد نحوهٔ تأثیر حالت گفتار بر فرکانس گام و شیب آن مورد تأیید قرار گرفته است [٤۳].



شکل ۱ هیستوگرام کلی فرکانس گام برای گفتار بدون حالت، حالت اندوه و خشم



شکل ۲ هیستوگرام کلی فرکانس فرمنت اول تا سوم برای گفتار حالت اندوه و خشم

بدیهی است که در یک دید کلی‌تر، می‌توان گفت که ایجاد حالت در گفتار سبب تغییراتی در ضرایب کپسترال خواهد شد. با توجه به نقش اصلی ضرایب کپسترال در سیستم بازشناسی، کاهش نتایج بازشناسی گفتار با حالت، با

استفاده از مدل‌های عادی قابل پیش‌بینی خواهد بود. برای جبران کاهش بازده سیستم بازشناسی، دو راه را می‌توان پیشنهاد کرد. گزینه اول جبران‌سازی اثر حالت گفتار بر ضرایب کپسترال است، به‌گونه‌ای که بتوان ضرایب کپسترال را به‌گونه‌ای در جهت عکس تغییر داد که به‌صورت اول (به‌صورت گفتار بدون حالت) برگردند. با توجه به رفتار پیچیده‌ای که اصولاً ضرایب کپسترال از خود نشان می‌دهند، به‌نظر می‌رسد که این کار مشکل خواهد بود. در گزینه دوم می‌توان پارامترهای دیگری را علاوه بر ضرایب کپسترال به بردار ویژگی اضافه کرد. این ضرایب به شرطی می‌توانند در بهبود نتایج بازشناسی مفید باشند که نسبت به اعمال حالت در گفتار پایدار بوده یا لااقل تغییرات منظم داشته باشند.

در این تحقیق، با استفاده از پارامترهای مذکور در بردار ویژگی به بررسی میزان تأثیر فرمنت‌ها، فرکانس گام و شیب آنها در بازشناسی گفتار با حالت خواهیم پرداخت. نکته‌ای که باید به آن اشاره شود آن است که اگر چه بررسیها نشان داد که طول زمانی نیز بر اثر اعمال حالت در گفتار دارای تغییرات منظمی است، اما استفاده از طول زمانی در سیستم بازشناسی چندان آسان نخواهد بود. لذا در این تحقیق در این مورد بحثی نخواهد شد.

برای کاهش میزان تأثیر مقدار متوسط ضرایب کپسترال، از روش CMS[1] استفاده شده است. مقایسهٔ نتایج نشان داد که استفاده از روش CMS باعث بهبود نتایج بازشناسی خواهد شد.

بر اساس نکاتی که به آن اشاره شد، مطابق جدول ۱ مدل‌هایی برای بازشناسی گفتار تشکیل شد. پارامتر اصلی این مدل‌ها همان ضرایب کپسترال و انرژی نرمالیزه شده است که در مدل M0 این جدول دیده می‌شود. در مدل‌های M1 تا M8، علاوه بر ضرایب کپسترال از فرمنت‌ها، فرکانس گام و شیب آنها در ساختن مدل‌های بازشناسی استفاده شده است. روشن است که آموزش کلیه این مدل‌ها با استفاده از دادگان D1TR انجام شده است.

جدول ۱ مشخصات مدل‌های ساخته شده با استفاده از فرمنت‌ها و فرکانس گام

مدل‌ها	بردار ویژگی
M0	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)$
M1	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+F1$
M2	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+F2$
M3	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+F3$
M4	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+\Delta F1$
M5	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+\Delta F2$
M6	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+\Delta F3$
M7	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+F0$
M8	$C+LE+\Delta(C+LE)+\Delta^2(C+LE)+\Delta F0$

برای ایجاد امکان مقایسه نتایج بازشناسی دادگان با حالت با دادگان عادی، در جدول ۲ این نتایج برای دادگان‌های D2NG و D2NA آورده شده است. مقایسهٔ این نتایج با نتایج بازشناسی گفتار با حالت، میزان تأثیر حالت گفتار را در بازشناسی نشان می‌دهد.

جدول ۲ نتایج بازشناسی با استفاده از مدل M0 برای دادگان‌های D2NG و D2NA

مدل	دادگان	تعداد عناصر مخلوط					
		۱	۲	۳	٤	٥	٦
M0	D2NG	٥٢/٩٩	٦١/٧٨	٦٨/٠٥	٦٩/٥٩	٧٤/١٣	٧٤/٩٠
	D2NA	٤٩/٣٢	٥٧/٢٧	٦٤/٨٩	٦٧/٣٩	٦٩/٥٥	٧١/٣٦

در جدول ۳ نتایج بازشناسی با استفاده از مدل‌های M0 تا M8 در دو حالت گفتاری مورد نظر مشاهده می‌شود. این نتایج برای مخلوط‌های دوم تا ششم در این جدول مشاهده می‌شود. نکته‌ای که باید به آن اشاره شود آن است که نتایج بازشناسی مربوط به‌حالت گفتاری

1. Cepstral Mean Subtraction

خشم، با توجه به تغییرات گسترده‌ای که گوینده در ادای این حالت نسبت به‌حالت عادی دارد، با مدل‌های عادی معمولاً قابل توجه نیست. اما به‌دلیل وجود امکان مقایسه با حالت گفتاری اندوه، این نتایج آورده شده است.

جدول ۳ نتایج بازشناسی گفتار با حالت با استفاده از مدل‌های M0 تا M8

حالت	مدل	تعداد عناصر مخلوط					
		۱	۲	۳	٤	٥	٦
اندوه	M0	۲۷/۵۱	۲۲/٤۹	۲۷/۹۹	۳۱/۲۷	۳٦/۳۹	۳۷/۱٦
	M1	۲۷/۵۰	۲۲/۷۳	۳۳/٤۰	۳٤/۲۷	۳٤/۵٦	۳۳/٤۹
	M2	۲۸/۳۰	۲۲/٦٤	۲۸/۹٦	۳۵/۵۲	۳۷/۹۳	۳۵/۸۹
	M3	۲۷/٤۱	۲۳/۲۲	۳۰/۵۰	۳۵/۲۷	۳۹/۵۳	۳٦/٦٦
	M4	۲۵/۸۰	۲۲/۵۰	۲۹/۰۰	۲۹/۸۵	۳۳/۸٦	۳٤/٤۱
	M5	۲۸/٦۳	۲٤/۲٤	۲٦/۸۱	۳۱/۸۱	۳۵/۷۰	۳۸/۲۲
	M6	۲۷/٦۹	۲۳/٦۸	۲۸/۵٦	۳۱/۰۸	۳٦/۸۷	۳۸/٤۲
	M7	۳۰/۳۵	۲٤/۰۸	۲۹/۷۸	۳٤/۲۷	۳٦/۷۸	۳۸/۰۳
	M8	۲۵/۲٤	۲۲/۵۱	۳۰/۲۱	۳٤/٦۳	۳٦/٦۸	٤۲/۸٦
خشم	M0	۵/۲۳	۰/۹۱	۰/۸۰	۲/۵۰	٦/۵۹	٦/۳٦
	M1	٦/۹۵	۱/۱۸	۰/۰۰	۰/۰۰	۱/۲۵	۲/۲۷
	M2	۵/۳۲	۱/۲۱	۰/۰۰	۰/۰۰	۲/۸٤	۱/۷۳
	M3	٤/۳۷	۱/۹۰	۰/۵۸	۰/۰۰	۱/۰٤	۰/۰۰
	M4	۵/۲٦	۳/۲۷	۲/۸۵	۰/۰۰	۲/۹۰	۳/۳۸
	M5	۵/۵۰	۲/۷۲	۰/۹۷	۱/۱٤	۲/۵۳	٤/٤٦
	M6	٦/۱۰	٤/۰۰	۳/٤۲	۰/۰۰	۱/۱٤	۳/۵٤
	M7	۵/۸۷	٤/۳۲	٤/۸۲	٤/۰۳	۸/٦۳	۸/٦۳
	M8	٦/۲۹	۵/۹٤	۰/۸۸	٤/۵۸	۱/۰٤	۳/۲۲

از مجموعهٔ نتایج ارائه شده می‌توان مطالب زیر را بیان کرد:

- همان‌گونه که انتظار داشتیم نتایج بازشناسی گفتار با حالت خشم خیلی پایین است. مقایسهٔ این نتایج با نتایج بازشناسی دادگان D2NA در جدول ۲ نیز بیانگر این موضوع است.
- با توجه به نرخ بسیار کم بازشناسی در گفتار با حالت خشم نمی‌توان در مورد تأثیر پارامترهای نوایی در بهبود کیفیت بازشناسی بحث کرد. اما به‌طور کلی می‌توان گفت که فرکانس گام در این حالت گفتاری، تأثیری مثبت در بهبود نرخ بازشناسی داشته است.
- نرخ بازشناسی گفتار در حالت اندوه نیز نسبت به‌حالت عادی کاهش داشته است. مقایسهٔ نتایج جدول ۲ در مورد دادگان D2NG با نتایج جدول ۳ برای این حالت گفتاری نشان می‌دهد که به‌صورت نسبی بیش از ۵۰٪ کاهش در نتایج بازشناسی به‌وجود آمده است. تغییرات ویژگیهای گفتاری در حالت اندوه و در نتیجه تغییرات عمده در ضرایب MFCC گفتار را با مدل‌های پایه با کاهش شدید نرخ بازشناسی مواجه کرده است.
- با توجه به تغییرات زیاد فرمنت اول در حالت اندوه و تغییرات کمتر فرمنت‌های دوم و سوم (به‌ویژه فرمنت سوم) تأثیر آنها در بازشناسی این حالت گفتاری بیشتر است. شیب فرمنت دوم نیز در بازشناسی گفتار تأثیر مثبت داشته است. شیب فرمنت سوم بیشترین تأثیر را در بازشناسی این حالت گفتاری دارد. این تأثیر با توجه به پایداری بیشتر فرکانس‌های فرمنت دوم و سوم و تغییرات کمتر آنها نسبت به گفتار عادی قابل توجیه است. فرمنت‌های دوم و سوم در شرایط استفاده از ٤ عنصر مخلوط، در حدود ٤٪ نرخ بازشناسی را بهبود داده‌اند.
- در مورد حالت اندوه نتایج جدول نشان می‌دهد که فرکانس گام نیز در بهبود نتایج مؤثر است و حداکثر ۳٪ در شرایط استفاده از چهار عنصر مخلوط، نتایج بازشناسی را بهتر کرده است. تغییرات منظم فرکانس گام در حالت گفتاری اندوه نسبت به گفتار عادی می‌تواند سبب تأثیر این ویژگی برای بهبود نتایج بازشناسی گفتار با حالت اندوه باشد.
- شیب فرکانس گام نیز در تعداد عناصر مخلوط بالا می‌تواند اثر مثبت داشته باشد. حداکثر میزان بهبود مربوط به استفاده از ٦ عنصر مخلوط و در حدود ۵/۷٪ است. شیب فرکانس گام با توجه به تغییرات بیشتر آن نسبت به فرکانس

گام با تعداد عناصر مخلوط بیشتر بهتر مدل شده و لذا نتیجهٔ بازشناسی با ٦ عنصر مخلوط بهترین مقدار را دارد.

لازم است ذکر شود که بررسیهای قبلی نشان داد که استفادهٔ ترکیبی از پارامترهای نوایی نه‌تنها منجر به بهبود نتایج بازشناسی نخواهد شد، بلکه دقت بازشناسی را حتی نسبت به مدل‌های پایه کاهش می‌دهد [٤٤]. در تحقیق فوق از مدلی با ضرایب MFCC و انرژی همراه با پارامترهای سرعت و شتاب و مقادیر سه فرکانس فرمنت در هر فریم استفاده شده است. یک دلیل قابل ارائه برای این رفتار می‌تواند احتمال وابستگی پارامترهای مذکور به‌یکدیگر باشد که سبب کاهش دقت سیستم بازشناسی گفتار شده است.

٤- شناسایی حالت گفتار

شناسایی حالت گفتار در جهت بازشناسی گفتار با حالت می‌تواند مفید باشد. در سیستم‌های بازشناسی گفتار با حالت، با شناسایی حالت گفتار می‌توان از مدل‌های مربوط به آن حالت گفتاری استفاده کرد و یا مدل زبان، متناسب با حالت گفتار شناسایی شده را در نظر گرفت. همچنین می‌توان مشابه کاری که در این تحقیق انجام شده، پارامترهای مفیدی را به بردار ویژگی اضافه کرد. به‌عنوان مثال اضافه کردن فرمنت سوم در حالت گفتاری غم و اندوه، بازده بازشناسی گفتار را در این حالت افزایش می‌دهد.

در تمامی این موارد باید ابتدا حالت گفتار را شناخت. با توجه به تأثیر حالت گفتار بر پارامترهای گفتاری نظیر فرمنت‌ها و فرکانس گام و شیب آنها [٣٣]، می‌توان از این پارامترها در شناسایی حالت گفتار استفاده کرد. دقت روش به‌کار رفته برای شناسایی حالت در گفتار به عوامل متعددی بستگی دارد. طبیعی است که روش به‌کار رفته باید در شناسایی حالت مورد نظر دارای دقت بالایی باشد. از طرف دیگر توانایی روش به‌کار رفته در تمیز بین دو حالت گفتاری نیز دارای اهمیت است. به‌عنوان مثال روشی ممکن است در شناسایی حالت اندوه دارای دقت بالایی باشد، اما در مورد حالت خشم دارای دقت بالایی نباشد و بسیاری از فریم‌های گفتاری در حالت خشم، به‌صورت حالت اندوه شناسایی شوند. در این‌صورت نمی‌توان در همهٔ موارد به برچسب‌هایی که فریم گفتاری را به‌حالت اندوه شناسایی می‌کنند، اعتماد کرد. در واقع بهترین روش، روشی است که در مورد تمامی حالتهای مورد نظر از دقت خوبی برخوردار باشد. در این تحقیق شناسایی حالت در گفتار به دو شکل مطرح شده است. در شکل اول می‌توان فرض کرد که گفتار حتماً دارای یکی از دو حالت خشم یا اندوه است و روش مورد نظر باید یکی از این دو حالت را انتخاب کند. در روشی کامل‌تر می‌توان گفت گفتار مورد نظر می‌تواند بدون حالت باشد یا یکی از این دو حالت گفتاری را داشته باشد. در این‌صورت روش مورد نظر باید یکی از سه گزینه بدون حالت، اندوه یا خشم را انتخاب کند. در این تحقیق در هر دو مورد، بررسی صورت گرفته است. همچنین برای شناسایی حالت، از دو روش استفاده شده که در ادامه به آن پرداخته می‌شود.

٤-١- شناسایی حالت گفتار با استفاده از معیار بیشینه درستنمایی

در این روش، با به‌دست آوردن مقدار درستنمایی بیشینه[1] در مورد هر فریم گفتاری، تصمیم‌گیری شد. در واقع، برای هر یک از پارامترهای نوایی مانند فرمنت‌ها، فرکانس گام و شیب آنها در دادگان آموزش، یک توزیع گوسی به‌دست آورده شده است. هر چند شاید فرض توزیع آماری پارامترها به‌صورت تک‌گوسی، کمی خوش‌بینانه باشد، اما نتایج نشان داد که با چنین فرضی،

1. Maximum Likelihood (ML)

تا حدی می‌توان حالت گفتار را تشخیص داد. نحوهٔ تشکیل این توزیع چنین است که در دادگان آموزش با فرض دانستن حالت گفتار، در مورد هر حالت گفتاری و برای هر فریم گفتاری، پارامترهای نوایی مورد نظر استخراج شده و مجموعهٔ داده‌های هر یک از این پارامترها در کل دادگان آموزش، تشکیل توزیع گوسی با مقدار متوسط و واریانس قابل محاسبه می‌دهند. برای بررسی شباهت فریم آزمایشی به داده‌های این توزیع، می‌توان از رابطه ۲ استفاده کرد.

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}exp\left(\frac{(x-m)^2}{2\sigma^2}\right) \qquad (2)$$

مقدار حاصل از رابطه ۲ می‌تواند معیاری برای شباهت فریم به یک حالت گفتاری باشد. در رابطهٔ فوق x پارامتر مورد نظر، m مقدار میانگین و σ انحراف معیار x است.

در جدول ٤ مقادیر آماری مربوط به فرمنت‌ها، فرکانس گام و شیب آنها در دادگان موجود نشان داده شده است. برای به‌دست آوردن شیب فرکانس‌های فرمنت یا شیب فرکانس گام از رابطهٔ (۱) استفاده شده است. در این جدول مقادیر متوسط و انحراف معیار توزیع‌ها برای پارامترهای نوایی در گفتار عادی، با حالت خشم و غم و اندوه مشاهده می‌شود. در مورد هر جمله از مجموعه دادگان آزمایش، مجموعهٔ پارامترهای فوق استخراج شده و مطابق جدول ۵ برای شناسایی حالت گفتار استفاده شده است. در این جدول استفاده از هر پارامتر نوایی یا ترکیبی از آنها مد نظر بوده است. در این جدول نتایج شناسایی یکی از دو حالت گفتاری اندوه و خشم از هم ارائه شده است. پارامترهایی که در این دو حالت گفتاری بیشترین تفاوت را با هم داشتند فرمنت اول و فرکانس گام بودند. هر چند تفاوت بسیار زیاد سطح انرژی در دو حالت گفتاری وجود دارد (در این جا لگاریتم انرژی در نظر گرفته شده است). در این جدول نتایج با استفاده از انرژی و همچنین بدون استفاده از آن ارائه شده است، زیرا دلیلی برای تفاوت انرژی در بین برخی از حالات گفتاری وجود ندارد. ردیفهای ۲، ۱۰ و ۱۱ که از فرکانس گام و فرمنت اول استفاده کرده‌اند، بهترین نتایج را داشته‌اند (با صرف‌نظر از مواردی که از انرژی استفاده کرده‌اند). در این میان نتایج شناسایی حالت عصبانیت بیشتر است. همان‌گونه که مشاهده شد این حالت گفتاری به‌نحو چشم‌گیری پارامترهای گفتار را تحت تأثیر قرار می‌دهد. مقایسهٔ نتایج سطرهای ۲، ٦ و ۱۱ جدول فوق نشان می‌دهد که استفادهٔ توأم از فرمنت اول و شیب آن دقت شناسایی حالت خشم را کاهش می‌دهد. با توجه به ردیف ٦ در جدول فوق می‌توان گفت که شیب فرمنت اول در شناسایی حالت غم و اندوه ناتوان است و لذا برخی از نمونه‌های گفتاری دارای حالت خشم که توسط فرمنت اول درست شناسایی می‌شدند، با اضافه شدن شیب فرمنت اول اشتباهاً به‌صورت اندوه برچسب خورده‌اند، هر چند درصد کمی نیز به‌دقت شناسایی حالت غم و اندوه در ردیف ۱۱ افزوده شده است.

جدول ٤ آمارگان پارامترهای استفاده شده در شناسایی حالت گفتار

پارامتر	m_s	σ_s	m_a	σ_a	m_n	σ_n
F0	۱۲۹/۲	۱۱/۹	۱٤۲/۲	۱۷/۱	۱۳۲/۱	۲۰/۳
F1	٤۲٦/٤	۱٤٥/۸	٥۸٤/۸	۱۳۹/٤	٤۳۲/۱	۱۷۲/۲
F2	۱٤۷۸/٤	۳٥۷/٥	۱٥۹٤/٥	۳۰۰/۸	۱٥۳۹/۰	۳٦٥/۳
F3	۲٤۹٥/۰	۲۱۲/۸	۲٤۹۸/۱	۲٥۰/۹	۲٥۰٥/۹	۲۱۰/۱
ΔF0	۰/۰۹	٤/۷	-۰/۸۷	٦/۷۳	-۰/٥٥	۱/۳۱
ΔF1	۲/۱۳	۲٤/۹	-۰/۰٤	۱٥/٤۷	۲/۱۰	۲۳/٥۰
ΔF2	٥/۱۹	٦۹/۷	-۰/۳۲	۳٦/٦۷	-۱/۳۰	٤۱/٦۰
ΔF3	-٦/٦٦	۱۰۹/۳	-۲/۲۷	٥٥/۹۱	-۰/۷	٤۲/۱۰

از نتایج جدول ۵ می‌توان گفت که انرژی به‌تنهایی هم برای تمیز بین این دو حالت مناسب است. در ردیفهای ۱۲، ۱۳ و ۱٤ نتایج ردیف‌هایی که بدون انرژی شناسایی خوبی داشتند همراه با انرژی ارائه شده است. همان‌گونه که در جدول مشاهده می‌شود، در نهایت می‌توان به‌دقت حدوداً ۹٤٪ برای حالت خشم و ۷۰٪ برای اندوه رسید. حالت اندوه همواره دارای سطح مشخص انرژی نیست و می‌تواند با انرژیهای مختلف در نظر گرفته شود و این در حالی است که لازمهٔ خشم، انرژی زیاد است. بررسیها نشان داده که حالت اندوه بیشترین امکان اشتباه شدن با حالت عادی، خستگی یا خشم را دارد [۸].

جدول ۵ نتایج شناسایی حالت گفتار با استفاده از پارامترهای مختلف

شماره	پارامتر	نرخ شناسایی	
		اندوه	خشم
۱	F0	٦۷/۸	۵۰/٦
۲	F1	٦۵/۸	۷۲/۰
۳	F2	٤۳/۵	۷۲/۰
٤	F3	۷۰/٦	٤۷/٤
۵	ΔF0	۸٤/۵	٤۸/۳
٦	ΔF1	۲۸/۵	۷۹/۲
۷	ΔF2	۲۱/۲	۸۰/۷۹
۸	ΔF3	۱٦/۵	۷۹/۸
۹	F0+ΔF0	٦۹/۲	٦۷/٦
۱۰	F0+F1	٦۵/٦	۷۱/۱۳
۱۱	F1+ΔF1	٦٦/٤	٦۸/۵
۱۲	E	۷۷/۸۸	۹۰/۷۹
۱۳	F1+E	۷۰/۵۷	۹٤/۳۹
۱٤	F0+F1+E	۷۰/۵۰	۹۳/۸
۱۵	F1+ΔF1+E	۷۰/۸۰	۹۳/۱

در جدول فوق هدف، تمایز بین یکی از دو حالت گفتاری اندوه و خشم بوده است. در جدول ٦، میزان اشتباه شدن با گفتار عادی نیز در نظر گرفته شده است. به‌عبارت دیگر در این جدول، گفتار با حالت، با گفتار عادی نیز مقایسه شده و درصد اشتباه شدن فریم گفتاری با حالت، با گفتار عادی مشخص شده است.

در این جدول به‌نظر می‌رسد که انرژی نمی‌تواند نقش چندان مهمی داشته باشد، زیرا در حالت گفتاری اندوه و گفتار عادی ما تقریباً سطح انرژی یکسانی را خواهیم داشت. مقادیر آماری استفاده شده برای گفتار عادی در قسمت آموزشی دادگان D۲N در دو ستون آخر جدول ٤ مشاهده می‌شود.

مجموع نتایج جدول ٦ نشان می‌دهد که امکان اشتباه شدن بین حالت گفتاری اندوه و گفتار عادی زیاد است. نتایج ردیفهای ۱، ۲، ۱۰ و ۱۱ بیانگر اهمیت F۰ و F۱ در تمایز بین این دو وضعیت است. تغییرات زیاد ناشی از حالت گفتار در این پارامترها سبب مفید بودن آنها در شناسایی حالت گفتار است. نکته‌ای که باید به آن اشاره شود این است که در بحث بازشناسی گفتار پارامترهایی می‌توانند بیشتر مفید باشند که دارای تغییرات کمی نسبت به گفتار عادی بوده و از طرفی بتوانند به‌خوبی توسط HMM مدل شوند. به‌گونه‌ای که استفاده توأم از آنها همراه با ضرایب MFCC بتواند نرخ بازشناسی را بهبود دهد. از طرف دیگر در مورد شناسایی حالت گفتار پارامترهایی می‌توانند نقش مهمتری داشته باشند که دارای تغییرات عمده‌ای در حالتهای مختلف گفتاری باشند. هر چند در این حالت مدل شدن مناسب آنها نیز توسط توزیع آماری شرط این تأثیر است. لذا با استفاده از این پارامترها می‌توان حالات گفتاری اندوه و عصبانیت را از هم تمیز داد.

جدول ٦ نتایج شناسایی حالت گفتار با استفاده از پارامترهای مختلف بین گفتار عادی و گفتار با حالت

شماره	پارامتر	حالت گفتار					
		اندوه			خشم		
		اندوه	خشم	عادی	اندوه	خشم	عادی
۱	F0	۷٤/۵	۱۹/۸	۵/٦	۲۵/۷	۵۹/۹	۱٤/۳
۲	F1	۵٦/۵	۳٦/۵	۷/۰	۳۰/۷	٦۹/۲	۱۰/۰
۳	F2	۳٤/٤	٤۹/۵	۷/۰	۳۲/۹	٦۱/۱	۵/۹
٤	F3	۲۰/۸	۲۲/۱	۵۷/۰	۱٦/۲	٤۲/۲	٤۱/٤
۵	ΔF0	۲/۳	۱۱/٦	۸۵/۹	۱۵/۸	٤۸/۱	۳٦/۱
٦	ΔF1	۲۱/۲	۷۱/٦	۷/۱	۱۸/۸	۷۵/٤	۵/٦
۷	ΔF2	۲۰/٤	۷۳/۲	٦/۳	۱۷/٤	۷۵/۲	۷/۳
۸	ΔF3	۷/۵	۱۵/۵	۷٦/۹	۲۰/۵	۱۱/۷	٦۷/۸
۹	F0+ΔF0	۳/۳	۱۲/۲	۸٤/٤	۱۰/۱	۵٤/٦	۳۵/۳
۱۰	F0+F1	٦۵/۹	۲۳/۹	۱۰/۰	۱٦/۵	٦۱/٦	۲۱/۸
۱۱	F1+ΔF1	۵۱/۹	۳۷/۰	۱۰/۹	۳۳/۵	٦۳/۸	۲/۷

٤-۲- شناسایی حالت گفتار با استفاده از GMM

همانطور که گفته شد، استفاده از توزیع گوسی برای شناسایی حالت در گفتار ممکن است کافی نباشد. لذا در این قسمت از GMM[1] برای شناسایی حالات گفتاری اندوه و خشم و همچنین گفتار عادی استفاده می‌شود.

GMM یکی از ابزارهای متداول و پر کاربرد در پردازش گفتار است. یکی از کاربردهای GMM در شناسایی گوینده به‌صورت ناوابسته به متن است [٤۵]. شناسایی حالتهای مختلف گفتار نیز می‌تواند به نوعی مشابه کاربرد GMM در شناسایی گوینده باشد، به‌نوعی که هر یک از حالتهای گفتار می‌تواند به‌عنوان گفتار گوینده فرض شود. بر این اساس مدل‌های آماری به‌کار رفته در GMM در حالت آموزش با داده‌های آموزشی هر یک از دو حالت گفتاری همراه با گفتار عادی آموزش داده شده‌اند. برای تخمین پارامترهای GMM مورد استفاده، از روش درستنمایی بیشینهٔ (ML)[2] استفاده شده است و تخمین پارامترهای ML با استفاده از الگوریتم EM[3] صورت گرفته است. همچنین برای تخمین اولیهٔ پارامترهای مدل‌ها نیز از الگوریتم K-Means در شرایط تقسیم باینری[4] استفاده شده است.

GMM مورد استفاده دارای ۳۲ عناصر مخلوط و بردار ویژگی ٤۳تایی است [٤۵]. این بردار ویژگی شامل ۱۲ ضریب کپسترال و لگاریتم انرژی همراه با پارامترهای سرعت و شتاب (جمعاً ۳۹ پارامتر) و همچنین فرمنت‌های سه‌گانه و فرکانس گام است. برای آموزش GMM از ۳۰ ثانیه گفتار در هر یک از حالات مورد نظر استفاده شده است. بنابر این دو مدل متفاوت GMM برای هر دو حالت گفتاری موجود و یک مدل برای گفتار عادی در نظر گرفته شده است. در حالت آموزش GMM با استفاده از داده‌های آموزشی مدل‌های مزبور به اندازهٔ کافی آموزش داده شد. پس از آن در حالت آزمایش، هر فریم گفتاری دادگان آزمایش به هر یک از این مدل‌های GMM اعمال شده است. در مجموع برای آزمایش از ٤۰ ثانیه داده استفاده شده است. نتایج شناسایی حالت گفتار در جدول ۷ مشاهده می‌شود.

نتایج جدول ۷ نشان می‌دهد که استفاده از GMM سبب افزایش دقت شناسایی حالت گفتار شده است. در این میان حالت خشم با توجه به تفاوتهای زیاد پارامترهای آن با گفتار عادی و حالت اندوه، به‌راحتی شناسایی می‌شود. دقت پایین‌تر دو نوع گفتار دیگر بیانگر شباهت آنها و خطا در GMM است. با مقایسهٔ عملکرد GMM با روش قبلی (معیار درستنمایی بیشینه با یک عنصر گوسی)، می‌توان گفت که افزایش تعداد عناصر مخلوط، دقت شناسایی گفتار با حالت را زیاد کرده است.

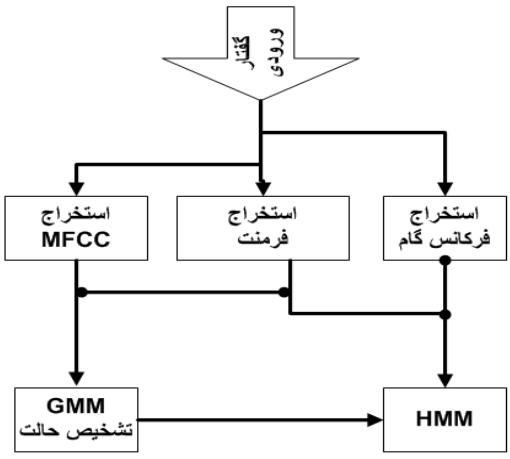
1. Gaussian Mixture Model

2. Maximum Likelihood
3. Expectation Maximization
4. Binary Splitting

جدول ۷ نتایج شناسایی حالتهای گفتاری و گفتار عادی با استفاده از GMM

حالت گفتاری	اندوه	خشم	عادی
نرخ شناسایی	۸۷/۳	۹۸/۷	۸۰/۷۲

شکل ۳ شمای کلی مجموعهٔ سیستم بازشناسی گفتار را همراه با سیستم شناسایی حالت نشان می‌دهد. به‌عبارت دیگر سیستم نهایی، ترکیبی از شناسایی‌کنندهٔ حالت همراه با سیستم بازشناسی گفتار است. همانطور که در این شکل مشاهده می‌شود از نتیجهٔ سیستم شناسایی حالت می‌توان برای انتخاب مدل بازشناسی مناسب برای به‌دست آوردن بهترین نتایج بازشناسی استفاده کرد. به‌عبارت دیگر نتایج جدول ۳ نشان می‌دهد که برای هر حالت گفتاری کدام پارامتر نوایی همراه با ضرایب کپسترال می‌توانند بیشترین بهبود را در نتایج بازشناسی ایجاد کنند. لذا نتیجهٔ ارائه شده از طرف سیستم شناسایی‌کنندهٔ حالت، می‌تواند در انتخاب مدل مناسب برای سیستم بازشناسی مفید باشد. لازم است ذکر شود که با توجه به این نکته که کار انجام شده در این تحقیق، زمان واقعی نیست در واقع فرایند در نظر گرفته شده در این شکل به‌صورت قسمتهایی جدا از هم دنبال شده است.



شکل ۳ شمای کلی سیستم بازشناسی گفتار با حالت

۵- نتیجه‌گیری

بازشناسی گفتار در دو حالت گفتاری اندوه و خشم، با استفاده از مدل‌های پایه (عادی بدون هر گونه تأکید و حالت گفتار) نشان داد که دقت بازشناسی در حالت گفتاری اندوه نسبت به گفتار عادی کاهش قابل ملاحظه‌ای دارد و در حالت خشم دقت بازشناسی به‌شدت کاهش می‌یابد. در حالت اندوه، اضافه کردن پارامترهای نوایی، خصوصاً فرمنت دوم و سوم، دقت بازشناسی گفتار را افزایش داد. همچنین شناسایی حالت گفتار با استفاده از معیار درستنمایی بیشینه و GMM صورت گرفت. بررسیها نشان داد که به‌دلیل تفاوت زیاد حالت خشم از حالت اندوه معمولاً شناسایی آنها از یکدیگر به‌راحتی صورت می‌گیرد. هر چند ممکن است در تمایز بین حالت گفتاری اندوه با گفتار عادی خطا ایجاد شود. مجموعهٔ بررسیهای صورت گرفته در این تحقیق نشان داد که برای بهبود نتایج بازشناسی گفتار با حالت، باید سیستم بازشناسی گفتار همراه با شناسایی‌کنندهٔ حالت به‌صورت توأم استفاده شوند، به‌گونه‌ای که بتوان از نتایج سیستم شناسایی‌کنندهٔ حالت برای انتخاب مدل صحیح برای بازشناسی گفتار استفاده کرد.

٦- منابع

[1] O.Pierre-Yves, "The Production and Recognition of Emotion in Speech: Features and Algorithms", Int. J. Human-Computer Studies 59, 157-183, 2003.

[2] T. Pao et al. "Detecting Emotions in Mandarin Speech", Computational Linguistics and Chinese Language Processing, Vol. 10, No. 3, Sep. 2005, pp. 347-362.

[3] M. Weintraub, K. Taussig, K. H. Smith and A. Snodgrass, "Effect of Speaking Style on LVCSR Performance", in Proc. ICSLP'96.

[4] P. Taylor, S. King, S. Isard, H. Wright and J. Kowtko, "Using Intonation to Constrain Language Models in Speech Recognition", in Proc. EUROSPEECH'97.

[5] G. Rigoll, R. Muller and B. Schuller, "Speech Emotion Recognition Expoiting and Linguistic Information Sources", in Proc. SPECOM'05.

[6] V. Hozjan and Z. Kacic, "Improved Emotion Recognition with Large Set of Statistical Features", in Proc. EUROSPEECH'03.

[7] V. A. Petrushin, "Emotion Recognition in Speech Signal: Experimental Study, Development and Application", in Proc. ICSLP'00.

[8] R. Tato, R. Santos, R. Kompe and J. M. Pardo, "Emotional Space Improves Emotion Recognition", in Proc. ICSLP'02.

[9] H. Wright, M. Desio and S. Isard, "Using High Level Dialogue Information for Dialogue Act Recognition Using Prosodic Features", in Proc. of an ESCA Tutorial Research Workshop on Dialogue and Prosody, pp. 139-143, 1999.

[10] P. Taylor, H. Shimodaira, S. Isard, S. King and J. Kowtko, "Using prosodic information to constrain language models for spoken dialogue", in Proc. ICSLP'96.

[11] F. Gallwitz, A. Batliner, J. Buckow, R. Huber, H. Niemann and E. Noth, "Integrated Recognition of Words and Phrase Boundaries", in Proc. ICSLP'98.

[12] H. Wright and P. Taylor, "Modeling Intonational Structure Using Hidden Markov Models", ESCA Workshop on Intonation: Theory, Models and Application, 1997.

[13] I. Cohen, A. Garg and T. S. Huang, "Emotion Recognition from Facial Expressions Using Multilevel HMM", Neural Information Processing Systems Conference, 2000.

[14] A. Nogueiras, A. Moreno, A. Bonafonte and J. B. Marino, "Speech Emotion Recognition Using Hidden Markov Models", in Proc. EUROSPEECH'01.

[15] T. Pao, Y. Chen, J. Yeh and W. Liao, "Detecting Emotions in Mandarin Speech", Computational Linguistics and Chinese Language Processing, Vol. 10, No. 3, pp. 347-362, September 2005.

[16] J. Toivanen, T. Seppanen and E. Vayrynen, "Automatic Recognition of Emotions in Spoken Finnish: Preliminary Results and

Applications", in Proc. Prosodic interfaces, pp. 85-89, France, 2003.

[17] J. Yuan, L. Shen and F. Chen, "The Acoustic Realization of Anger, Fear, Joy and Sadness in Chinese", in Proc. ICSLP'02.

[18] J. Yuan, C. Shih and G. P. Kochanski, "Comparison of Declarative and Interrogative Intonation in Chinese", in Proc. International Conference on Speech Prosody, 2002.

[19] O. W. Kwon, K. Chan and T. W. Lee, "Emotion Recognition by Speech Signals", In Proc. EUROSPEECH'03.

[20] B. Schuller, R. Muller, M. Lang and G. Rigoll, "Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles", in Proc. EUROSPEECH'05.

[21] Y. Kao and L. Lee, "Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language", in Proc. ICSLP'06.

[22] M. Slaney, G. McRoberts, "Baby Ears: A Recognition System for Affective Vocalizations", in Proc. ICASSP'98.

[23] S. zhang, P. C. Ching, F. Kong, "Automatic Emotion Recognition of Speech Signal in Mandarin", in Proc. ICSLP'06.

[24] B. Schuller, R. Muller, M. Lang and G. Rigoll,"Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles", in Proc. EUROSPEECH'05.

[25] B. Schuller and G. Rigoll, "Timing Level in Segment-Based Speech Emotion Recognition", in Proc. ICSLP'06.

[26] J. Cichosz, K. Slot, "Low-Dimensional Feature Space Derivation for Emotion Recognition", in Proc. EUROSPEECH'05.

[27] [27] K. P. Truong and D. A. Van Leeuwen, "Automatic Detection of Laughter", in Proc. EUROSPEECH'05.

[28] P. Outeyer, "Novel Useful Features and Algorithms for the Recognition of Emotions in Human Speech", In Proc. 1th International Conference on Speech Prosody, 2002.

[29] D. Neiberg, K. Elenius and K. Laskowski, "Emotion Recognition in Spontaneous Speech Using GMM", in Proc. ICSLP'06.

[30] I. Luengo, E. Navas, I. Hernaez and J. Sanchez, "Automatic Emotion Recognition Using Prosodic Parameters", in Proc. EUROSPEECH'05.

[31] T. Athanaselist, S. Bakamidis, I. Dologlou, R.Cowie, E. Douslas and C. Cox, "ASR for Emotional Speech: Clarifying the Issues and Enhancing Performance", Journal of Neural Network, Vol. 18, pp. 437-444, 2005.

[32] D. Gharavian and S. M. Ahadi, "Recognition of Emotional Speech and Speech Emotion in Farsi", in Proc. ISCSLP'06.

[33] D. Gharavian and S. M. Ahadi, "The Effect of Emotion on Farsi Speech Parameters: A Statistical Evaluation", In Proc. SPECOM'05.

[34] A. Paeschke, W. F. Sendlmeier, "Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements", in Proc. of the ISCA ITRW on Speech and Emotion, Newcastle, Belfast, September, 2000.

[35] I. Linnankoski et al. "Conveyance of Emotional Connotations by a Single Word in English", Speech Communication 45, 27-39, 2005.

[36] M. Bijankhan et al., "The speech database of Farsi spoken language", in Proc. SST'94.

[37] S. J. Young et al., The HTK Book (ver 3.2), Cambridge University Eng. Dept. 2002.

[38] S. S. McCandless, "An Algorithm for Formant Extraction Using Linear Prediction Spectra", IEEE Transactions on acoustics, speech and signal processing, ASSP-22, No. 2, pp. 135-141, April 1974.

[39] Y. Medan, E. Yair and D. Chazan, "Super Resolution Pitch Determination of Speech Signals", IEEE Trans. Sig. Proc., Vol. 39, No. 1, January 1991.

[40] Edinburgh Speech Tools Library, available at http://festvox.org/docs/speech_tools-1.2.0/x2152.htm.

[41] S. Zhang, P. C. Ching, F. Kong, "Acoustic Analysis of Emotional Speech in Mandarin Chinese", in Proc. ISCSLP'06.

[42] J. Yuan, L. Shen, F. Chen, "The Acoustic Realization of Anger, Fear, Joy and Sadness in Chinese", in Proc. Speech Prosody, France, 2002.

[٤٣] غرویان، داود، جانی‌پور، محسن، شیخان، منصور، «بررسی آماری نحوهٔ تغییرات فرکانس گام در گفتار با حالت زبان فارسی»، ارائه شده به کنفرانس مهندسی برق ۱۳۸۷.

[44] D. Gharavian and S. M. Ahadi, "Use of Formants in Stressed and Unstressed Continuous Speech Recognition", in Proc. ICSLP'04.

[٤٥] خیاط‌زاده، م، «شناسایی گوینده به‌صورت ناوابسته به متن با استفاده از مدلهای مخلوط گوسی»، پایان‌نامه کارشناسی ارشد، دانشکده مهندسی برق، دانشگاه صنعتی امیرکبیر، ۱۳۸۱.