بهبود عمل‌کرد سامانهٔ بازشناسی گفتار پیوسته با ویژگی‌های استخراج‌شده از مانیفولدهای گفتاری در فضای بازسازی‌شدهٔ فاز

یاسر شکفته و فرشاد الماس گنج

آزمایشگاه پردازش گفتار، دانشکدهٔ مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران

چکیده

یکی از رویکردهای مؤثّر در بهبود کارایی سامانه‌های بازشناسی گفتار، طراحی روش‌های متنوع استخراج ویژگی از سیگنال گفتار و ترکیب اطّلاعات به‌دست آمده از آنهاست. تحقیقات اخیر نشان می‌دهد که سیگنال گفتار رفتار غیرخطی و آشوبی دارد؛ ولی از این مشخّصه سیگنال گفتار در سامانه‌های بازشناسی پیوسته گفتار استفاده نمی‌شود. یکی از حوزه‌های مناسب برای نمایش مشخصه‌های پویا و غیرخطّی سیگنال آشوبی، فضای بازسازی شده فاز (RPS) است، از این‌رو در این مقاله یک روش جدید استخراج ویژگی مبتنی‌بر RPS (LLRPS) پیشنهاد شده است. این ویژگی‌ها از امتیاز شباهت تراژکتوری سیگنال گفتار جاسازی‌شده در RPS با مجموعه‌ای از مانیفولدهای واجی از پیش تعیین شده محاسبه می‌شوند. سپس مقادیر احتمال پسین واجی به‌وسیلهٔ ساختار شبکهٔ عصبی TMLP از روی ویژگی‌های LLRPS تخمین‌زده می‌شود. ساختار شبکهٔ عصبی استفاده شده، به‌صورتی است که علاوه‌بر توانایی استخراج اطّلاعات پویا، قابلیت پیاده‌سازی روش‌های متنوع ترکیب خروجی دارد. نتایج آزمایش‌ها برروی مجموعه‌دادگان گفتاری فارس‌دات نشان می‌دهد که ترکیب غیرخطّی خروجی سامانه‌های بازشناسی، شامل ویژگی‌های متداول کپستروم MFCC و ویژگی‌های پیشنهادی LLRPS، به‌ترتیب منجر به بهبود ۳/۹۴ درصد در دقّت بازشناسی قاب و ۴/۰۲ درصد در دقّت بازشناسی واج نسبت به عمل‌کرد سامانه بازشناسی پایه شده است.

واژگان کلیدی: بازشناسی گفتار پیوسته، استخراج ویژگی، فضای بازسازی شدهٔ فاز، مانیفولدهای واجی، امتیاز درست‌نمایی، شبکهٔ عصبی.

۱- مقدّمه

سامانهٔ تولید گفتار انسان شامل یک فرآیند پویای چندمتغیره است که خروجی آن به‌طور معمول به شکل یک مشاهدهٔ تک‌متغیره[1] یا تک‌بُعدی تحت عنوان سیگنال زمان-گسسته گفتار توسط میکروفون ضبط می‌شود. در فرآیند تولید سیگنال گفتار، به علّت بروز برخی عوامل مانند ارتعاش غیرخطّی تارهای صوتی و یا حرکات لایه‌ای هوا در مجرای صوتی، تظاهرات آشوب‌گونه[2] در شکل سیگنال گفتار مشاهده شده است (Awrejcewicz, 1990; Berry, 1994; Herzel,1995; Jiang, 2002; Jiang, 2006). از این‌رو با توجّه به خواص سیگنال‌های آشوب‌گونه، سیگنال گفتار را به‌صورت یک سری زمانی با توصیف چندمتغیره، در فضای حالت[3] یا فضای فاز[4] می‌توان بیان نمود (Kokkinos, 2005; Hagmuller, 2006; Sun, 2007). این انتقال سیگنال به فضای حالت می‌تواند به‌صورتی انجام شود که رفتار تراژکتوری متناظر با آن، از لحاظ هندسی معادل با تراژکتوری واقعی آن باشد (Kantz, 1997).

در همین زمینه، *تیکنز* روش‌های جاسازی فضایی[5] در حوزه‌ای مشابه با فضای حالت (که فضای بازسازی شدهٔ فاز[6] (RPS) نامیده می شود) را به‌عنوان یکی از تکنیک‌های مطالعه مشخصه‌های پویای غیرخطی و آشوبی معرّفی کرد

[1] Univariate
[2] Chaotic

[3] State Space
[4] Phase Space
[5] Embedding Space
[6] Reconstructed Phase Space

(Takens, 1980). بر مبنای این روش‌ها، ثبت یک متغیر زمانی از رفتار سامانه پویا می‌تواند برای کسب اغلب خواص پویایی[1] آن سامانه "کافی" باشد (Kantz, 1997). درنتیجه با انتقال سیگنال گفتار به حوزهٔ RPS، می‌توان امید داشت که اطّلاعات بیشتر و متمایزکننده‌تری از آن نسبت به دیگر روش‌های متداول استخراج کرد.

از همین اواخر تحقیقات متعددی در حوزهٔ پردازش سیگنال گفتار برمبنای استفاده از روش جاسازی سیگنال گفتار در RPS انجام شده است (Narayanan, 1995; Lindgren, 2004). اساس این تحقیقات، نمایش و استخراج اطّلاعات مربوط به حوزهٔ RPS سیگنال گفتار است که می‌تواند حاوی اطّلاعات متمایزکننده‌تری نسبت به روش‌های متداول آنالیز سیگنال گفتار (مانند روش‌های استخراج ویژگی MFCC و PLP که مبتنی بر طیف هستند Davis, 1980; Hermansky, 1990)) باشد. بخشی از این اطّلاعات متمایز متأثّر از مشخّصه‌هایی است که به‌واسطه حذف اطّلاعات فاز در روش‌های استخراج ویژگی مبتنی‌بر اندازهٔ طیف سیگنال از بین رفته‌اند. با اینکه برخی از تحقیقات الهام گرفته‌شده از دستگاه شنوایی انسان ادعا می‌کنند که دستگاه شنوایی انسان به اطّلاعات فاز سیگنال حساس نیستند؛ امّا برخی تحقیقات منتشریافتهٔ اخیر، نشان دهندهٔ مؤثّر بودن اطّلاعات فاز در سامانهٔ بازشناسی گفتار است (Paliwal, 2005; Hegde, 2007; Alsteris, 2007).

در حوزهٔ پردازش سیگنال گفتار، برخلاف انجام تحقیقات متعدد با روش جاسازی، متأسفانه تاکنون روش‌های محدودی برای استفاده از قابلیت آن در سامانه‌های بازشناسی گفتار پیوسته ارائه شده است. عمده فعالیت‌های صورت‌گرفته در این حوزه، شامل استخراج ویژگی‌های آشوب‌گونه مانند بُعد هم‌بستگی[2]، بُعد فرکتال[3]، بُعد فرکتال تعمیم‌یافته[4] و نماهای لیاپانوف[5] برای رده‌بندی واجی است (Pitsikalis, 2002&2009; Vaziri, 2010). در (Yu, 2006) دو مشخّصه جدید از سیگنال جاسازی‌شده در RPS معرفی شده است. این ویژگی‌ها مقادیر عددی هستند که مقدار جابه‌جایی و دایروی‌بودن مسیر تراژکتوری سیگنال در فضای فاز را بازنمایی می‌کنند.

برای اوّلین‌بار در سال ۲۰۰۲، از توزیع آماری سیگنال گفتار جاسازی‌شده در RPS با استفاده از روش هیستوگرام، برای رده‌بندی واج‌های مجزا[6] در مجموعه دادگان TIMIT استفاده شد (Ye, 2002). سپس در (Ye, 2003) استفاده از آنالیز مؤلّفه‌های اساسی (PCA) در متعامدکردن و کاهش بُعد سیگنال جاسازی شده در فضای فاز بررسی شد. در ادامه با استفاده از تحلیل پویای تولید گفتار در RPS، روشی برای رده‌بندی واکه‌ها بر مبنای فاصلهٔ اندازه‌گیری شده بین جاذب‌های[7] هر واج در فضای فاز مطرح شد (Liu, 2003). همچنین *لیندگرن* با استفاده از ویژگی‌های استخراج‌شده از سیگنال گفتار جاسازی شده در فضای RPS، یک سامانه بازشناسی گفتار واج مجزا بر روی مجموعه‌دادگان TIMIT معرّفی کرد (Lindgren, 2003). مشابه این سامانه، از یک مجموعه مدل مخلوط گوسی[8] (GMM) برای مدل‌سازی جداگانهٔ توزیع هر واج استفاده شده است (Povinelli, 2004&2006). در (Indrebo, 2006) یک روش جدید ترکیبی با استفاده از فضای بازسازی‌شدهٔ فاز اعمال شده بر روی زیرباندهای به‌دست آمده از تجزیهٔ فیلتر بانکی سیگنال، برای طبقه‌بندی سیگنال گفتار ارائه شد. از طرف دیگر در (Narayanan 2012; Thasleema 2012) نیز طبقه‌بندی واحدهای CV گفتاری با توسعهٔ روشی مبتنی بر هیستوگرام (SSPD) و طبقه‌بندی‌کننده‌های شبکهٔ عصبی مصنوعی (ANN) و ماشین بردار پشتیبان (SVM) پیشنهاد شده است. در (جعفری، ۱۳۸۹) و (Jafari, 2010; Jafari, 2012) چند نوع بردار ویژگی ترکیبی پیشنهاد شده است که عناصر آن مخلوطی از ویژگی‌های متداول MFCC و ویژگی‌های خاص استخراج شده از RPS هستند. در این روش، ویژگی‌های مبتنی بر RPS از پارامترهای GMM یادگرفته شده از توزیع سیگنال جاسازی شده در RPS و یا قطع پوانکارهٔ[9] آن به دست آمده‌اند. سپس با انتخاب برخی پارامترهای مدل GMM تعلیم‌یافته، بردار ویژگی نهایی تولید شده است. متأسفانه هزینهٔ محاسباتی بالا در پیاده‌سازی این روش‌ها یکی از نقاط ضعف اساسی آنها به‌شمار می‌رود.

درهمین اواخر در (Shekofteh, 2013) روشی پیشنهاد شد تا ویژگی‌های مناسبی از RPS استخراج شود که علاوه‌بر هزینهٔ محاسباتی و کارایی مناسب آن، قابلیت

[1] Dynamic
[2] Correlation Dimension
[3] Fractal Dimension
[4] Generalized Fractal Dimension
[5] Lyapunov Exponents
[6] Isolated Phoneme
[7] Attractor
[8] Gaussian Mixture Model
[9] Poincare Section

استفاده در سامانه‌های بازشناسی گفتار پیوسته را داشته باشند. این روش برمبنای جاسازی سیگنال زمانی گفتار در فضای RPS (که علاوه‌بر اطّلاعات دامنه و فاز سیگنال حاوی اطّلاعات مانیفولدهای گفتاری است) بود که در آن بایستی یک‌سری مدل‌های از پیش‌تعلیم یافتهٔ گفتاری مرجع آماده شوند. این مدل‌ها که می‌توان آنها را "مانیفولد" یا "جاذب" گفتاری تلقی کرد، وظیفهٔ مدل‌سازی توزیع نمونه‌های هر یک از واحدهای گفتاری واجی را در RPS برعهده دارند. سپس برای هر قاب گفتاری، ویژگی‌های PPRPS، بر مبنای رابطهٔ بیز و محاسبهٔ مقدار شباهت (درست‌نمایی شرطی[1]) نمونه‌های قاب گفتاری جاسازی شده در RPS با هر یک از مدل‌های مرجع محاسبه شده است. بنابراین بُعد بردار ویژگی PPRPS برابر با تعداد کل مدل‌های مرجع گفتاری است. در (Shekofteh, 2013) نشان داده شد که استفاده از این ایده منجر به بهبود عمل‌کرد سامانه بازشناسی واج مجزا در مقایسه با روش ارائه شده در (Povinelli, 2006) می‌شود. همچنین کاربرد این روش برای استفاده در سامانه بازشناسی گفتار پیوسته با مدل مخفی مارکوف[2] (HMM) و با مجموعهٔ مجموعهٔ محدودی از مانیفولدهای انتخاب شدهٔ واجی بررسی شده است. برای انجام این‌کار نگاشت خطّی آنالیز متمایزگر (LDA) بر روی تعدادی ویژگی اعمال شده است.

در (Furui, 1986) نشان داده شده است که با توجّه به ماهیت پویای گفتار، استخراج اطّلاعات دینامیک از توالی بردارهای ویژگی گفتاری می‌تواند بهبود چشم‌گیری در کارایی سامانه‌های بازشناسی گفتار داشته باشد. متاسفانه یکی از مشکلات روش پیشنهادی در (Shekofteh, 2013) عدم تولید ویژگی‌های مناسب پویا در مقایسه با ویژگی‌های پویای به دست آمده از روش‌های متداولی مانند MFCC (ضرایب دلتا و دلتادلتا) است. از این‌رو در این مقاله برای استخراج مناسب‌تر اطّلاعات پویا، استفاده از ساختار شبکهٔ عصبی با ورودی قطعه‌ای[3] پیشنهاد شده است که شامل توالی چندین بردار ویژگی از قاب‌های گفتاری است. همچنین ساختار شبکهٔ عصبی مورد استفاده منجر به یادگیری الگوهای زمانی[4] و ترکیب غیرخطی تمامی ویژگی‌های ورودی شده و تخمین مناسبی از احتمال پسین طبقه‌های واج[5] گفتاری در خروجی خود تولید خواهد کرد.

[1] Conditional Likelihood
[2] Hidden Markov Model
[3] Segmental
[4] Temporal Pattern
[5] Phoneme Classes

به‌علاوه در این مقاله عمل‌کرد روش پیشنهادی در حالت استفاده از اطّلاعات تمامی مانیفولدهای گفتاری واج بررسی شده است. درنهایت بوسیله سامانه بازشناس مبتنی بر شبکهٔ عصبی و تخمین خروجی‌های احتمال پسین آن، روش‌های متنوعی از ترکیب[6] خروجی سامانه‌ها و ساختار TANDEM در جهت بهبود کارایی ASR بررسی شده است.

ادامهٔ مقاله به‌این صورت شکل یافته است: در بخش دوم تئوری جاسازی سیگنال در RPS معرّفی خواهد شد. بخش سوم مدل‌سازی مانیفولدهای گفتاری را شرح می‌دهد. در بخش چهارم روش استخراج ویژگی پیشنهادی آورده شده است. بخش پنجم و ششم به‌ترتیب روش بازشناسی و مجموعه دادگان مورد استفاده را معرّفی می‌کنند. در بخش هفتم نتایج آزمایش‌های مربوط به سامانه بازشناس پایه ارائه شده است. بخش‌های هشتم و نهم به‌ترتیب شامل معرّفی روش‌های ترکیب و ارائه نتایج آزمایش‌ها است. در انتها نیز نتیجه‌گیری مقاله آورده شده است.

۲- جاسازی سیگنال در فضای بازسازی‌شدهٔ فاز

جاسازی سیگنال تک‌متغیرهٔ زمانی به‌وسیلهٔ روش محورهای تأخیری[7] برمبنای تئوری‌های مطرح شده در (Takens, 1980) و (Sauer, 1991) است. در این روش، پس از تعیین پارامترهای مناسب تأخیر زمانی و بُعد جاسازی، نمونه‌های سیگنال زمانی اولیه را می‌توان به گونه‌ای به فضای چند بُعدی RPS انتقال داد که مشخصه پویای واقعی سامانه تولید کنندهٔ سیگنال، به‌صورت مناسب در آن فضا نمایش داده شود (Kantz, 1997). در ادامه، روش جاسازی سیگنال یک‌بُعدی در فضای با بُعد بالاتر بیان می‌شود:

فرض کنیم که سری زمانی s ($s=\{s[1],..,s[N]\}$) (نمونه‌های سیگنال زمانی یک قاب گفتاری) با تعداد N نمونه، سیگنال زمانی تک‌متغیره مورد بحث باشد. متناظر با این سری زمانی، یک مجموعه متوالی از نقاط جاسازی‌شده در فضای چندبُعدی RPS تعریف می‌شود. به‌عنوان مثال می توان S[i] را به‌عنوان نمونه iام جاسازی‌شده در فضای RPS با رابطهٔ (۱) به‌صورت یک بردار d بُعدی نشان داد:

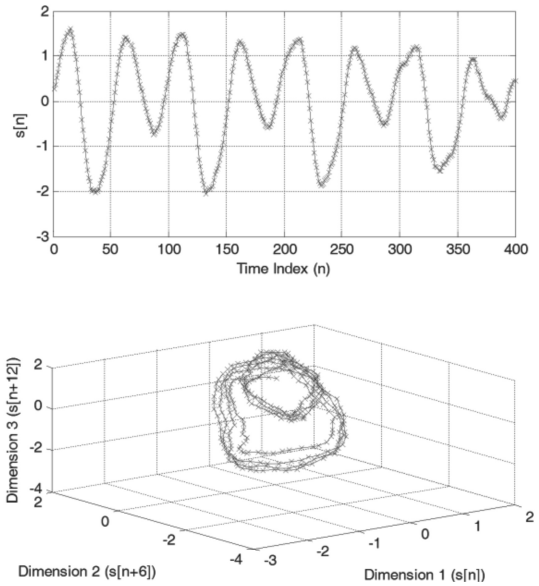
$$\mathbf{S}[i] = \big[s[i], s[i+\tau], \ldots, s[i+(d-1)\tau]\big], \qquad (۱)$$

[6] Combination
[7] Delay-Coordinat

به‌طوری‌که پارامتر **d**، بُعد جاسازی و پارامتر τ معادل با زمان تأخیر است. واضح است که برای یک قاب با طول **N**، تعداد نقاط جاسازی‌شده در حوزهٔ RPS به‌صورت زیر است:

$$L = N - (d-1)\tau, \qquad (۲)$$

روش‌های مختلفی برای انتخاب مقدار بهینهٔ بُعد جاسازی و تأخیر زمانی وجود دارند (Kennel 1992; Abarbanel, 1996; Povinelli, 2004; Johnson, 2005). یک انتخاب مناسب برای سیگنال گفتار میکروفونی با نرخ نمونه‌برداری شانزده کیلوهرتز، مقادیر d=8 و $\tau = 6$ است (Jafari, 2010). در شکل (۱) نمونه‌ای از یک قاب سیگنال گفتار (شامل واج/u/) و همچنین تراژکتوری متناظر با آن را که در فضای سه بُعدی RPS جاسازی شده‌اند، نشان می‌دهد.



(شکل ۱) – واکه /u/ در نمایش معمولی تک‌بُعدی (شکل بالا) و نمایش تراژکتوری سه‌بُعدی آن (شکل پایین) با روش جاسازی در RPS.

رفتار تراژکتوری سیگنال گفتار در فضای **RPS** برای اکثر واج‌های واکدار مشابه با فرآیند قبض[1] در سیگنال‌های آشوبی است. به‌عنوان مثال در شکل (۱) نمونه‌ای از این رفتار برای تراژکتوری واج واکدار /**u**/ نشان داده شده است. نتیجهٔ این رفتار تولید مانیفولدهایی با الگوهای خاص و مشخّص در RPS خواهد بود. از طرف دیگر برای واج‌های گفتاری انفجاری (مانند واج های /b/ و /t/) رفتار بسط[2] سیگنال آشوبی در فضای فاز قابل مشاهده است. برای سایر کلاس‌های واجی رفتاری بین موارد اشاره شده امکان‌پذیر است (Banbrook, 1996; Indrebo, 2006).

۳- مدل‌سازی مانیفولدهای گفتاری در فضای بازسازی‌شدهٔ فاز

در (Povinelli, 2006) یک سیستم بازشناسی واج مجزا[3] براساس مدل‌سازی آماری توزیع نقاط تراژکتوری سیگنال گفتار در RPS توسط *پاوینلی* پیشنهاد شده است. توالی زمانی تراژکتوری یک سیگنال در حوزه RPS منجر به تشکیل "مانیفولد گفتاری" می‌شود.

سیستم بازشناسی واج مجزای پیشنهاد شده توسط پاوینلی شامل دو مرحلهٔ تعلیم و آزمون بود. در مرحلهٔ تعلیم، برای هر یک از واج‌های گفتاری در RPS، یک توزیع چندمتغیره پارامتری با استفاده از GMM آموزش داده می‌شد. برای این‌منظور در ابتدا متناظر با هر واج، کلیه سیگنال‌های مربوط به آن واج (از مجموعه داده‌های تعلیم) در فضای RPS جاسازی‌شده و سپس برای توزیع نقاط تشکیل شده در RPS یک مدل کلی GMM تعلیم داده می‌شد. استفاده از GMM باعث می‌شد که ساختار هندسی مانیفولد شکل‌یافته در RPS به صورت نَرم مدل شودبه‌عنوان مثال تصویری از مدل‌سازی GMM مانیفولد واج واکدار /u/ در RPS دو بعدی در شکل (۲) نشان داده شده است.

پس از تعلیم مدل‌های GMM که دربردارنده اطّلاعات هندسی مانیفولد هر واج در RPS است؛ پاوینلی از معیار بیشینه درست نمائی[4] برای طبقه‌بندی در مرحلهٔ آزمون استفاده کرده است که در رابطه (۳) آورده شده است:

$$\hat{c} = \arg \max_{i=1,\ldots,K} \; p(\mathbf{X} \mid c_i) \;,$$
$$p(\mathbf{X} \mid c_i) = \frac{1}{L} \sum_{n=1}^{L} \log \; p(\mathbf{x}_n \mid c_i) \;, \qquad (۳)$$

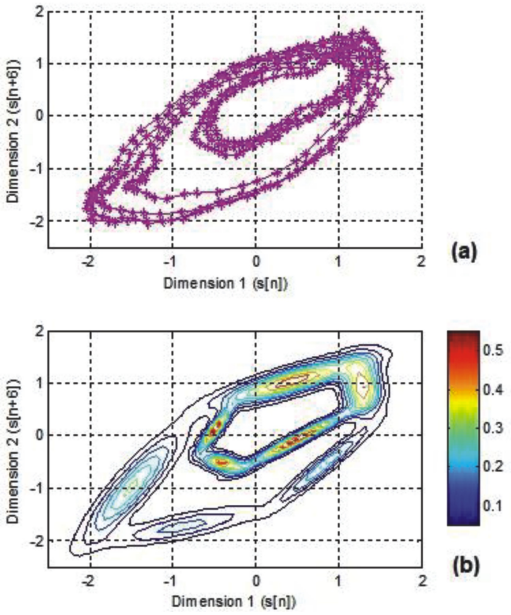
که در آن K تعداد کل مدل‌های واجی GMM (که هر کدام متناظر با یک کلاس واجی c_i هستند) و L تعداد نقاط جاسازی‌شده در RPS است. ماتریس X نیز شامل تعداد L نمونه جاسازی‌شده $\mathbf{x}_n$ در فضای RPS است که معمولاً حاوی اطّلاعات ماتریس تراژکتوری S و مشتقات زمانی آن، **ΔS**، می‌باشد (Lindgren, 2003). همچنین مقدار درست

[1] Folding Or Squeezing
[2] Stretching
[3] Isolated Phoneme Recognition
[4] Max-Likelihood

نمایی $p(\mathbf{x_n} \mid c_i)$ برای تعداد M گوسین مدل مخلوط گوسی (GMM) از رابطه (۴) محاسبه می‌شود:

$$p(\mathbf{x_n} \mid c_i) = \sum_{m=1}^{M} \omega_m^i N(\mathbf{x_n}; \mu_m^i, \Sigma_m^i), \qquad (۴)$$

که در آن $N(\mathbf{x}_n; \mu_m^i, \Sigma_m^i)$ نشان‌دهنده درست‌نمایی مؤلّفه **m**ام توزیع گوسی با وزن ω_m^i از کلاس واج **i**ام (c_i) است.



(شکل ۲) - (a) توزیع تراژکتوری مانیفولد واج /u/ نشان‌داده شده در شکل (۱) در فضای RPS دوبُعدی. (b) مدل‌سازی تابع چگالی احتمال آن به‌وسیلهٔ یک مدل GMM با تعداد هشت مؤلّفهٔ گوسی.

جانسن در (Johnson, 2005) نشان داد که قدرت تمایزگری بین واج‌ها که از فضای RPS به‌دست می‌آید، کمتر از ویژگی‌های متداول کپستروم MFCC است؛ امّا با استفاده از روش ترکیب جمع وزن‌دار در سطح خروجی رده‌بندی‌کننده‌ها (ادغام امتیاز درست‌نمایی به‌دست آمده از ویژگی‌های حوزه RPS با امتیاز درست‌نمایی حاصل شده از ویژگی‌های MFCC) کارایی سامانه بازشناسی واج مجزا بر روی مجموعه داده TIMIT افزایش خواهد یافت. این افزایش کارایی مؤیّد هم‌افزا بودن اطّلاعاتی است که توسط روش استخراج ویژگی MFCC و ویژگی‌های حوزه RPS به‌دست آمده‌اند.

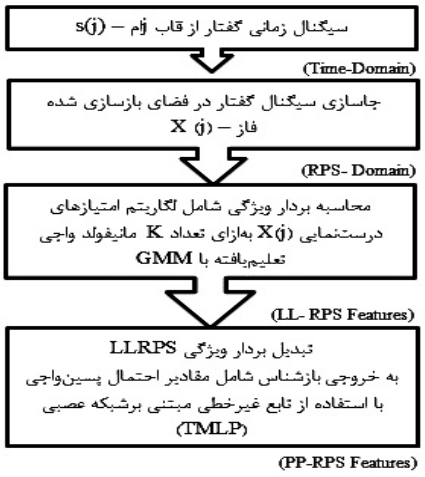
۴- محاسبه تخمین احتمال پسین واجی در بازشناسی گفتار به‌وسیلهٔ ویژگی‌های مبتنی بر RPS

در این مقاله، مبنای روش استخراج ویژگی مبتنی بر RPS، براساس ایده‌ای در توسعهٔ روش *پاوینلی* است. متأسفانه روش *پاوینلی* به‌انحصار قابلیت استفاده در کاربرد بازشناسی واج مجزا دارد. اما با روش توسعه‌ای مطرح شده در (Shekofteh, 2013)، توانایی استفاده از این روش در کاربرد بازشناسی گفتار پیوسته حاصل شده است.

همان‌طور که در رابطهٔ (۳) در بخش (۳) دیدیم، در مرحلهٔ آزمون سامانه *پاوینلی*، به تعداد کلاس‌های واجی موجود، امتیاز آکوستیکی درست‌نمایی تولید خواهد شد. سپس مدل واجی که بالاترین امتیاز درست‌نمایی را به دست آورده باشد، به‌عنوان کلاس واج برنده انتخاب می‌شود. بنابراین در بخش تصمیم‌گیری روش پاوینلی، از اطّلاعات فضای RPS هر مدل واجی به‌طور مستقل در تعیین طبقه واجی برنده استفاده می‌شود.

ایده مطرح‌شده در توسعهٔ این روش، پیشنهاد استفاده هم‌زمان از زیرمجموعه‌ای از امتیازهای درست‌نمایی به دست آمده از مانیفولدهای واجی، به عنوان یک بردار ویژگی مناسب است. با این روش می‌توان به‌راحتی از اطّلاعات مانیفولدهای گفتاری در RPS در پیاده‌سازی سامانه‌های بازشناسی گفتار پیوسته استفاده کرد.

از آنجا که هر یک از مؤلّفه‌های بردار ویژگی به‌دست آمده از RPS دارای اطّلاعات هم‌بسته با یکدیگر هستند، انتخاب یا ترکیب مناسب این ویژگی‌ها برای تولید اطّلاعات غنی‌تر مفید به‌نظر می‌رسد. یک روش مناسب برای دستیابی به این هدف، تبدیل بردارهای ویژگی اوّلیه به تخمین مناسبی از احتمالات پسین واجی است که درهمین‌اواخر به طور گسترده در کاربرد بازشناسی گفتار مطرح‌شده است (Park, 2011; Pinto, 2011; Ikbal, 2012). در اینجا استفاده از ساختار یک شبکهٔ عصبی که قابلیت یادگیری الگوهای زمانی را دارد، برای تخمین احتمالات پسین پیشنهاد شده است. در شکل (۳) روندنمای روش پیشنهادی آورده شده است:



(شکل ۳) – روندنمای روش پیشنهادی

روندنمای شکل (۳) شامل مراحل زیر است:
مرحلهٔ اوّل: جاسازی سیگنال زمانی گفتار در RPS؛

در اینجا برای فرآیند جاسازی سیگنال در فضای RPS، از مقادیر d=8 و $\tau = 6$ استفاده شده است. به‌طور متداول نمونه‌های سیگنال قبل از جاسازی در RPS، بهنجار می‌شوند (Povinelli, 2004). اعمال روش‌های هنجارسازی باعث می‌شود که اثر شدت یا همان اندازهٔ دامنهٔ سیگنال گفتار ضبط‌شدهٔ کم‌رنگ شده و تنها شکل هندسه سیگنال مورد بررسی قرار گیرد. از طرفی دیگر، اعمال روش‌های هنجارسازی منجر به قرارگیری مانیفولدهای گفتاری در محدوده نسبتاً مشخصی از RPS خواهند شد که باعث مقاوم‌سازی[1] ویژگی‌های استخراج شده از آن می‌شود. در این مقاله نمونه‌های هر قاب گفتاری به‌گونه‌ای بهنجار می‌شوند که مقدار میانگین نمونه‌ها صفر و انحراف معیار آن‌ها برابر با مقدار واحد شود.

مرحله دوم: محاسبه بردار ویژگی اولیه شامل امتیاز درست نمایی متناظر با هر مدل مانیفولد واجی مبتنی بر GMM؛

در این مرحله، از رابطهٔ (۳) برای محاسبه امتیازات درستنمایی ازهر مدل مانیفولد واجی برای سیگنال قاب ژام گفتاری جاسازی‌شده در RPS، ***X(j)*** استفاده خواهد شد. بنابراین بایستی مشابه با روش پاوینلی، در یک مرحله جداگانه تعلیم، به تعداد واج‌های مجموعه‌دادگان (K=30) مدل GMM از روی داده‌های تعلیم آموزش داده‌شود. در (Shekofteh, 2013) نشان دادیم که استفاده از GMM با ۴ مؤلّفهٔ گوسی (M=4) مصالحهٔ مناسبی بین زمان پردازش (حجم محاسبات) و دقّت سامانه بازشناسی خواهد بود. پس از محاسبه امتیازات درستنمایی، بردار ویژگی اولیه LLRPS***(j)*** برای قاب گفتاری ژام به‌صورت زیر حاصل می‌شود:

$$LLRPS(j) = \begin{bmatrix} p(\mathbf{X}(j) \mid c_1) \\ p(\mathbf{X}(j) \mid c_2) \\ \vdots \\ p(\mathbf{X}(j) \mid c_K) \end{bmatrix}. \qquad (۵)$$

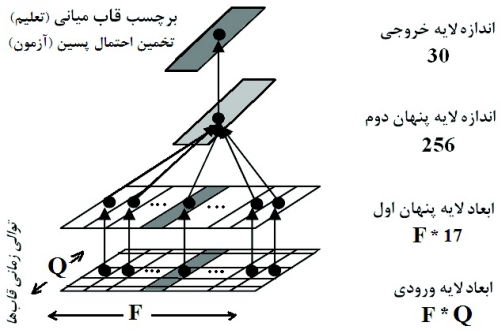
از آنجا که ماهیت ویژگی‌های تولیدشده مبتنی بر لگاریتم درستنمایی[2] (LL) است، به‌اختصار آن را LLRPS می‌نامیم.
مرحلهٔ سوم: تخمین مقادیر احتمال پسین واجی با استفاده از ساختار شبکهٔ عصبی؛

در این مرحله از یک ساختار شبکهٔ عصبی جلوسو برای ترکیب غیرخطی تمامی ویژگی‌های LLRPS و تخمین مقادیر احتمال پسین[3] واجی استفاده شده است. همچنین ساختار شبکهٔ مورد استفاده، به‌گونه‌ای طراحی می‌شود که ورودی آن قطعه‌ای و شامل چند بردار ویژگی متوالی باشد تا بتواند به‌طور هم‌زمان اطّلاعات پویای گفتاری و الگوهای زمانی مربوط به آن را از روی توالی بردارهای ویژگی ایستا[4] متوالی استخراج کند.

نحوهٔ تعلیم شبکهٔ عصبی هم به‌گونه‌ای است که در خروجی آن تخمینی از مقدار احتمال پسین (PP) طبقه واجی متناظر با بردار ویژگی ورودی (LLRPS) از قاب گفتاری میانی قطعهٔ ورودی به شبکهٔ عصبی را تولید کند. از این‌رو بردار تولیدشده در خروجی شبکه را PPRPS می‌نامیم. در این مقاله برای به‌دست آوردن تخمین احتمال پسین، از مدل شبکهٔ عصبی TMLP[5] استفاده شده است که ساختار آن الهام گرفته شده از خاصیت *تونوتوپیک* دستگاه شنوایی انسان است (Chen, 2005). ساختار این شبکه باعث می‌شود که توالی زمانی هر یک از ویژگی‌های ورودی در لایه‌های پایینی شبکه به‌طور مستقل پردازش شده و سپس اطّلاعات به‌دست آمده از آنها در لایه‌های بالاتر آن ترکیب شوند. این ساختار شبکهٔ عصبی به یادگیری بهتر الگوهای زمانی از هر یک از ویژگی‌های ورودی به شبکه کمک می‌کند. در (شکفته، ۱۳۸۶) نشان داده شد که قدرت یادگیری این مدل تمایزی[6] از شبکه‌های عصبی متداول MLP و یا TDNN

[1] Robustness
[2] Log Likelihood
[3] Posterior Probability
[4] Static
[5] Tonotopic Multi Layer Perceptron
[6] Discriminative

(رحیمی نژاد، ۱۳۸۲) در کاربرد بازشناسی گفتار بیشتر است. بنابراین در این مقاله از ساختار شبکهٔ عصبی TMLP استفاده شده است که در شکل (۴) نمایش داده شده است.



(شکل ۴) - ساختار شبکهٔ عصبی TMLP با توالی Q=21 قاب گفتاری و بردار ویژگی F بُعدی ورودی برای هر قاب گفتاری.

در ساختار شبکهٔ عصبی TMLP مورد استفاده، یک قطعه از بردارهای ویژگی متوالی (در اینجا تعداد Q قاب متوالی گفتاری) به ورودی شبکه اعمال خواهند شد تا طبقه واجی متناظر با قاب میانی ورودی آن به‌وسیلهٔ یک برچسب باینری در خروجی شبکهٔ عصبی پیش‌بینی شود. همچنین هر یک از بردارهای ویژگی ورودی دارای بُعد F (در اینجا F=K=30 و معادل با تعداد واج‌های دادگان می‌باشد) است. ساختار شبکهٔ عصبی TMLP، برای توالی Q=21 قاب از بردارهای ویژگی ورودی LLRPS (با بُعد ۳۰) به‌صورت زیر در نظر گرفته شده است:

30*(21*17)-30*(17*256)-256*30

که در برگیرندهٔ اطّلاعات پویا از مجموعهٔ ۲۱ قاب گفتاری متوالی (بیش از ۲۰۰ میلی‌ثانیه) است. انتخاب اندازهٔ لایه‌های پنهان شبکه براساس بیشینه‌سازی دقّت بازشناسی بر روی داده‌های اعتبارسنجی[1] و رعایت نسبت تعداد پارامترهای شبکه به تعداد پارامترهای دادهٔ تعلیم (برای جلوگیری از برازش بیش از حد[2] شبکه) انجام شده است.

در اینجا برای تعلیم شبکه TMLP، از برچسب‌دهی کدشدهٔ باینری نوع سخت (One-Hot) استفاده شده است (رحیمی نژاد، ۱۳۸۲؛ ولی، ۱۳۸۵). در این نوع برچسب‌دهی، یک خروجی ۳۰ نرونی (به تعداد طبقه‌های واجی) به‌عنوان خروجی مطلوب شبکه تعریف می‌شود که یک نرون آن حاوی مقدار یک (متناظر با برچسب طبقه واجی مربوط به قاب میانی مجموعه بردار ورودی) و بقیهٔ نرون‌های آن مقدار صفر دارند (Tohidypour, 2012). اگر الگوریتم تعلیم شبکه براساس کمینه‌سازی میانگین مجذور خطا[3] (MSE) باشد، تخمینی از مقدار احتمالاتی پسین طبقه‌های واجی در خروجی شبکه به شرط قطعه قاب‌های گفتاری ورودی اعمال شده به آن تولید خواهد شد (White, 1989; Zavaliagkos, 1994).

۵- روش بازشناسی گفتار

با توجّه به استفادهٔ روش پیشنهادی از مدل شبکهٔ عصبی به‌عنوان یک تخمین‌زنندهٔ مقدار احتمال پسین واجی، در این حالت سامانه بازشناسی گفتار می‌تواند در چارچوب‌های متنوعی پیاده‌سازی شود که در ادامه به شرح هر یک از آنها خواهیم پرداخت:

۵-۱- بازشناسی گفتار به‌وسیلهٔ مقدار احتمالات پسین واجی تخمین زده شده در خروجی مدل شبکهٔ عصبی

در این حالت ارزیابی عمل‌کرد سامانه بازشناسی گفتار با دو معیار دقّت بازشناسی در سطوح آوایی قاب (frame) و واج (phoneme) انجام خواهد شد. دقّت بازشناسی قاب، به‌طور مستقیم از خروجی احتمال پسین به‌دست آمده از هر قاب گفتاری قابل محاسبه است؛ به این صورت که طبقه واجی متناظر با نرونی از خروجی شبکه که بیشترین مقدار احتمال پسین را داشته باشد، به‌عنوان برچسب واج بازشناسی‌شده برای قاب متناظر با آن در ورودی شبکه درنظرگرفته می‌شود. از این رو معیار درصد دقّت بازشناسی قاب به‌صورت نسبت تعداد قاب‌های گفتاری درست تشخیص داده شده به تعداد کل قاب‌های گفتاری مورد استفاده تعریف می‌شود (Tohidypour, 2012).

برای محاسبه دقّت بازشناسی در سطح آوایی واج، از اعمال الگوریتم ویتربی بر روی بردارهای متوالی خروجی شبکهٔ عصبی که حاوی تخمینی از مقدار احتمال پسین هستند مشابه با (Pinto, 2011) استفاده خواهد شد. در این روش فرض می‌کنیم که هر واج با یک مدل چندحالته HMM مدل شده است که مقدار درست‌نمایی انتشاری[4] تمامی حالات[5] آن با یکدیگر یکسان و برابر با مقدار خروجی مقیاس[6] شده، احتمال پسین به‌دست آمده از نرون خروجی

[1] Validation
[2] Over-Training
[3] Mean Square Error
[4] Emission Likelihood
[5] States
[6] Sclaed

شبکهٔ عصبی متناظر با آن واج است. از این‌رو می‌توان این روش را مشابه با پیاده‌سازی روش هایبرید HMM/ANN تلقی کرد (Bourlard, 1994).

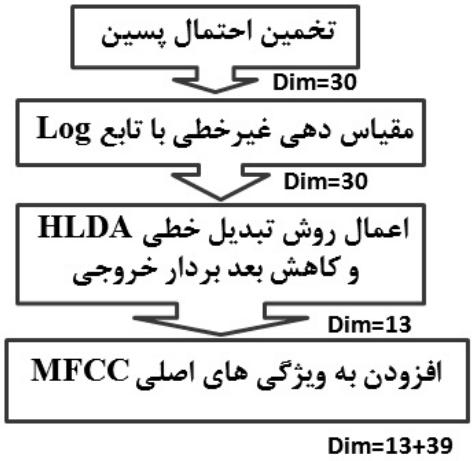
در خروجی الگوریتم ویتربی، دنباله‌ای از چند واج متوالی تولید خواهد شد که بالاترین احتمال وقوع را دارند. در اینجا برای گذر بین واج‌های مختلف، مقدار احتمال یکسان درنظرگرفته شده است که معادل با عدم استفاده از مقدار احتمالات مدل زبانی بایگرم[1] واجی در سامانه بازشناسی گفتار پیوسته است. همچنین مقدار بهینه جریمه درج[2] واج در الگوریتم ویتربی، با داده‌های توسعه[3] تعیین شده است. وظیفه جریمه درج، جلوگیری از پرش سریع خروجی الگوریتم بین واج‌های مختلف است که منجر به کاهش تعداد درج واج و درنتیجه افزایش مقدار دقّت بازشناسی خواهد شد. این فرآیند با اضافه‌کردن یک جریمه (در اینجا یک مقدار منفی) به مقدار لگاریتم درست‌نمایی انباشته شده واج‌هایی غیر از واج شناسایی شده فعلی انجام می‌شود. با استفاده از معیار **NIST**، درصد دقّت بازشناسی واج نهایی از رابطهٔ زیر محاسبه می‌شود:

$$\%Phoneme\ Accuracy = \frac{A-D-I-S}{A} * 100 \qquad (۶)$$

که در آن A تعداد واج‌های موجود در برچسب داده مرجع، D تعداد واج‌های حذف‌شده، I تعداد واج‌های درج‌شده و S تعداد واج‌های جانشین شده است.

۵-۲- بازشناسی گفتار با مدل مخفی مارکوف (HMM) در ساختار TANDEM

در روش TANDEM، از مقادیر تخمین احتمال پسین واجی که در خروجی شبکهٔ عصبی تولید شده است، به‌عنوان بردار ویژگی خام و اولیه برای تولید ویژگی گفتاری مناسب در مدل‌سازی صوتی مبتنی بر GMM/HMM استفاده می‌شود (Hermansky, 2000; Zhu, 2004; Ikbal, 2012). در شکل (۵) ساختار TANDEM مورد استفاده در این مقاله برای تولید بردار ویژگی گفتاریِ پس‌پردازش شده نشان داده شده است.



(شکل ۵) - بلوک دیاگرام ساختار TANDEM در تولید بردار ویژگی گفتاری مورد استفاده در مدل‌سازی صوتی GMM/HMM.

در اینجا پیاده‌سازی مدل‌سازی صوتی GMM/HMM توسط جعبه ابزار HTK انجام شده است (HTK, V.3.4). همچنین از واحدهای آوایی سه واجی[4] گره زده شده برای مدل‌سازی صوتی استفاده شده است (Sameti, 2011). در این حالت ارزیابی عمل‌کرد سامانه در سطح آوایی واج و با معیار NIST (رابطه ۶) انجام خواهد شد.

۶- مجموعه دادگان

دادگان گفتاری مورد استفاده در این مقاله از مجموعه "دادگان فارس‌دات کوچک میکروفونی" با حجم زمانی حدود شش ساعت است که نرخ نمونه‌برداری آن به 16kHz تقلیل یافته است (FARSDAT). این دادگان شامل جملات پیوستهٔ بیان شده از ۳۰۴ نفر گویندهٔ زن و مرد است که هر نفر تعداد بیست جمله را به زبان فارسی بیان کرده‌اند. جملات دادگان شامل برچسب‌دهی آوایی[5] با ۴۳ برچسب مختلف (بدون سکوت) است. در کار حاضر با تجمیع تعدادی از برچسب‌ها با یکدیگر، از تعداد سی برچسب واجی (۲۹ واج متداول فارسی به همراه یک برچسب سکوت؛ K=30) استفاده شده است (Bijankhan,2003). از داده‌های مربوط به ۲۲۰ گویندهٔ اوّل در بخش تعلیم مدل بازشناس، داده‌های ۵۴ گوینده آخر در بخش آزمون و برای بخش توسعه (Development) و اعتبارسنجی مدل شبکهٔ عصبی و الگوریتم ویتربی نیز از داده‌های سی گوینده باقیماندهٔ فارس‌دات استفاده شده است.

---

[1] Bigram
[2] Insertion Penalty
[3] Development
[4] Triphone
[5] Phonetic Labeling

۷- آزمایش‌های سامانه‌های پایه

در این بخش به بررسی نتایج بازشناسی به دست آمده از اعمال بردار ویژگی LLRPS به مدل بازشناس شبکهٔ عصبی خواهیم پرداخت. در جدول (۱) نتایج درصد دقّت بازشناسی قاب و واج آن آورده شده است.

(جدول ۱) – مقایسه نتایج آزمون درصد دقّت بازشناسی قاب و واج برای بردار ویژگی LLRPS، ویژگی متداول MFCC و روش TANDEM به‌وسیله مدل بازشناس شبکهٔ عصبی TMLP و یا مدل متداول HMM.

	ساختار مدل بازشناس	بردار ویژگی ورودی	بُعد ویژگی	٪ دقت قاب	٪ دقت واج
۱	TMLP	LLRPS	۳۰	۷۴/۲۷	۶۲/۶۰
۲	TMLP	MFCC13	۱۳	۷۹/۲۳	۶۹/۶۷
۳	TMLP	MFCC39	۳۹	۸۰/۹۸	۷۲/۲۵
۴	HMM	MFCC39	۳۹	-	۷۵/۲۲
۵	HMM	TANDEM (PPRPS13)	۱۳	-	۶۰/۰۵
۶	HMM	TANDEM (PPRPS13 +MFCC39)	۵۲	-	۷۵/۸۴

همچنین در جدول (۱)، نتایج سامانه بازشناسی پایه، شامل ویژگی‌های MFCC (به‌عنوان یک روش استخراج ویژگی پرکاربرد گفتاری) آورده شده است. برای محاسبهٔ این نتایج، بردار ویژگی ایستا MFCC، شامل سیزده ضریب کپستروم (MFCC13) و یا بردار ویژگی ایستا-پویا MFCC39، شامل سیزده ضریب کپستروم به همراه مشتقات زمانی اول و دوم آن (ضرایب دلتا و دلتادلتا) به ورودی شبکهٔ عصبی TMLP اعمال شده است. سپس مشابه با روش الگوریتم ویتربی مطرح شده در بخش ۴، نتایج بازشناسی آن به‌وسیلهٔ مقدار احتمال پسین به‌دست آمده از خروجی شبکه (که در اینجا آنها را PPMFCC می‌نامیم) به دست آمده است.

در ردیف چهارم جدول (۱)، نتایج دقّت بازشناسی واج به‌دست آمده از مدل‌سازی صوتی HMM با بردار ویژگی‌های MFCC39 (در حالت عدم استفاده از مدل زبانی بایگرم واجی در کدگشای آن) آورده شده است که به‌عنوان یک سامانه بازشناس پایه، متداول و پرکاربرد مطرح است. همچنین در دو سطر آخر جدول ۱، نتایج دقّت بازشناسی واج مدل‌سازی صوتی HMM با ویژگی‌های به‌دست آمده از روش TANDEM، که به‌ترتیب در حالات اضافه نشده یا اضافه شده به ویژگی متداول MFCC39 هستند، آورده شده است.

همان‌طور که از نتایج جدول (۱) مشاهده می‌شود، کارایی مدل بازشناس HMM (با ویژگی‌های متداول MFCC) از مدل بازشناس مبتنی بر شبکهٔ عصبی TMLP (با ویژگی‌های متداول MFCC) حدود سه درصد در دقّت بازشناسی واج بهتر است. همچنین در بین نتایج به‌دست آمده از مدل بازشناس شبکهٔ عصبی، بردار ویژگی MFCC39 که حاوی اطّلاعات ایستا و پویای گفتاری است، بالاترین درصد بازشناسی قاب و واج را به دست آورده است؛ این در حالی است که در این آزمایش بردار ویژگی‌های LLRPS و MFCC13 به‌انحصار شامل اطّلاعات ایستای هر قاب گفتاری هستند. البته لازم به ذکر است که ساختار شبکهٔ عصبی TMLP که در آن از مجموعه چندین قاب متوالی در ورودی خود استفاده می‌کند، قابلیت استخراج اطّلاعات پویا را دارد، با این وجود نتایج به‌دست آمده از جدول (۱) نشان می‌دهد که استفادهٔ مجزا از اطّلاعات دینامیک در بردار ویژگی MFCC39، به‌ترتیب منجر به بهبود ۱/۷۵٪ و ۲/۵۸٪ در نتایج دقّت بازشناسی قاب و واج نسبت به روش MFCC13 شده است. همچنین مدل بازشناس TMLP با بردار ویژگی ورودی LLRPS دقّت بازشناسی واج ۶۲/۶۰٪ را کسب کرده است که بیان‌گر کارایی پایین‌تر آن نسبت به روش متداول MFCC است.

از طرف دیگر نتیجهٔ سطر آخر جدول(۱) نشان می‌دهد که مدل‌سازی صوتی HMM با ویژگی‌های مبتنی بر RPS (PPRPS) محاسبه شده از روش TANDEM (که همان مقادیر احتمالاتی پسین پس‌پردازش شده از خروجی شبکهٔ عصبی TMLP هستند)، توانسته است در حالت افزوده‌شدن به‌ویژگی‌های متداول MFCC، دقّت بازشناسی واج ۷۵/۸۴ را کسب کند که بهبود کارایی حدود ۰/۶٪ نسبت به سامانه پایه شامل روش مدل‌سازی HMM و برداروِیژگی MFCC نشان می‌دهد.

۸- روش‌های ترکیب خروجی سامانه‌های بازشناس

استفادهٔ هم‌زمان از بردارهای ویژگی متمایز که حاوی اطّلاعات متفاوتی از یک سیگنال هستند، می‌تواند منجر به افزایش کارایی عمل‌کرد سامانه‌های بازشناسی شود (Kittler, 1998; Morgan, 2004; Nejadgholi, 2009) . در این بخش

نشان خواهیم داد که چگونه با استفاده از روش‌های مناسب ترکیب، تخمین به‌دست آمده از مقادیر احتمالات پسین در خروجی شبکه‌های حاوی بردار ویژگی LLRPS و MFCC39 (که بالاترین نرخ بازشناسی را دربرداشت) قابلیت بهبود یافتن دارد. شواهد تجربی نشان می‌دهند که اگر ماهیّت اطّلاعات کسب‌شده در روش‌های استخراج ویژگی با یکدیگر متفاوت باشند، کارایی سامانه‌های بازشناسی به‌وسیلهٔ روش‌های ترکیب مناسب بهبود خواهند یافت (Johnson, 2005; Valente, 2010).

اکنون فرض کنیم که بردارهای PPRPS(i) و PPMFCC(i) به‌ترتیب بردارهای شامل تخمین مقادیر احتمال پسین به‌دست آمده از خروجی شبکه‌های عصبی شامل ویژگی‌های LLRPS و MFCC39 برای قاب گفتاری **i**ام باشند، در این‌صورت خروجی ترکیب شده مقدارهای احتمال پسین، *PPcomb(i)*، می‌تواند به‌وسیلهٔ یکی از هفت روش ترکیب معرّفی شده در ذیل حاصل شود:

۱- روش جمع وزن‌دار:

$$PPcomb(i) = w \times PPRPS(i) + (1-w) \times PPMFCC(i)$$
$$where \quad 0 \le w \le 1$$

۲- روش ضرب وزن‌دار:

$$PPcomb(i) = PPRPS(i)^{w} \times PPMFCC(i)^{(1-w)}$$
$$where \quad 0 \le w \le 1$$

۳- روش میانگین: حالت خاص روش ۱ با مقدار **w=0.5**.

۴- روش ضرب: حالت خاص روش ۲ با مقدار **w=0.5**.

۵- روش جمع وزن‌دار متناسب با معکوس آنتروپی:

$$PPcomb(i) = w_{rps}(i) \times PPRPS(i) + w_{mfcc}(i) \times PPMFCC(i)$$
$$where \quad w_j(i) = \frac{h_j^{-1}(i)}{h_{rps}(i)^{-1} + h_{mfcc}^{-1}(i)}$$

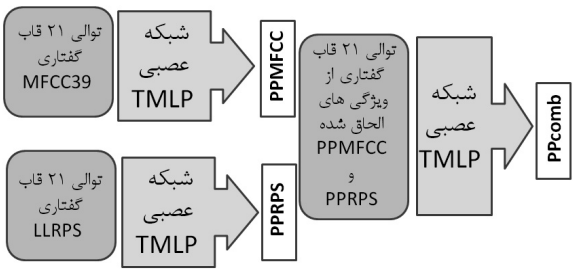
که در آن h(i) مقدار آنتروپی به‌دست آمده از بردار مقادیر احتمال پسین از قاب گفتاری **i**ام است (Misra, 2003; Kazemi, 2011).

۶- روش بیشینه (Max): در این روش برداری که حاوی بزرگ‌ترین مقدار احتمال پسین باشد، انتخاب خواهد شد.

۷- روش ترکیب غیرخطی با شبکهٔ عصبی TMLP:

$$PPcomb(i) = TMLP\{PPRPS(i), PPMFCC(i)\}$$

در این حالت، دوباره از یک ساختار شبکهٔ عصبی TMLP برای ترکیب غیرخطّی احتمالات پسین به‌دست آمده از خروجی دو شبکه MFCC39 (PPMFCC) و LLRPS (PPRPS) استفاده می‌شود. در شکل (۶) نمای کلی این روش ترکیب غیرخطی نشان داده شده است.



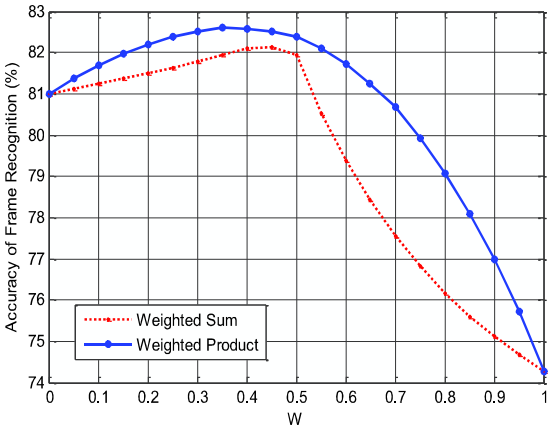
(شکل ۶) - روندنمای کلی روش ترکیب غیرخطی مقادیر احتمال پسین بوسیله ساختار شبکه عصبی TMLP.

پرواضح است که مشابه با مباحث مطرح شده در بخش ۴، خروجی این شبکه نیز تخمینی از مقدارهای احتمال پسین را تولید خواهد کرد. با توجّه به شکل (۶)، این روش شامل دو طبقه شبکهٔ عصبی است که در یک ساختار کلی سلسله‌مراتبی[1] گنجانده شده‌اند. در این ساختار، شبکه‌های عصبی موجود در طبقهٔ اوّل وظیفهٔ استخراج تخمین اوّلیهٔ مقادیر احتمال پسین (PPMFCC و PPRPS) را برعهده دارند و تنها شبکهٔ عصبی موجود در طبقهٔ دوم مسؤول ترکیب غیرخطّی تخمین‌های اوّلیه و تولید مقدار احتمال پسین نهایی (PPcomb) است.

۹- آزمایش‌ها بر روی روش‌های ترکیب و بحث و بررسی نتایج به‌دست آمده

این بخش مشتمل بر آزمایش‌های انجام گرفته برای بررسی عمل‌کرد هر یک از روش‌های ترکیب اشاره شده در بخش (۸) است. در نمودار شکل (۷)، درصد دقّت بازشناسی قابی برای روش‌های ترکیب جمع و ضرب وزن‌دار برحسب تغییرات وزن **w** نشان داده شده است. پُرواضح است که برای مقدار **w=0**، خروجی ترکیب معادل خروجی روش MFCC39 و برای **w=1**، خروجی ترکیب معادل خروجی روش LLRPS است.

[1] Hierarchical



(شکل ۷) – نمودار درصد دقّت بازشناسی قاب برحسب وزن W در حالت روش‌های ترکیب جمع (Sum) و ضرب (Product) وزن‌دار.

نتایج به‌دست آمده از نمودار شکل (۷) نشان می‌دهد که استفاده از روش‌های ترکیب جمع وزن‌دار و ضرب وزن‌دار می‌توانند منجر به بهبود نتایج سیستم بازشناسی ترکیبی نسبت به سیستم بازشناس پایه با ویژگی‌های متداول MFCC39 شود. همچنین کارایی روش ترکیبی ضرب وزن‌دار از روش جمع وزن‌دار بالاتر است.

با اعمال تابع لگاریتم در رابطه روش ترکیب ضرب وزن‌دار، این روش تبدیل به روش جمع وزن‌دار بر روی مقادیر لگاریتم احتمال پسین می‌شود. از آنجا که تابع لگاریتم، مقادیر کمتر از یک (محدوده مقادیر احتمالات پسین) را به صورت غیرخطی و با وضوح مناسبی بازنمایی می‌کند، نتایج به‌دست آمده از شکل (۷) مؤیّد این مطلب است که استفاده از تابع لگاریتم در آنالیز مقادیر احتمال پسین برای کاربردهای بازشناسی بسیار مؤثّرتر است. در ادامه و در جدول (۲) نتایج آزمون دقّت بازشناسی قاب و واج برای تمامی روش‌های ترکیب بردارهای PPRPS و PPMFCC آورده شده است. در اینجا تنها از اعمال الگوریتم ویتربی بر روی مقدار ترکیبی احتمال پسین (PPcomb) برای تعیین خروجی بازشناسی در سطح آوایی واج استفاده شده است.

(جدول ۲) – مقایسه نتایج آزمون دقت بازشناسی قاب و واج برای روش‌های متنوع ترکیب بردارهای PPRPS و PPMFCC.

	روش ترکیب خروجی شبکه‌ها	٪ دقت قاب	٪ دقت واج
۱	جمع وزن‌دار *(w=0.45)*	۸۲/۱۳	۷۳/۵۷
۲	ضرب وزن دار *(w=0.35)*	۸۲/۵۹	۷۴/۴۵
۳	میانگین *(w=0.5)*	۸۱/۹۵	۷۳/۲۶
۴	ضرب *(w=0.5)*	۸۲/۳۷	۷۴/۲۲
۵	جمع وزن‌دار متناسب با معکوس آنتروپی	۸۱/۷۸	۷۳/۴۲
۶	بیشینه *(Max)*	۸۱/۶۱	۷۳/۱۰
۷	غیرخطی *TMLP*	**۸۴/۹۲**	**۷۶/۲۷**

با توجّه به نتایج ارائه شده در جدول (۲)، استفاده از روش‌های ترکیب، همواره منجر به بهبود نتایج دقّت بازشناسی قاب و واج نسبت به نتایج به‌دست آمده از هر یک از سامانه‌ها در حالت بدون ترکیب جدول (۱) شده است. بیشترین بهبود کارایی با روش ترکیب غیرخطی TMLP (سطر آخر جدول) به دست آمده است که به‌ترتیب با درصد دقّت بازشناسی قاب و واج ۸۴/۹۲٪ و ۷۶/۲۷٪، توانسته است نسبت به روش پایه MFCC39 (با مدل بازشناس TMLP) به ترتیب بهبود ۳/۹۴٪ و ۴/۰۲٪ داشته باشد. همچنین با مقایسهٔ درصد دقّت بازشناسی واج این روش (ترکیب غیرخطی خروجی‌های PPMFCC و PPRPS با مدل شبکه عصبی TMLP) با نتیجه به‌دست آمده از سامانه پایه مبتنی بر HMM (ارائه شده در ردیف ۴ جدول (۱)، دیده می‌شود که روش ترکیب غیرخطی پیشنهادی (با مدل شبکه عصبی TMLP) توانسته است بهبود عمل‌کردی در حدود یک‌درصد در نتیجه دقّت بازشناسی واج نشان دهد.

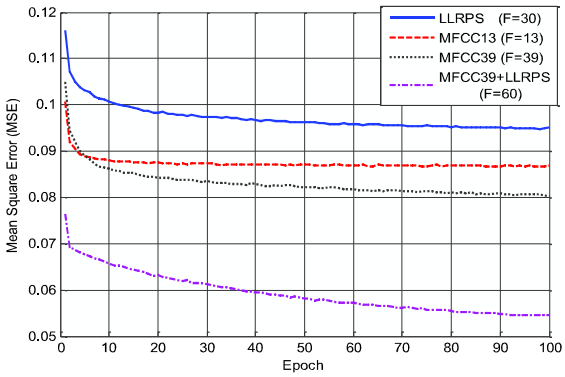
با توجه به نتایج جداول (۱) و (۲) موارد ذیل قابل نتیجه‌گیری است:

- در شرایط یکسان، عمل‌کرد بازشناسی گفتار با مدل HMM بهتر از مدل شبکهٔ عصبی TMLP است.
- ترکیب غیرخطی اطّلاعات در سطح خروجی با روش شبکه عصبی TMLP نتیجه بازشناسی بهتری نسبت به ترکیب غیرخطی اطّلاعات در سطح ویژگی با ساختار TANDEM و روش HMM داشته است.
- در مقایسه با نتایج ارائه‌شده در (Shekofteh, 2013) که در آن از اعمال نگاشت خطی LDA بر روی ویژگی‌های اوّلیهٔ

LLRPS اعمال شده است، استفاده از نگاشت غیرخطی ویژگی‌ها با روش شبکه عصبی و ساختار TANDEM، منجر به بهبود نتایج دقّت بازشناسی واج از ۵۶/۶۹ به ۶۰/۰۵ در بازشناسی با HMM شده است.

همان‌طور که از نتایج جدول (۲) مشاهده شد، عمل‌کرد روش ترکیب غیرخطی TMLP، با اعمال مجدد ویژگی‌های PPRPS و PPMFCC به شبکهٔ عصبی TMLP در یک ساختار سلسله‌مراتبی، توانست بالاترین نتیجه دقّت بازشناسی را کسب کند. یک توجیه مناسب برای بهبود عمل‌کرد این روش، بررسی نوع ویژگی‌های ورودی به شبکه‌های عصبی در هر یک از طبقات ساختار سلسله‌مراتبی شکل (۶) است. به‌عنوان مثال ویژگی متداول MFCC از نوع ویژگی‌های صوتی مبتنی بر مدل منبع-فیلتر[1] است که درجه بالایی از تغییرات غیرزبانی[2] مانند مشخّصه‌های گوینده و محیط (مانند نوفه و کانال) را شامل می‌شود. این در حالی است که ویژگی‌های مبتنی بر احتمالات پسین دارای توزیع تُنُک[3] بوده و به‌دلیل خاصیّت تفکیک‌پذیری بالاتر آن نسبت‌به ویژگی‌های صوتی، کمتر تحت تأثیر مشخّصه‌های هم‌تولیدی[4] گفتاری قرار می‌گیرند (Ellis, 2001; Sivadas, 2002). بنابراین یادگیری الگوهای گفتاری توسط شبکه‌های عصبی موجود در طبقهٔ اوّل ساختار شکل (۶)، دارای پیچیدگی فراوان‌تری نسبت‌به طبقه دوم آن است که باید آموزش الگوهای گفتاری را از روی ویژگی‌های مبتنی بر احتمالات پسین بر روی توالی قاب‌های گفتاری انجام دهد.

برای بررسی بهتر این موضوع، در شکل (۸) نمودار خطای تعلیم (MSE) شبکهٔ عصبی TMLP در حالت استفاده از ویژگی‌های ورودی LLRPS با بُعد ویژگی F=30، MFCC13 با F=13، MFCC39 با F=39 و شبکهٔ عصبی TMLP در طبقهٔ دوم ساختار شکل (۶) با بردار ترکیب PPMFCC+PPRPS و F=60، برای صد تکرار اوّل مرحلهٔ آموزش آنها آورده شده است.



(شکل ۸) - نمودار خطای تعلیم شبکه TMLP در صد تکرار اوّل برای بردار ویژگی‌های ورودی مختلف با بُعد ورودی F.

با توجّه به نمودارهای خطای رسم‌شده در شکل (۸)، شبکهٔ TMLP دربرگیرنده ویژگی‌های PPMFCC+PPRPS، خطای تعلیم کم‌تری را نسبت‌به دیگر سامانه‌های مورد استفاده کسب کرده است که این نیز می‌تواند ملاک مناسبی برای توجیه عمل‌کرد بهتر آن در فرآیند یادگیری و نتایج ارائه‌شده در جدول‌های (۱) و (۲) باشد.

همان‌طور که در بخش مقدّمه اشاره شد، در روش‌های متداول استخراج ویژگی از سیگنال گفتار به‌طور معمول از روش‌های مبتنی بر طیف و کپستروم (مانند LFBE، MFCC و PLP) استفاده می‌شود که در الگوریتم آنها اطّلاعات فاز سیگنال گفتار به‌طور کامل حذف شده است. نتایج به‌دست آمده در آزمایش‌های این بخش مؤیّد این مطلب است که ترکیب مناسب ویژگی‌های کپستروم (MFCC) که فاقد اطّلاعات فاز سیگنال است با ویژگی‌های پیشنهادی (LLRPS) استخراج شده از RPS که حاوی اطّلاعات حذف نشده فاز سیگنال است، توانسته است به‌علّت هم‌افزابودن اطّلاعات آنها، موجب بهبود عمل‌کرد نهایی سامانه بازشناس در حالت ترکیبی شود.

۱۰- نتیجه‌گیری

در این مقاله روشی مبتنی بر استفاده از اطّلاعات فضای بازسازی‌شدهٔ فاز (RPS) سیگنال گفتار برای بهبود عمل‌کرد یک سامانه بازشناسی گفتار پیوسته واج، مطرح و بررسی شد. در این روش علاوه‌بر استفاده از بردار ویژگی‌های متداول گفتاری، از اطّلاعات به‌دست آمده از یک روش نوین استخراج ویژگی مبتنی بر فضای بازسازی‌شده فاز استفاده شده است. این ویژگی‌ها (LLRPS) بر مبنای ساختار هندسی توزیع سیگنال گفتار جاسازی‌شده در حوزهٔ RPS حاصل شده‌اند.

[1] Source-Filter Model
[2] Nonlinguistic
[3] Sparse
[4] Coarticulation

برای این‌منظور باید در مرحلهٔ تعلیم، یک‌سری مدل‌های مانیفولد واجی با مدل‌سازی GMM تعلیم داده شوند. ویژگی‌های نهایی، از امتیاز شباهت (مبتنی بر لگاریتم درست‌نمایی) تراژکتوری سیگنال گفتار جاسازی‌شده در RPS با GMMهای گفتاری تعلیم‌یافته در RPS محاسبه شوند. در اینجا نشان دادیم که با استفاده از مدل بازشناس غیرخطی TMLP، عمل‌کرد ویژگی پیشنهادی LLRPS از ویژگی‌های متداول MFCC پایین‌تر است. اما با ترکیب مناسب خروجی‌های به‌دست آمده از دو سیستم بازشناس که حاوی اطّلاعات متمایز نسبت به یکدیگر هستند، می‌توانیم نتایج دقّت قاب و واج بازشناسی‌شده را به‌ترتیب به مقدار ۳/۹۴٪ و ۴/۰۲٪ نسبت به بهترین سامانه بازشناس پایه مبتنی بر مدل TMLP افزایش دهیم. همچنین با استفاده از ساختار TANDEM، نشان داده شد که الحاق ویژگی‌های مبتنی بر RPS می‌تواند کارایی ویژگی‌های MFCC را حدود ۰/۶٪ بهبود دهد.

مراجع

رحیمی نژاد، مهدی؛ سیّدصالحی، سیّدعلی؛ ۱۳۸۲. مقایسه و ارزیابی کارایی انواع روش‌های استخراج پارامترهای بازنمایی و هنجارسازی در بازشناسی مستقل از گویندهٔ گفتار، نشریه امیرکبیر، ش. ۵۵.

جعفری، ایوب؛ الماس‌گنج، فرشاد؛ نبی بیدهندی، مریم؛ ۱۳۸۹. مدل‌سازی غیرخطی قطع پوانکاره سیگنال گفتار در ترکیب با تحلیل حوزهٔ فرکانس به‌منظور افزایش صحّت عمل‌کرد سیستم‌های بازشناسی گفتار، نشریه فنی و مهندسی مدرس، دوره ۱۰، ش. ۳، ص. ۵۵ - ۷۰.

ولی، منصور؛ سیّدصالحی، سیّدعلی؛ ۱۳۸۵. بازشناسی مقاوم و توأم گفتار مستقیم و تلفنی با استخراج مناسب بردارهای بازنمایی و اصلاح آنها توسط معکوس‌سازی شبکه‌های عصبی، نشریه مهندسی برق و مهندسی کامپیوتر ایران، سال ۴، ش. ۱، ص. ۲۱ - ۲۹.

شکفته، یاسر؛ الماس گنج، فرشاد؛ ۱۳۸۶. بهبود بازشناسی گفتار با استفاده از شبکه های عصبی دربرگیرنده الگوهای زمانی، مجموعه مقالات سومین کنفرانس بین‌المللی فناوری اطّلاعات و دانش (IKT2007)، دانشگاه فردوسی مشهد، آذر ۱۳۸۶.

Abarbanel, H.D.I., 1996. Analysis of observed chaotic data. Springer, New York.

Alsteris, L.D., Paliwal, K.K., 2007. Short-time phase spectrum in speech processing: A review and some experimental results. Digital Signal Processing, vol. 17, pp. 578–616.

Awrejcewicz J., 1990. Bifurcation portrait of the human vocal cord oscillation. Journal of Sound Vibrations. vol. 136, pp. 151–156.

Banbrook, M., McLaughlin, S., 1996. Dynamical modelling of vowel sounds as a synthesis tool. In Proc. ICSLP, pp. 1981-1984.

Berry, D.A., Herzel, H., Titze, I.R., Krischer K., 1994. Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. The Journal of the Acoustical Society of America, vol. 95, pp. 3595–3604.

Bijankhan M., Sheykhzadegan J., Roohani M.R., Zarrintare R., Ghasemi S.Z. and Ghasedi M.E., 2003. TFarsDat - The Telephone Farsi Speech Database. In Proc. EuroSpeech, Geneva, Switzerland.

Bourlard, H., Morgan, N., 1994. Conn-ectionist speech recognition - a hybrid approach. Norwell, MA: Kluwer.

Chen, B., Zhu, Q., Morgan, N., 2005. Tonotopic multi-layer perceptron a neural network for Learning long-term temporal features for speech recognition. In Proc. ICASSP, pp. 945-948.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences. IEEE Trans. Speech and Audio Processing, vol. 28(4), pp. 357-366.

Ellis, D., Singh, R., Sivadas, S., 2001. Tandem acoustic modeling in large vocabulary recognition. In Proc. ICASSP, pp. 517–520.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoustics, Speech and Signal Processing, vol. 34 (1), pp. 52-59.

FARSDAT, Persian speech database: <http://catalog.elra.info/product_info.php?products_id=18>.

Hagmuller, M., Kubin, G., 2006. Poincare pitch marks. Speech Communication, vol. 48, pp. 1650–1665.

Hegde, R. M., Murthy, H.A., Gadde, V.R.R., 2007. Significance of the modified group delay feature in speech recognition, IEEE Trans. Audio, Speech and Language Processing, vol. 15 (1), pp. 190–202.

Kennel, M.B., Brown, R., Abarbanel, H.D.I., 1992. Determining embedding dimension for phase-space reconstruction using a geometrical construction. Physical review A, vol. 45 (6), pp. 3403–3411.

Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., 1998. On combining classifiers. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20(4), pp. 226–239.

Kokkinos, I., Maragos, P., 2005. Nonlinear speech analysis using models for chaotic systems. IEEE Trans. Speech Audio Processing, vol. 13, pp. 1098–1109.

Lindgren, A.C., Johnson, M.T., Povinelli, R.J., 2003. Speech recognition using reconstructed phase space features. In Proc. ICASSP, China, pp. 61–63.

Lindgren, A.C., Johnson, M.T., Povinelli, R.J., 2004. Joint frequency domain and reconstructed phase space features for speech recognition. In Proc. ICASSP, Montreal, Canada, pp. I-533–I-536.

Liu, X., Povinelli, R.J., Johnson, M.T., 2003. Vowel classification by global dynamic modeling. In Proc. NOLISP, pp. 111–114.

Misra, H., Bourlard, H., Tyagi, V., 2003. New entropy based combination rules In HMM/ANN multi-stream ASR. In Proc. ICASSP, pp. 741–744.

Morgan, N., Chen, B.Y., Zhu, Q., Stolcke, A., 2004. TRA-Pping conversational speech: Extending TRAP/Tandem approaches to conversational telephone speech recognition. In Proc. ICASSP, pp. 536–539.

Narayanan, N.K., Thasleema, T.M., Prajith, P., 2012. Reconstructed state space model for recognition of consonant - vowel utterances using support vector machines. International Journal of Artificial Intelligence and Applications, vol. 3(2), pp. 101-119.

Narayanan, S.S., Alwan, A.A., 1995. A nonlinear dynamical systems analysis of fricative consonants. Acoustical Society of America Journal, vol. 97, pp. 2511-2524.

Nejadgholi, I., Seyyedsalehi, S.A., 2009. Nonlinear normalization of input patterns to speaker variability in speech recognition neural networks. Neural Computing and Applications, vol. 18, pp. 45–55.

Paliwal, K., Alsteris, L., 2005. On the usefulness of STFT phase spectrum in human listening tests. Speech Communication, vol. 45, pp. 153–170.

Park, J., Diehl, F., Gales, M.J.F., Tomalin, M., Woodland, P.C., 2011. The efficient incorporation of MLP features into automatic speech recognition

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. Journal of the Acoustic Society of America, vol. 87(4), pp. 1738-1752.

Hermansky, H., Ellis, D.P., Sharma, S., 2000. Tandem connectionist feature extraction for conventional HMM systems. In Proc. ICASSP, pp. 1635-1638.

Herzel, H., Berry, D., Titze, I., Steinecke, I., 1995. Nonlinear dynamics of the voice: signal analysis and biomechanical modeling. Chaos, vol. 5, pp. 30–34.
HTK (v.3.4), Hidden Markov Model Toolkit: < http://htk.eng.cam.ac.uk/>

Indrebo, K.M., Povinelli, R.J., Johnson, M.T., 2006. Sub-banded reconstructed phase spaces for speech recognition. Speech Communication, vol. 48, pp. 760-774.

Ikbal, S., Misra, H., Hermansky, H., Magimai, M., 2012. Phase autocorrelation (PAC) features for noise robust speech recognition, Speech Communication, vol. 54, pp. 867–880.

Jafari, A., Almasganj, F., 2012. Using nonlinear modeling of reconstructed phase space and frequency domain analysis to improve automatic speech recognition performance. International Journal of Bifurcation and Chaos, vol. 22(3).

Jafari, A., Almasganj, F., NabiBidhendi, M., 2010. Statistical modeling of speech Poincaré sections in combination of frequency analysis to improve speech recognition performance. Chaos, vol. 20, pp. 033106:1-11.

Jiang, J.J., Zhang, Y., 2002. Chaotic vibration induced by turbulent noise in a two-mass model of vocal folds. The Journal of the Acoustical Society of America, vol. 112, pp. 2127–2133.

Jiang, J.J., Zhang, Y., McGilligan, C., 2006. Chaos in voice, from modeling to measurement. Journal of Voice, vol. 20(1), pp. 2-17.
Johnson, M.T., Povinelli, R.J., Lindgren, A.C., Ye, J., Liu, X., Indrebo, K.M., 2005. Time-domain isolated phoneme classification using reconstructed phase spaces. IEEE Trans. Speech Audio Processing, vol. 13(4), pp. 458–466.

Kantz, H., Schreiber, T., 1997. Nonlinear Time Series Analysis Cambridge University Press, Cambridge, England.

Kazemi, A.R., Sobhanmanesh, F., 2011. MLP refined posterior features for noise robust phoneme recognition. Scientia Iranica, Trans. D: Computer Science & Engineering and Electrical Engineering, vol. 18, pp. 1443–1449.

consonant classification. International Journal of Speech Technology, vol. 15(2), pp. 227-239.

Tohidypour, H.R., Seyyedsalehi, S.A., Behbood, H., Roshandel, H., 2012. A new representation for speech frame recognition based on redundant wavelet filter banks. Speech Communication, vol. 54(2), pp. 256-271.

Valente, F., 2010. Multi-stream speech recognition based on Dempster–Shafer combination rule. Speech Communication, vol. 52(3), pp. 213-222.

Vaziri, G., Almasganj, F., Behroozmand, R., 2010. Pathological assessment of patients' speech signals using nonlinear dynamical analysis. Computers in Biology and Medicine, vol. 40(1), pp. 54-63.

White, H., 1989. Learning in artificial neural networks: A statistical perspective, Neural Computation, pp. 425-464.

Ye, J., Johnson, M.T. M.T., Povinelli, R.J., 2003. Phoneme classification over reconstructed phase space using principal component analysis. In Proc. NOLISP, Le Croisic, France, pp. 11–16.

Ye, J., Povinelli, R.J., Johnson, M.T., 2002. Phoneme Classification Using naïve Bayes Classifier in Reconstructed Phase Space. In Proc. IEEE Digital Signal Processing Workshop, Atlanta, Georgia.

Yu, S., Zheng, D., Feng, X., 2006. A new time domain feature parameter for phoneme classification. In Proc. WESPAC IX 2006, Seoul, Korea.

Zavaliagkos, G., Zhao, Y., Schwartz, R., Makhoul, J., 1994. A hybrid seg-mental neural net/hidden Markov model system for continuous speech recognition, IEEE Trans. Speech Audio Processing, vol. 2 (1) pp. 151-160.

Zhu, Q., Chen, B., Morgan, N., Stolcke, A., 2004. On using MLP features in LVCSR. In Proc. ICSLP.

یاسر شکفته مدارک کارشناسی، کارشناسی ارشد و دکترای خود را در رشته مهندسی پزشکی (گرایش بیوالکتریک) به‌ترتیب در سال‌های ۱۳۸۴، ۱۳۸۷ و ۱۳۹۲ از دانشکدهٔ مهندسی پزشکی دانشگاه صنعتی امیرکبیر اخذ نموده است. زمینه‌های تحقیقاتی مورد علاقهٔ ایشان شامل پردازش

systems. Computer Speech and Language, vol. 25, pp. 519–534.

Pinto, J., Garimella, S., Magimai-Doss, M., Hermansky, H., Bourlard, H., 2011. Analysis of MLP-Based hierarchical phoneme posterior probability estimator. IEEE Trans. Audio Speech Language Processing, vol. 19(1), pp. 225-241.

Pitsikalis, V., Maragos, P., 2002. Speech analysis and feature extraction using chaotic models. In Proc. ICASSP, Orlando, Florida, pp. 533-536.

Pitsikalis, V., Maragos, P., 2009. Analysis and classific-ation of speech signals by generalized fractal dimension features. Speech Communication, vol. 51(12), pp. 1206-1223.

Povinelli, R.J., Johnson, M.T., Lindgren, A.C., Roberts, F.M., Ye, J., 2006. Statistical models of reconstructed phase spaces for signal classification. IEEE Trans. Signal Processing, vol. 54, pp. 2178–2186.

Povinelli, R.J., Johnson, M.T., Lindgren, A.C., Ye, J., 2004. Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Trans. Knowledge and Data Engineering, vol. 16, pp. 779–783.

Sameti, H., Veisi, H., Bahrani, M., Babaali, B., Hosseinzadeh, K., 2011. A large vocabulary continuous speech recognition system for Persian language. EURASIP Journal on Audio, Speech, and Music Processing, vol. (2011-1), pp. 1-12.

Sauer, T., Yorke, J.A., Casdagli, M., 1991. Embedology. Journal of Statistical Physics, vol. 65, pp. 579–616.

Shekofteh, Y., Almasganj, F., 2013. Feature extraction based on speech attractors in the reconstructed phase space for automatic speech recognition systems, ETRI Journal, vol. 35(1), pp. 100-108.

Sivadas, S., Hermansky, H., 2002. Hierarchical tandem feature extraction, In Proc. ICASSP, pp. 809–812.

Sun, J., Zheng, N., Wang, X., 2007. Enhancement of Chinese speech based on nonlinear dynamics. Signal Processing, vol. 87, pp. 2431–2445.

Takens, F., 1980. Detecting strange attractors in turbulence. In Proc. Dynamical System Turbulence, pp. 366–381.

Thasleema, T.M., Prajith, P., Narayanan, N.K., 2012. Time–domain non-linear feature parameter for

سیگنال‌های حیاتی، شناسایی الگو، مدل‌سازی سیستم‌های بیولوژیکی و سیستم‌های بازشناسی گفتار است.

نشانی رایانامهٔ ایشان عبارتست از:

y_shekofteh@{aut.ac.ir,yahoo.com}

فرشاد الماس‌گنج مدارک کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی برق (گرایش الکترونیک) به‌ترتیب در سال‌های ۱۳۶۳ و ۱۳۶۷ از دانشگاه صنعتی امیرکبیر اخذ نموده است. وی سپس در سال ۱۳۷۷ مدرک دکترای خود را در گرایش مهندسی پزشکی از دانشکدهٔ مهندسی برق دانشگاه تربیت مدرس اخذ نمود. ایشان هم اکنون عضو هیئت علمی دانشکدهٔ مهندسی پزشکی دانشگاه صنعتی امیرکبیر با سمت و درجهٔ علمی دانشیار هستند. زمینه‌های تحقیقاتی مورد علاقه ایشان شامل پردازش سیگنال‌های رقمی، شناسایی الگو و مدل‌سازی صوتی و زبانی در سیستم‌های بازشناسی گفتار است.

نشانی رایانامهٔ ایشان عبارتست از:

almas@aut.ac.ir